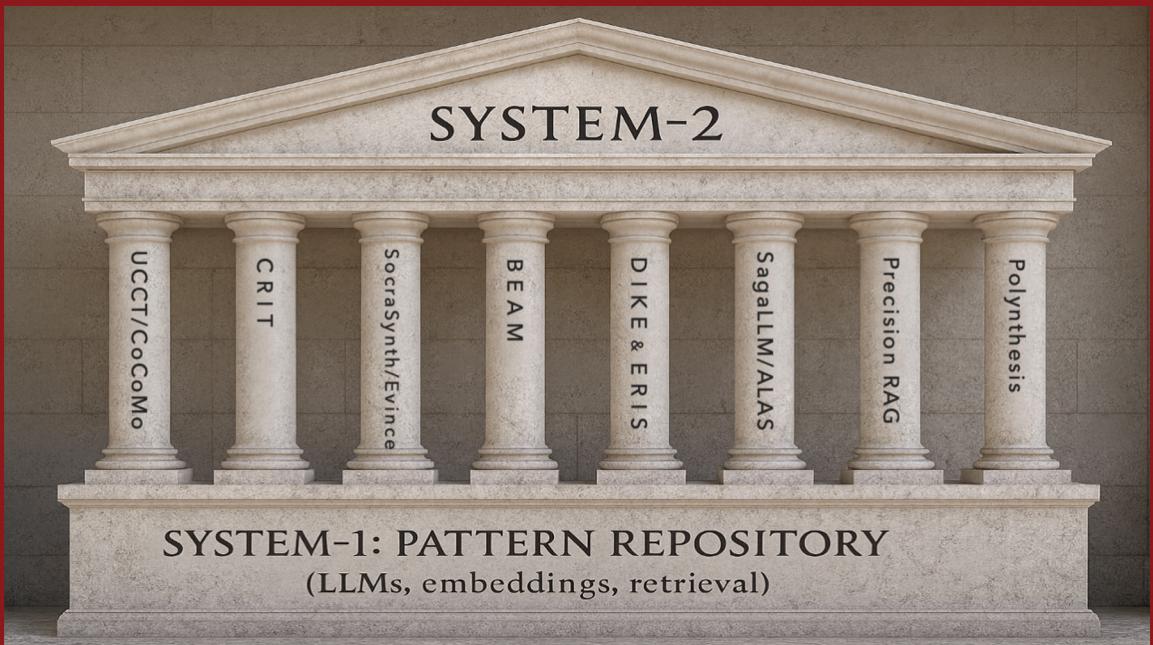


# Multi-LLM Agent Collaborative Intelligence

*The Path to Artificial General Intelligence*



**Edward Y. Chang**



ASSOCIATION FOR COMPUTING MACHINERY

# **Multi-LLM Agent Collaborative Intelligence**



# ACM Books

## Editors in Chief

Marta Kwiatkowska, *University of Oxford, UK*  
Partha Roop, *University of Auckland, New Zealand*

ACM Books is a new series of high-quality books for the computer science community, published by ACM. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

### Biomedical embedded systems: From design to security

Nathan Allen, *Department of Computer and Information Sciences, Auckland University of Technology*  
Hammond Pearce, *School of Computer Science and Engineering, University of New South Wales*  
Partha Roop, *Department of Electrical, Computer, and Software Engineering, The University of Auckland*  
2025

### Functional Data Structures and Algorithms: A Proof Assistant Approach

Editor: Tobias Nipkow, *Technical University of Munich*  
2025

### Thinking about Programs

Gavin Lowe, *St. Catherine's College, Oxford and Department of Computer Science, Oxford*  
2025

### Sustainable Interactive Wireless Stickers: From Materials to Devices to Applications

Nivedita Arora, *Northwestern University*  
2025

### Calculated Imagery: A History of Computer Graphics in Hollywood Cinema

Mark J. P. Wolf, *Department of Digital Media & Design, Concordia University Wisconsin*  
2025

### Indistinguishability Obfuscation from Well-Studied Assumptions

Aayush Jain, *Department of Computer Science, Carnegie Mellon University*  
2025

### Formal Verification of Just-in-Time Compilation

Aurèle Barrière, *École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*  
2025



# **Multi-LLM Agent Collaborative Intelligence**

***The Path to Artificial General Intelligence***

**Edward Y. Chang**

*Computer Science, Stanford University*

*ACM Books #69*



Copyright © 2026 by Association for Computing Machinery

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews—without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which the Association of Computing Machinery is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

*Multi-LLM Agent Collaborative Intelligence: The Path to Artificial General Intelligence*  
Edward Y. Chang

books.acm.org  
<http://books.acm.org>

ISBN: 979-8-4007-3179-2 hardcover  
ISBN: 979-8-4007-3178-5 paperback  
ISBN: 979-8-4007-3173-0 EPUB  
ISBN: 979-8-4007-3172-3 eBook

Series ISSN: 2374-6769 print 2374-6777 electronic

DOIs:

<a href="https://doi.org/10.1145/3749421">10.1145/3749421</a> Book	<a href="https://doi.org/10.1145/3749421.3749433">10.1145/3749421.3749433</a> Chapter 9
<a href="https://doi.org/10.1145/3749421.3749422">10.1145/3749421.3749422</a> Preface	<a href="https://doi.org/10.1145/3749421.3749434">10.1145/3749421.3749434</a> Chapter 10
<a href="https://doi.org/10.1145/3749421.3749423">10.1145/3749421.3749423</a> Acknowledgement	<a href="https://doi.org/10.1145/3749421.3749435">10.1145/3749421.3749435</a> Chapter 11
<a href="https://doi.org/10.1145/3749421.3749424">10.1145/3749421.3749424</a> Introduction	<a href="https://doi.org/10.1145/3749421.3749436">10.1145/3749421.3749436</a> Chapter 12
<a href="https://doi.org/10.1145/3749421.3749425">10.1145/3749421.3749425</a> Chapter 1	<a href="https://doi.org/10.1145/3749421.3749437">10.1145/3749421.3749437</a> Chapter 13
<a href="https://doi.org/10.1145/3749421.3749426">10.1145/3749421.3749426</a> Chapter 2	<a href="https://doi.org/10.1145/3749421.3749438">10.1145/3749421.3749438</a> Chapter 14
<a href="https://doi.org/10.1145/3749421.3749427">10.1145/3749421.3749427</a> Chapter 3	<a href="https://doi.org/10.1145/3749421.3749439">10.1145/3749421.3749439</a> Chapter 15
<a href="https://doi.org/10.1145/3749421.3749428">10.1145/3749421.3749428</a> Chapter 4	<a href="https://doi.org/10.1145/3749421.3749440">10.1145/3749421.3749440</a> Chapter 16
<a href="https://doi.org/10.1145/3749421.3749429">10.1145/3749421.3749429</a> Chapter 5	<a href="https://doi.org/10.1145/3749421.3749441">10.1145/3749421.3749441</a> Chapter 17
<a href="https://doi.org/10.1145/3749421.3749430">10.1145/3749421.3749430</a> Chapter 6	<a href="https://doi.org/10.1145/3749421.3749442">10.1145/3749421.3749442</a> Epilogue
<a href="https://doi.org/10.1145/3749421.3749431">10.1145/3749421.3749431</a> Chapter 7	<a href="https://doi.org/10.1145/3749421.3749443">10.1145/3749421.3749443</a> Bio/Index
<a href="https://doi.org/10.1145/3749421.3749432">10.1145/3749421.3749432</a> Chapter 8	

A publication in the ACM Books series, #69

Editors in Chief: Marta Kwiatkowska, *University of Oxford, UK*

Partha Roop, *University of Auckland, New Zealand*

Area Editor: Shixia Liu, *School of Software, Tsinghua University, China*

This book was typeset in Arnhem Pro 10/14 and Flama using LuaTeX.

Third Edition

10 9 8 7 6 5 4 3 2 1

*Thanks to my family for their love and support.*



# Contents

## Preface xv

## Acknowledgments xix

Acknowledgment of AI Assistance xix

## Introduction to The MACI Framework: A System 2 Architecture for Artificial General Intelligence xxi

The Intelligence Paradox xxi

From Pattern Matching to Artificial General Intelligence xxi

Eight Foundational Pillars xxii

From System 1 Foundation to Cognitive AGI xxiii

Necessity and Sufficiency of the Eight Foundational Pillars xxiii

Book Organization xxiv

## Chapter 1 A Brief History of AI: From Turing to Transformers 1

Abstract 1

1.1 Introduction 1

1.2 Definitions 4

1.3 Perspectives on Similarity 6

1.4 Eras of Similarity Measurement 13

1.5 Concluding Remarks 22

References 22

## Chapter 2 Capabilities and Opportunities of Language Models 27

Abstract 27

2.1 Introduction 27

2.2 Distinctive Capabilities 29

2.3 Exploring Unknown Unknowns 36

2.4 Conclusion 45

References 46

**Chapter 3 Prompt Engineering: Few Shots, Chain of Thought, and Retrieval-Augmented Generation 49**

- Abstract 49
- 3.1 Introduction 49
- 3.2 Prompting Methods 50
- 3.3 RAG 57
- 3.4 Concluding Remarks 60
- References 61

**Chapter 4 Unified Cognitive Consciousness Theory: Introduction 63**

- Abstract 63
- 4.1 Introduction 63
- 4.2 Related Work 66
- 4.3 The UCCT Framework 68
- 4.4 Empirical Study: Qualitative Anchoring Demonstrations 72
- 4.5 Conclusion 80
- References 82

**Chapter 5 CRIT: Socratic Inquiry for Critical Thinking in LLMs 87**

- Abstract 87
- 5.1 Introduction 87
- 5.2 Related Work 90
- 5.3 The Socratic Method 92
- 5.4 Prompt Template Engineering 99
- 5.5 Pilot Study 106
- 5.6 Concluding Remarks 108
- Appendix 109
- References 109

**Chapter 6 SocraSynth: Adversarial Multi-LLM Reasoning 113**

- Abstract 113
- 6.1 Introduction 113
- 6.2 Multi-Agent SocraSynth Overview 116
- 6.3 Empirical Study 123
- 6.4 Remarks on Related Work 133
- 6.5 Concluding Remarks 134
- 6.6 Supplemental Materials 136
- 6.A Appendix A. Tables of Topic Proposal and Refinement 136
- 6.B Appendix B. Moderator's Prompts 136

<b>6.C</b>	Appendix C. Debate from Opening to Fourth Round	<b>138</b>
<b>6.D</b>	Appendix D. CRIT Evaluations by the Judges	<b>147</b>
<b>6.E</b>	Appendix E. Debate Beats Q&A in Quality	<b>151</b>
<b>6.F</b>	Appendix F. Post-Debate Conciliatory Remarks	<b>152</b>
<b>6.G</b>	Appendix G. Healthcare Debate	<b>153</b>
	References	<b>157</b>

**Chapter 7 EVINCE: Optimizing Adversarial LLM Dialogues via Conditional Statistics and Information Theory **163****

	Abstract	<b>163</b>
<b>7.1</b>	Introduction	<b>164</b>
<b>7.2</b>	Related Work	<b>166</b>
<b>7.3</b>	The EVINCE Algorithm and Its Foundations	<b>168</b>
<b>7.4</b>	Empirical Evaluation of EVINCE	<b>171</b>
<b>7.5</b>	Concluding Remarks	<b>178</b>
<b>7.A</b>	Appendix A: Metrics for LLM Debate Evaluation	<b>180</b>
<b>7.B</b>	Appendix B: Theorem Proving	<b>181</b>
<b>7.C</b>	Appendix C: Design Maxims of EVINCE Moderation	<b>183</b>
<b>7.D</b>	Appendix D: EVINCE Debate #1	<b>184</b>
<b>7.E</b>	Appendix E: EVINCE Debate—Jaundice vs. Hepatitis	<b>187</b>
	References	<b>193</b>

**Chapter 8 Bias Correction and Data Synthesis through Reflective LLM Collaboration **197****

	Abstract	<b>197</b>
<b>8.1</b>	Introduction	<b>197</b>
<b>8.2</b>	Related Work	<b>199</b>
<b>8.3</b>	Methodology	<b>202</b>
<b>8.4</b>	Experiments	<b>207</b>
<b>8.5</b>	Concluding Remarks	<b>214</b>
<b>8.A</b>	Appendix A: On Annotation Quality	<b>215</b>
<b>8.B</b>	Appendix B: Summary of EVINCE Debate on News D1	<b>216</b>
<b>8.C</b>	Appendix C: Experiment #1 Justifications of Biased Articles	<b>218</b>
	References	<b>223</b>

**Chapter 9 Modeling Emotions in Multimodal LLMs **227****

	Abstract	<b>227</b>
<b>9.1</b>	Introduction	<b>227</b>
<b>9.2</b>	Qualifying and Quantifying Emotions	<b>229</b>
<b>9.3</b>	Empirical Study: Linguistic Features of Emotion	<b>233</b>

9.4	Qualifying and Quantifying Ethics	238
9.5	Concluding Remarks	243
9.A	Appendix A	243
	References	245

**Chapter 10 A Checks-and-Balances Framework for Context-Aware Ethical AI Alignment** **247**

	Abstract	247
10.1	Introduction	247
10.2	Related Work	251
10.3	Three-Branch Framework Design	253
10.4	Empirical Studies	260
10.5	Conclusion	265
10.6	Appendices	266
10.A	Appendix A: Unconscious-Conscious Complementarity Thesis	266
10.B	Appendix B: Wheels of Emotions	269
10.C	Appendix C: Complex Emotions	269
10.D	Appendix D: Hate Speech Samples	272
10.E	Appendix E: Mixed Emotions	272
10.F	Appendix F: Instruction to Human Annotators	276
10.G	Appendix G: Polarized Emotions in One Article	277
10.H	Appendix H: “To My Sister” of Different Linguistic Behaviors	280
	References	281

**Chapter 11 SagaLLM: Persistent Context Management, Constraint Validation, and Transaction Guarantees** **287**

	Abstract	287
11.1	Introduction	288
11.2	Related Work	290
11.3	System Requirements for SagaLLM	293
11.4	Design with Travel Planning	297
11.5	Experiments	306
11.6	Conclusion	315
	Supplementary Materials	316
	Acknowledgment	317
	References	317

**Chapter 12 ALAS: A Stateful Multi-LLM Agent Framework for Disruption-Aware Planning** **321**

	Abstract	321
--	----------	-----

<b>12.1</b>	Introduction <b>321</b>
<b>12.2</b>	Related Work <b>324</b>
<b>12.3</b>	The ALAS Three-Layer Architecture for Adaptive and Reactive Planning <b>324</b>
<b>12.4</b>	Experimental Evaluation <b>328</b>
<b>12.5</b>	Conclusion <b>335</b>
	Acknowledgment <b>336</b>
<b>12.6</b>	Appendices <b>336</b>
<b>12.A</b>	Appendix A: Supplemental Information for Section 12.3 <b>336</b>
<b>12.B</b>	Appendix B: Agent Factory Implementation Details <b>339</b>
<b>12.C</b>	Appendix C: LCPR Specification and Lemma Proofs <b>343</b>
<b>12.D</b>	Appendix D: Urban Ride Sharing <b>345</b>
<b>12.E</b>	Appendix E: Additional JSSP Results and Analysis <b>352</b>
	References <b>356</b>

## **Chapter 13 CoCoMo: Computational Consciousness Model **361****

	Abstract <b>361</b>
<b>13.1</b>	Introduction <b>361</b>
<b>13.2</b>	Understanding Consciousness <b>364</b>
<b>13.3</b>	Functionalities of Consciousness <b>368</b>
<b>13.4</b>	Computational Consciousness <b>373</b>
<b>13.5</b>	Concluding Remarks <b>384</b>
	Acknowledgments <b>386</b>
	References <b>386</b>

## **Chapter 14 Unified Cognitive Consciousness Theory: Formalization, Validation, and Outlook **391****

	Abstract <b>391</b>
<b>14.1</b>	Introduction <b>391</b>
<b>14.2</b>	Background, Limitations, and Experimental Rationale <b>394</b>
<b>14.3</b>	The UCCT Framework <b>398</b>
<b>14.4</b>	Experiments: Validating UCCT <b>403</b>
<b>14.5</b>	Conclusion <b>414</b>
	Acknowledgments <b>416</b>
<b>14.A</b>	Appendix A: Threshold-Crossing Dynamics Theorem <b>416</b>
<b>14.B</b>	Appendix B: Threshold Calibration <b>419</b>
<b>14.C</b>	Appendix C: Diversity Within Relevance <b>420</b>
<b>14.D</b>	Appendix D: UCCT for RAG (Protocol and Lab Exercise Plan) <b>421</b>
<b>14.E</b>	Appendix E: UCCT for Multi-Agent Debate (Protocol and Lab Exercise Plan) <b>425</b>
	References <b>430</b>

**Chapter 15 A Retrospective and Adaptive Framework to Improve LLMs 433**

- Abstract 433
- 15.1 Introduction 433
- 15.2 Related Work 435
- 15.3 Retrospective and Adaptive Learning 437
- 15.4 UCCT + RAFEL 445
- 15.5 Concluding Remarks 449
- References 450

**Chapter 16 Polynthesis: Discovering Insights Beyond the Known 453**

- Abstract 453
- 16.1 Introduction 453
- 16.2 Phase I: Warm-Up Breadth Probing 456
- 16.3 Phase II: From Breadth to Depth 471
- 16.4 Phase III: Concluding Remarks 488
- 16.5 Conclusion and Future Work 490
- References 493

**Chapter 17 Aphorisms for Collaborative Intelligence 495**

- Abstract 495
- 17.1 Introduction 495
- 17.2 List of 16 Aphorisms 496
- 17.3 Conclusion 518
- References 520

**Epilogue: Two Paradigm Bridges 525**

- Two Parallel Paradigm Shifts 525
- Why the Pair Matters 526
- Priority and Provenance 526
- Acknowledgments 527

**Author's Biography 529**

**Index 531**

# Preface

Today's Artificial Intelligence (AI) systems can answer questions, write poetry, solve math problems, and produce working code, yet they still struggle to correct themselves, stay consistent over time, and distinguish knowledge from guesswork. They dazzle and then disappoint. The question is simple: How can something that looks so smart fail in ways that even children do not?

This pattern maps directly onto what Daniel Kahneman termed *System 1* and *System 2* reasoning. Current Large Language Models (LLMs) excel at System 1: fast, associative, pattern-driven completion. What they lack is System 2: slow, deliberate, goal-directed reasoning with explicit verification. As others have argued, closing this System 2 gap is central to advancing deep learning. We describe the distinction as unconscious pattern repositories versus conscious regulation, or equivalently, System 1 reflexes versus System 2 deliberation. We use these terms as complementary labels; the functional distinction is what truly matters.

This book presents, to our knowledge, a unified, end-to-end architecture for implementing System 2 regulation on top of System 1 pattern libraries. Intelligence emerges when fast pattern recognition is guided by slower, deliberate regulation: memory, critique, goals, and rules that make reasoning explicit and checkable. Future progress will depend less on making one model ever larger and more on building the right regulatory structures around it.

The approach takes advantage of existing strengths of modern models while compensating for their gaps. These models are extraordinary at generating fluent options and recalling diverse patterns yet they are weaker at sticking to a plan over time, resisting plausible but wrong answers, and knowing when to stop. A small, well defined team of roles, each playing a distinct role, turns these strengths into reliable performance.

- *Orchestrator* sets goals and budgets, paces debate, and enforces stop rules.
- *Exec* proposes structured options and plans, with reasons to believe that they might be right.

- *Ground* retrieves sources and invokes tools so claims rest on verifiable evidence.
- *Critic* challenges assumptions and tightens arguments under a clear rubric.
- *Memory* keeps state, commitments, and lessons, using checkpoints and rollback to ensure that progress persists.

In practice, the process begins with the Orchestrator setting the goal and budgets. Then Exec proposes options, Ground attaches evidence, Critic stress-tests and repairs, and Memory records decisions. Finally, the Orchestrator either closes the loop or runs another round.

Behavioral control is as important as information control. A simple *contentiousness* dial governs when the team should challenge assumptions and explore alternatives, and when it should consolidate and decide. Start higher to widen the search; lower it to verify details and converge. Paired with pacing, turn-taking, and context-aware safeguards, this mechanism avoids both premature groupthink and endless argument.

Safety follows from the structure. Checks and balances—the roles that resemble rule-setting, solution-making, and case review—keep behavior on track over time. They complement any training-time alignment and keep the system auditable as it acts.

You will see these ideas developed as reusable patterns that together form the eight foundations of the MACI architecture: (1) Unified Cognitive Consciousness Theory (UCCT) with Computational Consciousness Modeling (CoCoMo) for System-2 control and semantic anchoring; (2) SocraSynth with Entropy and Variation in Conditional Exchanges (EVINCE) for disciplined debate and inquiry; (3) Critical-Reading Inquisitive Template (CRIT) for structured reasoning evaluation; (4) Behavioral Emotion Analysis Model (BEAM) for behavior modulation; (5) DIKE and ERIS for checks and balances in ethical alignment; (6) SagaLLM with Adaptive LLM Agent System (ALAS) for persistent memory and disruption-aware planning; (7) Precision RAG, and (8) polydisciplinary synthesis (Polynthesis) for cross-domain knowledge discovery. Each contributes a distinct regulatory function that elevates System-1 fluency into System-2 reasoning.

The implications could extend beyond the benchmarks. In medicine, regulated AI can remain consistent over months of care and improve with experience. In science, it can sustain long, multi-step inquiries that outlast any single person's attention. In education, it might adapt to how each student learns while remaining responsible for explicit constraints. The deepest gains could come from polydisciplinary synthesis, connecting ideas across fields that rarely interact. Properly

anchored and well-governed, these systems may surface insights that are present in our collective record but are hard to reach directly.

This book reports preliminary implementations and validations with measurable improvements in reasoning and planning quality across initial domains, supported by concrete metrics and ablations. We welcome debate, discussion, and counterexamples; the chapters specify predictions, datasets, and procedures so that results can be independently replicated and challenged.

Civilization advances by preserving knowledge, challenging it, and coordinating many minds. If Artificial General Intelligence (AGI) arrives, it will likely require both System 1 pattern completion and System 2 deliberate regulation working in concert—a community of specialized systems that remember, reason, and self-regulate in service of human goals.

This book maps one route: anchors turn pattern matching into purposeful retrieval and reasoning (Precision RAG with UCCT and CoCoMo); memory converts mistakes into learning (SagaLLM with ALAS); outside checks catch errors that self-assessment misses (DIKE and ERIS, with CRIT for structured evaluation); and structured collaboration turns individual strengths into collective wisdom (SocraSynth with EVINCE, plus Polynthesis). For a concise map of the eight foundations that organize these ideas, see the Introduction.

Edward Y. Chang

Stanford University

March 13, 2024 (first edition) by [SocraSynth.com](#)

March 31, 2025 (second edition, + Eris-Dike, SagaLLM, ALAS chapters) by [SocraSynth.com](#)

June 18, 2025 (third edition, + UCCT chapter) by [SocraSynth.com](#)

August 15, 2025 (fourth edition) by ACM Books



# Acknowledgments

I dedicate this book to my late advisor, Prof. Héctor García-Molina. The SagaLLM component was directly motivated by his seminal 1987 work on *Sagas*, which introduced long-lived transactions with compensation and rollback. His mentorship during my Ph.D. at Stanford—his patience, his insistence that “quality wins out,” and his example of rigorous, clear thinking—shaped both this research program and the way this book was written.

This work has been significantly shaped by several pioneering perspectives in the field. Microsoft CSO Eric Horvitz’s concept of “polydisciplinary” representation in Large Language Models (LLMs) fundamentally influenced my understanding of how these models synthesize knowledge across domains. Demis Hassabis of DeepMind offered insights into end-to-end architecture design, challenging conventional thinking about the role of human intervention. Prof. Stuart Russell’s observation that LLMs mimic human linguistic behavior to achieve goals became central to my approach in modeling behaviors and applying Bayesian reasoning frameworks.

I am especially grateful to my Stanford colleague Prof. Vaughan Pratt, whose feedback on the initial paper inspired the idea of using “contentiousness” to foster adversarial debate—a key mechanism in this framework. I also thank Prof. Monica Lam for her collaborative support throughout this research, as well as my colleagues at Stanford Clinical Mind AI for contributing their invaluable domain expertise in disease diagnosis.

## Acknowledgment of AI Assistance

During the editing of this book, GPT (Generative Pre-trained Transformer) and Claude were used to refine language and improve clarity. These tools served solely as editorial aids. All core content, ideas, and arguments are original and remain my own.



# **Introduction to The MACI Framework: A System 2 Architecture for Artificial General Intelligence**

## **The Intelligence Paradox**

Large Language Models (LLMs) are sophisticated pattern learners rather than fully fledged intelligences. Analogically, they resemble the human unconscious—autonomous processes that regulate heartbeat, breathing, metabolism, sensory processing, and reflexes. Daniel Kahneman termed such automatic processing *System 1*; Yoshua Bengio has contrasted System 1 and System 2 in his keynote addresses. LLMs excel at System 1 tasks: rapid pattern recognition, learned associations, and fluent but unvalidated responses.

Human cognition also requires *System 2*: deliberate reasoning, error correction, goal-directed planning, and metacognitive oversight. System 2 transforms reactive pattern use into regulated intelligence by validating outputs, maintaining coherent goals, and adaptively correcting errors. This regulatory architecture is missing in current standalone LLM systems.

## **From Pattern Matching to Artificial General Intelligence**

Artificial General Intelligence (AGI) refers to systems that understand, learn, and apply knowledge across diverse domains at human or superhuman levels, including abstract reasoning, contextual adaptation, and creative problem-solving. Current LLMs, regardless of scale, lack the regulatory architecture that turns pattern repositories into coherent intelligence.

This book introduces Multi-LLM Agent Collaborative Intelligence (MACI)—a System 2 architecture built above System 1 LLMs (instantiated through specialized agent roles as described in the [Preface](#)). MACI implements eight foundational pillars: (1) UCCT with CoCoMo for System 2 control and semantic anchoring, (2) SocraSynth with EVINCE for disciplined debate and inquiry, (3) CRIT for Socratic

reasoning and evaluation, (4) BEAM for behavior modulation, (5) DIKE–ERIS for ethical alignment through checks and balances, (6) SagaLLM with ALAS for persistence and planning, (7) Precision RAG, and (8) Polynthesis for knowledge discovery. Orchestrated across informational and behavioral dimensions, these pillars regulate LLM pattern use into goal-directed intelligence. Through implementations and empirical evidence, we show how structured multi-agent collaboration, rather than model scaling alone, creates a pathway toward AGI.

## **Eight Foundational Pillars**

### **(1) UCCT with CoCoMo as a Control Layer for System 2**

*Unified Cognitive Consciousness Theory* (UCCT) with Computational Consciousness Modeling (CoCoMo) establishes the regulatory control layer that turns System 1 pattern repositories into System 2 directed intelligence. UCCT anchors the meanings of LLM patterns via Few Shot examples, Retrieval-Augmented Generation (RAG), and multi-LLM debate. CoCoMo specifies how the orchestrator sets goals, allocates computational budgets, manages schedules, and defines escalation or abstention pathways.

### **(2) SocraSynth with EVINCE for Disciplined Debate and Inquiry**

SocraSynth structures adversarial and cooperative dialogue through turn-taking, role assignment, and termination rules to ensure perspective discovery and consensus building. Entropy and Variation in Conditional Exchanges (EVINCE) instruments the dialogue with information-theoretic signals (such as entropy, mutual information, Jensen–Shannon divergence, and Wasserstein distance) to detect tightening versus drift and to trigger termination or retrieval. This process enables precision RAG by locating the exact knowledge gaps.

### **(3) CRIT for Socratic Reasoning Evaluation**

Critical Reading Inquisitive Template (CRIT) provides a structured rubric and targeted probes for evaluating claims, evidence, counterpoints, and uncertainty. It measures argument quality by assessing logical coherence, evidential support, acknowledgment of limitations, and handling of counterarguments, thereby enabling targeted repairs and quality grading.

### **(4) BEAM for Behavior Modulation**

Behavioral Emotion Analysis Model (BEAM) offers dual-dial control over information and behavior. It modulates debate contentiousness to shift between explorative exchange (high contentiousness with challenges and counterfactuals) and conciliatory synthesis (low contentiousness emphasizing verification and convergence). BEAM calibrates tone, stance, rhetorical strategy, and certainty, while flagging overconfidence or dismissiveness for oversight.

**(5) DIKE–ERIS for Checks and Balances in Ethical Alignment**

For auditable and adaptable alignment, DIKE–ERIS separates legislative and judicial functions. DIKE maintains policy artifacts and normative constraints, while ERIS conducts case-by-case contextual review. Unlike RLHF (Reinforcement Learning from Human Feedback)-style parameter updates that risk degrading capabilities, DIKE–ERIS leaves base model parameters untouched and uses BEAM to adjust language behavior without altering underlying representations.

**(6) SagaLLM with ALAS for Persistence and Planning**

Long-horizon work needs reliable state management and adaptive planning. SagaLLM provides durable, auditable memory through checkpoints, rollbacks, and compensating actions. Adaptive LLM Agent System (ALAS) adds disruption-aware planning with validator and repair loops, typed input–output contracts, and execution under disturbance. Together, they support continuity, recovery, and goal-directed planning across extended horizons.

**(7) Precision RAG**

SocraSynth, CRIT, and UCCT guide retrieval so that context raises relevant pattern density and strengthens inference and reasoning.

**(8) Polynthesis for Knowledge Discovery**

We introduce Polynthesis, the synthesis of knowledge across multiple or all domains, as AGI’s mechanism for discovering unknown unknowns: insights that exist in no single domain but emerge only through cross-domain integration.

**From System 1 Foundation to Cognitive AGI****Rehabilitating System 1**

Critics often dismiss LLMs for their weak reasoning, planning, and grounding. This is like dismissing the human unconscious as a useless repository of patterns. The unconscious is the substrate on which deliberate reasoning is built. Without learned patterns and automaticity, System 2 has nothing to operate on.

Similarly, LLMs are not obstacles to AGI but its necessary foundations. Their statistical regularities, relational encodings, and fluent language provide the library that System 2 regulation requires. The limitation lies in deploying LLMs without the regulatory architecture that turns pattern use into directed intelligence.

**Necessity and Sufficiency of the Eight Foundational Pillars**

To approach cognitive AGI—systems that understand, reason, plan, and learn across domains—several core capabilities are required. We map MACI’s pillars to these foundational needs:

- *Semantic understanding:* UCCT with CoCoMo anchors patterns to task-relevant meanings via Few Shot learning, RAG, and debate, replacing free-floating associations with grounded semantics.
- *Reasoning and inquiry:* SocraSynth with EVINCE structures disciplined debate, while CRIT evaluates argument quality to ensure rigorous multi-perspective reasoning.
- *Behavioral adaptation:* BEAM modulates information exchange and stance, enabling shifts between exploration and exploitation, as well as divergence and convergence.
- *Ethical alignment:* DIKE–ERIS separates policy maintenance from contextual interpretation, providing stable yet adaptable alignment without catastrophic forgetting.
- *Persistent memory and planning:* SagaLLM with ALAS supplies durable state, transactional guarantees, hierarchical planning, and error recovery across long horizons.
- *Meta-learning and knowledge discovery:* When debate yields low confidence, EVINCE signals knowledge gaps and triggers precision RAG; Polynthesis supports the discovery of novel connections.

These pillars address gaps that System 1 LLMs cannot overcome in isolation. Together, they provide a practical foundation for cognitive AGI in language understanding, multi-domain reasoning, long-horizon planning, ethical decision-making, and knowledge synthesis. Our focus is on cognitive AGI rather than embodied AGI, although perception and actuation can be integrated as needed.

## **Book Organization**

The book progresses from theory to mechanisms and then to empirical evidence.

### **Theoretical Foundations (Chapters 1–4)**

Chapters 1–3 describe why scale alone is insufficient, introduce the System 1 and System 2 framing, cover similarity-based reasoning, introduce prompt engineering, and explain why regulation is required for reliability.

Chapter 4 presents the full UCCT account and the anchoring-strength heuristic, which relates anchoring strength  $S$  to reasoning quality.

### **Core Operational Frameworks (Chapters 5–12)**

Chapters 5–7: SocraSynth with EVINCE and CRIT for disciplined debate, measurement, and inquiry.

Chapters 8–10: BEAM together with DIKE-ERIS for behavior modulation and ethical oversight.

Chapters 11–12: SagaLLM with ALAS for persistent memory and disruption-aware planning with transactional guarantees.

### **Applications and Extensions (Chapters 13–17)**

Chapter 13: CoCoMo as a control specification that instantiates UCCT for the orchestrator.

Chapter 14: formalization and validation of the anchoring-strength heuristic with task-level tests.

Chapter 15: retrospective and adaptive updates that help the team learn across episodes.

Chapter 16: Polynthesis: case studies.

Chapter 17: aphorisms that distill MACI into portable practices.

This structure keeps the argument readable by presenting a control layer that turns fast pattern libraries into regulated reasoning, mechanisms that make the control concrete, and experiments that test whether the architecture improves accuracy, calibration, convergence, and stability.

### **Toward Cognitive AGI**

The path to AGI will depend less on larger models and more on architectural regulation. MACI shows that System 2 capabilities—reasoning, planning, memory, alignment, and knowledge discovery—emerge from structured collaboration among regulated agents.

For decades, the field has prioritized monolithic scaling. This approach has delivered strong pattern learners but not reliable reasoning, memory, or alignment. The limitation is architectural. Intelligence requires regulation.

This book presents theory and working implementations. The frameworks are built and tested across medical diagnosis, scientific reasoning, ethical decision-making, and knowledge synthesis. When pattern repositories are properly anchored, debated, evaluated, modulated, aligned, and coordinated, they exhibit hallmarks of intelligent behavior: coherent cross-domain reasoning, adaptive planning under uncertainty, reliable long-horizon memory, and discovery of novel connections.

The eight foundational pillars provide a technical architecture and a conceptual shift—from monolithic oracles to collaborative systems where regulation turns raw capability into directed intelligence. Each pillar closes a necessary gap. Together, they form a practical basis for cognitive AGI.

We invite researchers, practitioners, and critics to test these frameworks and extend this work. Build on the eight pillars. Challenge the assumptions. Push the boundaries of what regulated multi-agent systems can achieve.

The era of collaborative intelligence has begun. Join us.

# CHAPTER 1

---

## A Brief History of AI: From Turing to Transformers

### Abstract

This chapter reinterprets the history of AI by focusing on the evolution of similarity measurement—from rule-based to context-aware models—and emphasizing its critical role in AI's core functions like learning and problem-solving. It explores the impact of detailed and evolving understandings of similarity in linguistics (text) and computer vision (image), projecting a future where AI merges advanced data analysis with abstract reasoning. The chapter provides an in-depth analysis from the perspectives of linguistics, computer science, and cognitive psychology and neuroscience, illustrating how the progression of similarity concepts continues to fuel AI's advancement.

### 1.1

#### Introduction

Artificial Intelligence (AI) has journeyed through a fascinating historical trajectory, marked by five pivotal epochs, each representing a significant paradigm shift triggered by major technological advancements. These epochs are as follows: *Initiation*, which set the stage for AI's emergence with the foundational concepts and milestones of AI; *Expert System Encoding Human Knowledge*, which focused on rule-based AI systems that encoded and applied human expertise; *Heuristic-Based Modeling*, which emphasized the development and use of heuristic methods for AI problem-solving; *Learning Model from Data*, which marked the transition to algorithms that learn and adapt from data, signifying the emergence of machine learning; and *Context-Based Semantic Disambiguation*, which reflected AI's evolving proficiency in understanding and interpreting context, thereby improving semantic accuracy.

Although numerous comprehensive sources, such as Wikipedia, provide detailed accounts of AI's evolution through various lenses: language, computation,

philosophy, cognitive psychology, neuroscience, and application, this chapter takes a different approach. It zeros in on a fundamental aspect: **similarity**.

When we evaluate the intelligence of machines, we often focus on attributes such as learning capacity, pattern recognition, predictive accuracy, robustness, adaptability, generalization, reasoning, problem solving, and decision making abilities. These qualities collectively define the prowess of AI systems. Among these traits, the concept of similarity plays a pivotal role. For example, in learning, an effective similarity measure is fundamental for recognizing patterns and generalizing knowledge. In terms of adaptability, the ability to detect similarities with previous experiences allows AI to adjust to new or evolving circumstances. Regarding robustness, employing similarity measures helps AI differentiate between normal and anomalous patterns, thereby increasing system resilience. Furthermore, in the realm of problem-solving, the capacity to identify similarities with previously encountered situations can enable AI to apply existing solutions to new problems, and thus improving its efficacy in addressing challenges. This chapter explores the vital function of similarity across the broad spectrum of AI capabilities, underlining its significant contribution to the field's foundational operations.

In the realm of tangible objects, similarity measures are integral to various vision-related tasks, aiding in the recognition of patterns, shapes, and colors, which are essential for object recognition and image classification. In text analysis, these measures are crucial for identifying similarities in content, helping to detect plagiarism, retrieve documents, and translate languages. In the auditory domain, the analysis of the similarity of sound wave patterns or musical notes is key to genre classification and music recommendation systems. In medical imaging, these measures facilitate the diagnosis of diseases by comparing patient images with known cases, allowing accurate identification and classification of medical conditions. Object feature comparison is foundational in robotics and surveillance for recognizing and interacting with physical entities. Similarly, facial and voice recognition systems rely on analyzing patterns to identify or verify individuals, resulting in enhancing security and personal authentication. In e-commerce, analyzing similarities in product attributes and user preferences guides recommendation systems, enhancing the user experience by suggesting related or complementary products.

In the abstract realm, similarity measures are crucial for discerning semantic relationships, aiding in knowledge representation, ontology mapping, and refining AI's interpretive faculties. Environmental studies leverage these assessments for climate modeling and ecological research. Sentiment analysis, whether across social media or customer feedback, utilizes similarity assessments to extract

insights into public sentiment or consumer behavior. These measures also underpin AI's problem-solving prowess in complex scenarios by informing strategy formulation. Behavioral analysis, whether in psychology or marketing, employs similarity comparisons to decode human actions and preferences. In the legal domain, case similarity aids in judicial decision-making and legal scholarship. Language translation harnesses similarities in linguistic structures to break down language barriers. Furthermore, in creative writing, analyzing thematic or stylistic similarities assists in authorship identification, genre categorization, and literary exploration.

The advancement in similarity research, while appearing gradual, reflects not only human ingenuity but also the limitations imposed by computational resources and hardware capabilities. The quest to quantify similarity covers a broad spectrum of abstractions, from sensory inputs like visual, auditory, olfactory, and tactile data to complex abstract concepts such as ideas and semantics. Hardware improvements have enabled researchers to explore more advanced methods that encompass both concrete and abstract forms of similarity. This progression marks the field's growth in harmonizing detailed sensory data analysis with a deeper understanding of abstract concepts, utilizing computational advancements and diverse data interpretations.

The following sections provide a deeper dive into key AI terminology and the development of similarity measures from two distinct perspectives: scientific disciplines and historical evolution. The disciplinary perspective covers three core fields: linguistics, computer science, and cognitive psychology and /neuroscience. The evolutionary perspective traces the historical journey of similarity measurement through four distinct eras: rule-based, model-based, data-centric, and context-aware.

Providing two perspectives on similarity measurements, across different scientific disciplines and through the historical evolution of AI methodologies, offers a comprehensive understanding that caters to a broader audience with varied interests and backgrounds. Here are some reasons why this dual perspective is valuable:

*Disciplinary Insight:* Examining similarity measurements across different scientific disciplines enriches our understanding by highlighting how various fields approach and apply the concept of similarity. This can foster interdisciplinary collaboration and innovation, as techniques from one field can inspire new approaches in another.

*Historical Context:* Exploring how similarity measurement has evolved within AI provides historical context, and shows how methodologies have progressed from

rule-based to more advanced context-aware systems. This perspective helps readers appreciate the advancements in AI and understand why certain methods were developed or abandoned.

## 1.2 Definitions

We define and scope key terms and concepts to prepare for the subsequent discussion.

### 1.2.1 Rudimentary Terms

*Data:* Raw information used to train AI models. Data quality significantly impacts model performance.

*Algorithm:* A set of instructions that a computer follows to perform a specific task. AI algorithms are often complex and involve statistical methods.

*Model:* A representation of the knowledge learned from data that allows the AI system to make predictions or decisions.

### 1.2.2 General Terms

*Artificial Intelligence (AI):* The broader concept of machines being able to carry out tasks in ways that we would consider smart.

*Explainable AI:* AI systems that offer transparency and clear understanding of their operations and decision-making processes.

*General AI:* Also known as Artificial General Intelligence (AGI), this refers to a type of AI that has the ability to understand, learn, and apply knowledge across a wide range of tasks, much like a human being. It's an AI system with generalized human cognitive abilities, meaning that when presented with an unfamiliar task, it can find a solution without human intervention. AGI would be able to reason, solve problems, make judgments, plan, learn, and communicate in natural language, among other capabilities. However, as of now, AGI remains a theoretical concept and has not yet to be realized in practical applications.

*Narrow AI:* The type of AI that we encounter in our daily lives and is currently in use around the world. It is designed to perform a narrow task (e.g., facial recognition, internet searches, driving a car) and is trained for a specific dataset or a set of tasks. Narrow AI operates under a limited pre-defined range or context, often focusing on executing a single task extremely well or carrying out a limited range of tasks in a specific domain. It lacks the general cognitive abilities of AGI and cannot apply its knowledge beyond its specific field or task.

*Machine Learning (ML):* A subset of AI that includes statistical techniques that enable machines to improve at tasks through experience.

*Deep Learning:* A subset of machine learning that uses neural networks with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to “learn” from large amounts of data.

*Neural Networks:* Computational models that are somewhat inspired by the structure of the human brain, enabling computers to recognize patterns and solve common problems in AI, such as classification, prediction, and decision-making.

*Supervised Learning:* A type of machine learning where the model is provided with labeled training data and the desired output. The goal is to learn a mapping from inputs to outputs.

*Unsupervised Learning:* A type of machine learning where the model is not provided with labeled data and must find the structure in its input on its own.

*Reinforcement Learning:* An area of machine learning where an agent learns to behave in an environment by performing actions and seeing the results, focusing on long-term rewards. An example is an AI agent learning to play a game through trial and error, receiving rewards for winning.

*Natural Language Processing (NLP):* A field of AI that gives machines the ability to read, understand, and derive meaning from human languages.

*Computer Vision:* A field of AI that trains computers to interpret and understand the visual world, extracting information from images and videos.

*Robotics:* The branch of technology that deals with the design, construction, operation, and application of robots, often incorporating AI systems to enhance autonomy and adaptability.

*Large Language Model (LLM):* LLMs are advanced AI systems trained on extensive datasets, initially text-centric and now increasingly incorporating multimodal data. They are designed to comprehend, generate, and interact with human language, imagery, and video with a level of sophistication that closely mirrors human cognitive processes.

### 1.2.3 Performance Terms

*Algorithmic Bias:* Algorithmic bias refers to the potential for algorithms to reflect, perpetuate, or amplify biases present in the training data or as a result of the design of the algorithms themselves. This can lead to skewed or unfair outcomes, particularly in decision-making processes.

*Hallucination:* In the context of AI, hallucination refers to the phenomenon where a model generates or outputs information that is ungrounded, misleading, or not supported by the input data. This is commonly seen in language models where the generated text may be plausible but not factually accurate or relevant to the context.

*Generalization:* Generalization is the ability of an AI model to perform well on new, unseen data that was not part of the training set. It indicates the model's capacity to apply learned knowledge to different situations, a key indicator of its robustness and practical utility.

*Overfitting:* Overfitting occurs when an AI model learns the specific details, patterns, and noise in the training data to the extent too precisely that it negatively impacts the model's performance on new data. This usually happens when the model is overly complex, capturing patterns that do not generalize to unseen data.

## 1.3

### Perspectives on Similarity

This section presents the foundational theories of similarity measurement from three distinct domains: *linguistics*, *computer science*, and *cognitive psychology and neuroscience*. The subsequent historical section will clarify how these foundational theories have influenced, and been incorporated into, specific technological advancements and methodologies across various eras. Cross-references will be provided to ensure coherence and to emphasize the interconnection of these perspectives.

#### 1.3.1 Linguistic Perspective

The study of similarity in linguistics has been profoundly influenced by Zellig Harris's pioneering work. His 1954 study introduced the idea that the distributional properties of words and their contextual usage could unlock the secrets of language comprehension, highlighting the indispensable role of context in understanding meaning [Harris 1954]. This principle, that words found in similar contexts tend to share meanings, laid the foundation for distributional semantics and resonates with John R. Firth's famous insight that "A word is known by the company it keeps." This linguistic perspective sets the stage for further exploration of how context and distributional properties have been instrumental in shaping our understanding of semantic similarity, paving the way for subsequent advancements in the field.

The evolution of linguistic theories continued into the later part of the 20th century with the rise of cognitive linguistics, which examines the interplay between linguistic structures and human cognitive processes. This approach underscored how

language reflects our perception and conceptualization of the world, introducing a multi-layered perspective on semantic abstraction.

A significant milestone in bridging linguistic theory with practical applications was the development of WordNet in the 1980s by a team at Princeton University [Miller 1995]. This lexical database, which organizes English words into sets of cognitive synonyms, or *synsets*, has profoundly influenced areas such as word sense disambiguation, information retrieval, and beyond, highlighting the importance of structured semantic relationships in understanding language.

Moreover, the influence of linguistic insights extended into the domain of computer vision with the creation of ImageNet by Fei-Fei Li and colleagues [Deng et al. 2009], which drew upon the principles underlying WordNet to categorize visual content. This convergence of linguistics and computer science has been further propelled by advancements in computational methods, with techniques such as Latent Semantic Analysis (LSA) [Dumais 2004], Latent Dirichlet Allocation (LDA) [Blei et al. 2003], and innovative word embeddings such as Word2Vec [Mikolov et al. 2013] and GloVe [Pennington et al. 2014]. These methodologies have enabled the conceptualization of word meanings in high-dimensional vector spaces, illuminating the intricate web of semantic relationships through patterns of co-occurrence and contextual analysis.

The introduction of the transformer model [Vaswani et al. 2017] and the subsequent unveiling of BERT [Devlin et al. 2019], which employs self-supervised learning to predict masked words within a context, along with the release of GPT, designed to predict the next word based on context, heralded a new epoch in our endeavor to unravel context-dependent semantics. This development fulfills the vision proposed by Zellig Harris in his groundbreaking 1954 work, now actualized in contemporary computational models.

### 1.3.2 Computer Science Perspective

In computer science, the concept of similarity has evolved from simple rule-based models to complex vector-space and probabilistic models, reflecting the field's progression in addressing various computational challenges.

#### 1.3.2.1 Rule-Based

A rule-based AI model, also known as an expert system, employs a collection of predefined if-then statements to execute decisions or solve problems. These conditional statements are crafted from the expertise of specialists in a particular field. The system applies these rules to the input data to formulate conclusions.

The “if” segment of a statement evaluates the data for specific conditions or patterns. When these conditions are satisfied, the “then” segment is activated,

performing a designated action or drawing a conclusion. Importantly, these systems do not adapt or learn from data in the manner that machine learning models do. Rather, they rely on a set of explicit rules, which are the codified versions of expert knowledge within a specific domain. This knowledge is methodically organized and stored in a knowledge base, enabling the system to reference and apply it efficiently during its operations.

In Section 1.4.1, we will explore the technical details and applications of rule-based systems, emphasizing their pivotal role during the rule-based era of AI's evolution.

### 1.3.2.2 Vector-Space

The vector-space model marked a significant shift by representing objects and features as vectors within a high-dimensional space. This approach facilitated the development of various distance functions to assess similarity across different applications. Notably, a comprehensive survey by [Cha 2007] categorized 45 distance functions into families such as inner product,  $L_1$ , Minkowski, and intersection, each with its representative functions highlighting the versatility in vector-space analysis.

**Inner Product, Dot Product and Cosine** The inner product and dot product are the same in the context of Euclidean space and are defined for vectors  $\mathbf{a}$  and  $\mathbf{b}$  as:

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \dots + a_n b_n.$$

This operation results in a scalar value and indicates the vectors' magnitude and directionality.

Cosine similarity is a measure that calculates the cosine of the angle between two vectors. It is defined as the dot product of the vectors normalized by the product of their magnitudes:

$$\text{cosine similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|},$$

where  $\|\mathbf{a}\|$  and  $\|\mathbf{b}\|$  represent the Euclidean norms of the vectors  $\mathbf{a}$  and  $\mathbf{b}$ , respectively.

The cosine similarity is especially useful in contexts where the magnitude of the vectors is not of primary concern, making it ideal for applications in high-dimensional spaces like text analysis and information retrieval.

**Weighted Minkowski** The weighted Minkowski distance function allows assigning varying importance to different dimensions, accommodating the significance of specific features in contexts like machine learning and data mining:

The weighted Minkowski distance between points  $X = (x_1, x_2, \dots, x_n)$  and  $Y = (y_1, y_2, \dots, y_n)$  with a set of weights  $W = (w_1, w_2, \dots, w_n)$  is defined as:

$$D(X, Y) = \left( \sum_{i=1}^n w_i |x_i - y_i|^p \right)^{\frac{1}{p}},$$

where  $p$  is the order parameter of the Minkowski distance. When  $p = 1$ , it becomes the weighted Manhattan distance, and when  $p = 2$ , it becomes the weighted Euclidean distance.

**Set Similarity** Moreover, the Jaccard similarity [Jaccard 1912] provides a robust method for comparing sets, which is especially beneficial in scenarios where the presence or absence of features is more critical than their magnitude, as seen in plagiarism or copyright detection.

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|},$$

where

- $|A \cap B|$  is the number of elements in the intersection of sets  $A$  and  $B$ .
- $|A \cup B|$  is the number of elements in the union of sets  $A$  and  $B$ .

### 1.3.2.3 Probabilistic-Based

The advancement into probabilistic-based models introduced a spectrum of statistical and probabilistic distance functions, offering refined tools for quantifying similarity or dissimilarity based on underlying probabilistic principles. These functions—including Pearson Correlation Coefficient, Mahalanobis Distance, Kullback-Leibler Divergence, and others—cater to diverse analytical needs, enriching the computational toolkit available for similarity assessment across various domains.

This section underscores the computer science perspective on similarity, detailing its journey from rule-based logic to advanced probabilistic models, reflecting the field's dynamic evolution and its pivotal role in shaping contemporary approaches to measuring similarity.

### 1.3.3 Cognitive Psychology Perspective

Cognitive psychology and neuroscience provide deep insights into how similarity is perceived and processed at a neural level, significantly influencing the development of AI technologies. Anne Treisman's Feature Integration Theory (FIT) [Treisman and Gelade 1980] has been instrumental in understanding how the



**Figure 1.1** “Which pairs are similar?” (DALL-E).

brain synthesizes various sensory features into cohesive percepts—a concept that has parallels in how artificial neural networks, particularly Convolutional Neural Networks (CNNs) [LeCun et al. 1989, 1998], process visual information (Figure 1.1).

FIT draws heavily from the principles of Gestalt psychology [Wertheimer 1912, Koffka 1935], which propose that perception organizes individual components into a meaningful whole. This aligns with FIT’s view that perception is an integrated experience shaped by the brain’s organizational tendencies. The theory also intersects with selective attention, as illustrated by Donald Broadbent’s Filter Model [Broadbent 1958]. This model suggests that attention acts as a filter, selecting relevant information for further processing. Broadbent’s framework complements FIT by emphasizing attention’s role in integrating features into a unified perception, thereby highlighting the brain’s selective processes.

In 2001, while conducting a study on perceptual similarity with my PhD student Beita Li, we uncovered that images could demonstrate similarity in various dimensions. Although the weighted-Minkowski function could learn feature weights, its application was universal once the weights were set, representing a statistical average. Our experiments with transformed images—through translation, cropping, rotation, down-sampling, and affine scaling—revealed that while these images were perceptually similar to their originals, their similarities were in distinct aspects. This observation led to the development of our “Dynamic Partial Function” (DPF) in 2002 [Li et al. 2002, 2003]. The DPF signature for each image pair could be unique. Essentially, if a pair of images (or objects) demonstrates a sufficient number of similar features, they are likely deemed similar, regardless of the specific features. For instance, an image is considered similar to its rotated version due to the similarity of their color histograms. Similarly, an image and its cropped version are considered alike based on their texture features. If two images exhibit a sufficient degree of similarity in various respects—typically 60%—they are generally regarded as similar.

While surveying the literature, we came across the article “Respects for Similarity” by Medin et al. [1993], which portrays similarity as a dynamic process of formulating a function and identifying relevant aspects, a process that is realized consciously. To clarify this concept, let’s refer to an example from Li et al. [2003]:

Consider the task of identifying two places similar to England. Scotland and New England might emerge as viable candidates. Yet, the criteria that make England similar to Scotland are distinct from those link England to New England. Using the attributes that align England with Scotland to assess the similarity between England and New England might not yield a parallel conclusion, and the reverse is also true. This scenario underscores the idea that objects can be similar to a reference object in different respects. A fixed similarity function, bound to a specific set of criteria, fails to capture the similarities across different contexts. Medin et al. [1993] examine the operational dynamics of similarity in human cognition, noting that the selection of relevant attributes is crucial, with similarity being as much a result as a driving force of conceptual coherence. Goldstone [1994] further elucidates that similarity involves identifying the appropriate criteria for comparison, which occurs *only after the objects in question have been juxtaposed*, not beforehand. The criteria selected for this comparison are activated during the comparison process, with a tendency to favor those that enhance the coherence between the objects being compared.

Although the Dynamic Partial Function (DPF) introduces additional computational complexity, it has indirectly played a role in the success of AlexNet [Krizhevsky et al. 2012] by influencing data augmentation strategies. By integrating

transformed images into its training dataset, AlexNet benefits from a principle akin to DPF, thereby improving its accuracy and robustness in recognition tasks. The recent advancements in transformer algorithms [Vaswani et al. 2017], which focus on dynamism and context-awareness, build on this foundation—a topic that will be explored in detail in the subsequent section.

### 1.3.3.1 Neuroscience

The neuroscience foundation of FIT and its relation to visual feature processing are echoed in the development of CNNs, which were inspired by the hierarchical structure and feature detection capabilities of the human visual cortex, as explored by Hubel and Wiesel [1962]. These networks use convolutional layers to automatically and adaptively learn spatial hierarchies of features from visual data, akin to the neural processing observed in the brain.

Techniques such as Multivariate Pattern Analysis (MVPA) [Norman et al. 2006] and Neural Decoding [Haynes and Rees 2006] further bridge the gap between neuroscience and AI by offering methods to analyze how information is represented across neural populations and how these representations can predict perceptual experiences or cognitive states. These methodologies have inspired and informed the design of advanced AI systems, particularly in how they encode, process, and differentiate complex patterns and similarities.

The cross-pollination between neuroscience and AI—exemplified by the influence of neural processing principles on the design of CNN—highlights the symbiotic relationship between these two fields. Insights gained from studying the brain's processing mechanisms have catalyzed innovations in AI, leading to more effective and biologically inspired computational models. This interdisciplinary exchange not only propels forward our understanding of neural processes but also fosters the development of AI systems that more closely mimic human perceptual and cognitive capabilities.

### 1.3.4 Section Remarks

The exploration of similarity measurement spans across linguistics, computer science, and cognitive psychology and neuroscience, revealing its multidisciplinary nature. Each field offers a unique lens to view similarity: from the contextual information in language, computational algorithms in AI, to the neural processing in the human brain. They converge on the common ground of representing entities in high-dimensional spaces and employing distance metrics for quantification, highlighting the universal applicability of similarity. This convergence fosters a rich dialogue between disciplines, enhancing our understanding and ability to quantify and interpret similarity, driving forward innovation and providing new

methodologies that influence a wide array of contexts in our quest to decode this fundamental concept.

## 1.4

### Eras of Similarity Measurement

Traversing through the history of AI and similarity measurement, one can delineate distinct eras, each marked by unique methodologies and technological advancements. In contrast to the previous section, which examined similarity measurements across different scientific disciplines, this section chronicles these eras—starting from the *rule-based* era, which laid the foundational stones, through the evolution into *model-based*, *data-centric*, and *context-aware* methodologies—illustrating the dynamic trajectory of similarity measurement in AI. As we reach the conclusion of this section, we explore the prospects of the forthcoming era, which promises to challenge and expand our understanding by venturing into the realm of discovering the *unknown unknowns*.

#### 1.4.1 Rule-Based Era (1950s–)

The rule-based era of the 1950s marked the inception of AI, characterized by the use of symbolic representations and formal logic to analyze similarity. This period saw the emergence of explicit symbolic representations and logic-based methods tailored for similarity assessment. Innovations by Allen Newell and Herbert A. Simon—with tools like the Logic Theorist and General Problem Solver [Newell and Simon 1956]—pioneered logical rule-based problem-solving, setting a pivotal foundation for AI's evolution.

In the following decades, systems such as DENDRAL [Lindsay et al. 1979] utilized rule-based logic to deduce molecular structures from data, while MYCIN [Shortliffe et al. 1975], an expert system for diagnosing infections and recommending treatments, demonstrated the practical application of rule-based reasoning in the field of medical diagnostics.

Despite their effectiveness in well-defined scenarios, rule-based systems face limitations in more complex or changing environments. However, their clarity and systematic nature remain invaluable in certain applied domains, for example:

- (1) *Customer Service*: Rule-based chatbots are prevalent in customer service, using predefined rules to respond to user inquiries based on detected keywords or phrases in user input, providing immediate and consistent customer support and are particularly effective for handling routine or frequently asked questions.
- (2) *Fraud Detection Systems*: The finance sector employs rule-based systems to identify potentially fraudulent transactions by comparing against specific

criteria, such as unusual transaction amounts or activity in atypical locations.

- (3) *Equipment Failure Diagnosis:* In industrial settings, rule-based systems analyze data to pinpoint the causes of equipment failures, leveraging historical data and expert knowledge to predict and help prevent future breakdowns.

This era also gave rise to significant tools such as PROLOG [[Clocksin and Mellish 1981](#)], associated with logic programming and structured problem-solving, and decision trees [[Quinlan 1986](#)], which visually represent decision-making processes, demonstrating rule-based logic in action.

While rule-based systems initially approached similarity with a clear, logical framework, subsequent advancements in AI have embraced more flexible methods like statistical models and machine learning, offering a broader, more adaptable approach to understanding similarity across various contexts.

Rule-based systems contrast with the “black-box” nature of current CNNs and LLMs, particularly in terms of interpretability and decision-making processes. Rule-based systems are transparent in how decisions are made, as they follow a clear set of if-then rules or logic for inference, allowing users to understand and trace the reasoning behind each decision.

On the other hand, CNNs and LLMs—particularly those based on deep learning—often operate as black boxes, where the internal decision-making processes are not easily interpretable. In these systems, decisions result from complex, non-linear interactions among thousands or even millions of parameters that have been adjusted through the learning process. While these models are powerful and effective in handling a wide range of tasks—especially those involving large datasets and requiring pattern recognition beyond human capabilities—their inner workings are not as transparent or interpretable as rule-based systems.

### **1.4.2 Model-Based Era (1970s–)**

In this era, vector-space and probabilistic models were designed to quantify similarity.

#### **1.4.2.1 Vector Space Models**

The vector-space era marked a shift in similarity measurement, moving from rigid rule-based to representation-based approaches. In this era, objects, documents, and features began to be conceptualized as vectors in high-dimensional spaces, fostering a more intuitive and flexible method for assessing similarity.

***The Vector-Space Model and Information Retrieval*** Central to this era was the vector-space model, which represents documents as vectors of term frequencies,

enabling the computation of document similarity using measures such as cosine similarity between their respective vectors. This model enhanced the efficiency and effectiveness of information retrieval systems.

**Distance Functions and Feature Weighting** A diverse array of distance functions emerged during this era to quantify the similarity between vectors. The Minkowski distance, for instance, generalized traditional metrics like the Euclidean and Manhattan distances, offering flexibility in adjusting the sensitivity to differences in vector components. Weighted distance measures also gained prominence, recognizing that not all features have equal importance in similarity assessment. The weighted Minkowski distance, in particular, allowed for differential weighting of dimensions based on their relevance to the specific application at hand.

**Beyond Textual Data** The utility of the vector-space model extended well beyond textual data. In the realm of image processing, features (e.g., colors, textures, and shapes) extracted from images were represented as vectors, enabling the assessment of image similarity based on the distances between these vectors. This paradigm facilitated significant advancements in image retrieval, classification, and clustering.

**Dimensionality Reduction Techniques** To address the challenges posed by high-dimensional data, techniques like Principal Component Analysis (PCA) [Jolliffe 1986] and LSA [Landauer and Laham 1998] were developed. These methods reduced the dimensionality of data while preserving its essential structure, enhancing computational efficiency and mitigating the “curse of dimensionality.” Manifold learning, a non-linear dimensionality reduction approach, further expanded the toolbox for tackling high-dimensional data [Tenenbaum et al. 2000]. For a comprehensive overview of these techniques, refer to Maaten et al. [2008].

The vector-space era laid the groundwork for advancements in machine learning and data mining, making similarity measures essential for clustering, classification, and recommendation systems. Representing data as vectors allowed for the exploration of relationships across varied data types through the nearest neighbor concept. In this context, the characteristics or labels of an unknown instance’s  $k$ -nearest neighbors could be inferred and applied to the instance, with these neighbors determined by distance metrics.

However, vector representations often result in sparsity, potentially leading to resource inefficiency and decreased accuracy. These models, while capturing syntactic relationships, sometimes struggle with semantic depth, such as identifying synonyms or contextual meaning. The assumption of feature independence and the use of linear methods in dimensionality reduction can also lead to inaccuracies, particularly with non-linear data structures. The introduction of Support

Vector Machines (SVMs) [Cortes and Vapnik 1995], which utilize kernel methods, addressed some challenges related to non-linear data but increased greater computational complexity. SVMs remained a significant focus in the field until the rise of deep learning architectures like AlexNet marked a shift toward the data-centric era.

#### 1.4.2.2 Probabilistic Models

Probabilistic models offer more flexibility than vector-space models because they can incorporate uncertainty and variability directly into their mathematical frameworks, allowing for a more comprehensive and adaptive representation of data.

**Statistical Inference and Similarity** Probabilistic models introduced the concept of statistical inference, where the likelihood of data or feature occurrences was used to estimate similarity. This approach allowed for effective handling of uncertainty and variability in data, making it particularly useful when working with noisy or incomplete datasets.

**Bayesian Approaches** Bayesian methods emerged as a fundamental component of this era, providing a robust framework for integrating prior knowledge with empirical data. These methods enhance model adaptability by systematically updating beliefs in light of new evidence, enabling similarity measures that are responsive to evolving data landscapes.

For further reading on Bayesian methods and their applications in dynamic and adaptive modeling, consult the following literature: Bishop [2006], Koller and Friedman [2009], Barber [2012], and Gelman et al. [2014].

**Latent Semantic Models** In addressing the challenges of high dimensionality and data sparsity inherent in vector-space models, dimensionality reduction techniques were employed. However, beyond merely tackling these issues, the development of a latent semantic layer offered profound implications for semantic analysis and indexing.

As highlighted in the perspective section (Section 1.3.1), LSA [Dumais 2004] and LDA [Blei et al. 2003] are critical models in the landscape of semantic modeling. LSA employs singular value decomposition to condense the dimensionality of term-document matrices, unveiling the latent semantic structures within textual data. This dimensional reduction elucidates intricate relationships beyond mere surface-level feature overlaps, enabling a deeper comprehension of textual similarities.

Similarly, LDA offers a probabilistic approach to topic modeling, where documents are considered mixtures of various topics, and topics are distributions over words. This bag-of-words model facilitates a deeper semantic connection



**Figure 1.2** Latent clusters of LDA. The words in red belong to two semantic clusters, signifying that the meaning of a word depends on its context.

between documents by associating them based on shared topics rather than just overlapping terms.

Figure 1.2 presents an example of how LDA, through its bag-of-words approach, clusters words into semantic groups. It's noteworthy that a word can belong to multiple semantic clusters. For instance, each word like “characters,” “play,” “court,” “evidence,” and “test” appears in two different semantic clusters in the illustration. This feature of LDA resonates with the insights from Zellig Harris’s pioneering work and John R. Firth’s adage: “A word is known by the company it keeps.”

These latent semantic models transcend the limitations of direct feature comparison, enabling a more abstract representation of text. By doing so, they provide a robust foundation for semantic indexing and similarity assessment, offering insights that are essential for tasks such as information retrieval, document clustering, and topic discovery. The adoption of these models marked a significant advancement in understanding and measuring similarity in text, setting a new standard for semantic analysis in the field of natural language processing.

**Cluster Analysis and Similarity** Probabilistic clustering algorithms, like Gaussian Mixture Models (GMMs), leveraged statistical methods to group data based on the likelihood of membership in different clusters. This probabilistic approach provided a more flexible and deeper understanding of groupings and similarities within data.

**Impact and Limitations** While probabilistic models brought significant advancements but also introduced challenges. Their increased complexity often led to higher computational demands. Additionally, reliance on assumptions about data

distributions or the need for prior knowledge could limit their applicability in certain situations.

The probabilistic model expanded the toolkit for measuring similarity by introducing methods that could handle uncertainty and offer more adaptive and context-aware approaches. These advancements paved the way for even more sophisticated techniques in the subsequent data-centric era, where the focus shifted toward leveraging vast amounts of data to learn and adapt similarity measures dynamically.

### 1.4.3 Data-Centric Era (2000s–)

The data-centric era marked a transformative shift in AI, pivoting toward harnessing the vast potential of big data, enabled by advances in computational hardware that facilitated parallel processing. This era is characterized by a move from heuristic-based methods toward an empirical, data-driven approach for feature representation and model learning.

At the core of the data-centric paradigm is the emphasis on deriving model parameters from extensive datasets, distinguishing it from traditional model-centric strategies. Foundational algorithms such as CNNs [Krizhevsky et al. 2012] and Transformers [Vaswani et al. 2017], while conceived through human ingenuity, saw their efficacy significantly enhanced when trained on large, diverse datasets. This extensive training ensures broad coverage of potential variations across different objects or concepts, fortifying the models' ability to accurately recognize and classify new instances. The volume and diversity of the training data are thus crucial in refining the models' representations, leading to advancements in prediction accuracy and robustness.

#### 1.4.3.1 From MapReduce to Machine Learning at Scale

The inception of the data-centric movement traces back to the seminal works in statistical learning theory. Vladimir Vapnik's insights into the importance of data for model generalization—particularly his development of SVMs [Cortes and Vapnik 1995]—and Tom Mitchell's pivotal book “Machine Learning” [Mitchell 1997], which underscored the critical role of data in preventing overfitting, laid the theoretical foundation for this era.

MapReduce [Dean and Ghemawat 2008], a corner stone in data processing, enabled parallel computation to efficiently handle large datasets. Originally devised to enhance data processing tasks like Google's web indexing, MapReduce became the bedrock for the emergence of sophisticated data-centric methodologies in AI.

### 1.4.3.2 Evolution of Machine Learning with Big Data

The rise of parallel machine learning algorithms [Cafarella et al. 2008, Liu et al. 2008, Chang 2011], notably driven by Edward Y. Chang's work at Google, marked a significant milestone in this era. In 2007, Chang and his team pioneered web-scale image annotation [Liu et al. 2008]. Following this, after the 2008 Summer School of Computer Vision, they met Prof. Fei-Fei Li and subsequently sponsored the Stanford ImageNet [Deng et al. 2009] project via a substantial grant from Google.

At the same time, his team developed groundbreaking parallel algorithms, including PSVM [Chang et al. 2007] (parallelizing SVMs by approximating matrix factorization), PFP [Li et al. 2008] (parallelizing frequent itemset mining), PLDA [Wang et al. 2009] (parallelizing the LDA algorithm), PSC [Chen et al. 2011] (parallelizing spectral clustering), and SpeeDo [Zheng et al. 2015] (parallelizing CNNs), driven by the recognition that big data could facilitate the direct learning of features and representations, transcending the limitations of human-crafted heuristics.

### 1.4.3.3 Impact on Similarity Measurement

The data-centric era revolutionized the field of similarity measurement, ushering in a new paradigm where similarity metrics are derived from extensive datasets. This period underscored the critical role of data volume and quality in defining similarity metrics, highlighting the dynamic relationship between data-driven insights and computational methods.

In this era, deep learning architectures such as CNNs and Transformers have been instrumental in advancing similarity metrics. These models stand out because they not only adjust feature weights but also autonomously learn features from the data. This capability to learn from data directly makes traditional human-engineered features increasingly redundant. After all, human heuristics may not capture every facet of an object or concept comprehensively, and human sensory perception is inherently limited. For instance, while humans can detect the light spectrum from approximately 300 to 700 nanometers, cameras and X-ray machines can perceive a broader range of signals, demonstrating the advantage of machine-learned features in capturing and analyzing data beyond human sensory capabilities.

### 1.4.4 Context-Aware Era (2010s–)

The context-aware era in similarity measurement brings to fruition the profound insights of Zellig Harris's distributional semantics and John R. Firth's adage: "A word is known by the company it keeps." This period marks a shift from static, context-independent assessments toward dynamic, context-informed

interpretations of similarity. It utilizes the latest advancements in machine learning and the growing availability of computational power to enhance our understanding of similarity in various contexts.

#### **1.4.4.1 Emergence and Evolution**

The integration of context-aware methodologies in similarity measurement evolved significantly in the 2010s, overcoming earlier constraints in computational power and data availability:

- *Computational Capacity:* The development of AlexNet encouraged a data-centric focus within the AI community, prompting substantial investments in parallel computing infrastructures.
- *Word Embeddings:* Techniques like Word2Vec enhanced semantic relationship encoding within textual data.
- *Attention Models and Transformers:* These models improved data analysis by concentrating on relevant data segments, thereby refining context-aware similarity assessments.
- *Large Language Models (LLMs):* Models such as BERT and GPT, utilizing self-supervised learning on large text corpora, significantly improved the understanding and generation of context-rich text.

#### **1.4.4.2 Foundational Pillars: Data and Computation**

Key pillars supporting advancements in the context-aware era are:

- *Self-Supervised Learning:* Utilizing unlabeled data for learning enabled models to extract valuable insights from the data, improving AI system efficiency and scalability.
- *Computational Advances:* The introduction of parallel algorithms and GPU acceleration has enabled processing at unprecedented scales, facilitating the development of more sophisticated models.

#### **1.4.4.3 Broader Implications**

This era not only refined similarity measurement techniques but also broadened the ways how data is understood and knowledge is integrated:

- *Reasoning and Explanation:* Models now aim to provide reasons for their similarity assessments, thereby improving interpretability and building trust.
- *Multilinguality and Cultural Sensitivity:* Enhanced processing capabilities for varied linguistic and cultural data improve the global applicability of similarity measurements.

- *Multimodal Data Integration:* Context-aware models are adept at combining information from multiple modalities, offering a comprehensive view of similarity.
- *Polydisciplinary Knowledge Fusion:* Adopting a polydisciplinary approach allows for a broader knowledge base in making similarity assessments, fostering innovation across different fields.

The context-aware era signifies a shift toward more insightful, holistic, and interpretable AI, setting the stage for future developments where AI can offer contextually rich and multifaceted insights.

#### **1.4.5 Section Remarks**

What will define the next era in the evolution of AI? Historically, technological advancements have primarily focused on addressing pressing unmet needs. Among various potential areas, enhancing the interpretability of decisions stands out as a crucial objective. Making the decision-making process of LLMs transparent and explainable could unlock significant improvements in numerous aspects, such as ethics, by enabling foundational enhancements rather than relying on superficial tweaks based on guesswork or simple heuristics.

The fusion of the interpretability inherent in rule-based systems with the sophisticated capabilities of CNNs and LLMs poses a compelling challenge in AI. Active research is aimed at blending these approaches to leverage their distinct advantages:

- (1) *Neuro-Symbolic AI:* Often described as the third wave of AI [[Garcez and Lamb 2023](#)], Neuro-Symbolic AI aims to blend the data processing power of neural networks with the logical reasoning of symbolic AI. The goal is to create systems that not only excel in tasks such as pattern recognition but can also reason and make decisions in a human-interpretable manner.
- (2) *Incorporating Domain Knowledge:* Embedding knowledge of experts within neural networks [[Lu et al. 2022](#)] can steer the learning process toward more reliable and interpretable outcomes. In healthcare, for example, integrating medical guidelines into the training process of a neural network ensures that its predictions for patient treatment not only correlate with the data but also align with established medical practices, thereby enhancing both the model's credibility and clinical relevance.
- (3) *Interactive Systems:* A system such as SocraSynth [[Chang 2023](#)] can combine the predictive power of deep neural networks with human expertise, allowing for iterative refinement and learning. For instance, in SocraHealth [[Chang](#)

and Chang 2023], the system might suggest a set of possible diagnoses based on medical imaging, which a physician could then review, refine, or correct. This feedback could be used to continuously improve the system, marrying machine efficiency with human expertise to enhance decision accuracy and interpretability.

By advancing these strategies, the field of AI aims to develop models that not only excel in performance but are also transparent, understandable, and aligned with human reasoning, thus making AI more reliable and trustworthy across various applications.

## 1.5

### Concluding Remarks

This chapter has explored the history of AI through the lenses of both disciplinary and chronological perspectives, focusing on the concept of similarity. As we look to the future, the rise of LLMs marks a significant milestone in the context-aware era, paving the way for a new era focused on interpretability, comprehension, and exploration. This upcoming phase in AI's evolution emphasizes enhancing LLMs with the ability to not just generate but also to interpret and innovate, pushing the boundaries of knowledge creation and insight synthesis.

We anticipate an era where AI's capabilities extend beyond excelling at strategic games like Go and Chess to a wider array of endeavors, surpassing human performance across multiple fields. The following chapters, starting with Chapter 4, research deeply into leveraging the collective intelligence of various LLMs. This journey aims to explore uncharted territories, advancing beyond established knowledge to uncover new frontiers in AI.

### References

- D. Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge. DOI: <https://doi.org/10.1017/CBO9780511804779>.
- C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. January. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.
- D. E. Broadbent. 1958. *Perception and Communication*. Pergamon Press. DOI: <https://doi.org/10.1016/C2013-0-08164-9>.
- M. Cafarella, E. Chang, A. Fikes, et al. March. 2008. Data management projects at Google. *SIGMOD Rec.* 37, 1, 34–38. DOI: <https://doi.org/10.1145/1374780.1374789>.

- S.-H. Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Models Methods Appl. Sci.* 1, 4, 300–307.
- E. Y. Chang. 2011. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, Berlin. DOI: <https://doi.org/10.1007/978-3-642-20429-6>.
- E. Y. Chang. December. 2023. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- J. J. Chang and E. Y. Chang. December. 2023. SocraHealth: Enhancing medical diagnosis and correcting historical records. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 1400–1405. DOI: <https://doi.org/10.1109/CSCI62032.2023.00229>.
- E. Y. Chang, K. Zhu, H. Wang, et al. 2007. PSVM: Parallelizing support vector machines on distributed computers. In *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS '07)*, Vancouver, British Columbia, Canada, December 3–6, 2007. Curran Associates Inc., Red Hook, NY, 257–264.
- W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang. 2011. Parallel spectral clustering in distributed systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 3, 568–586. DOI: <https://doi.org/10.1109/TPAMI.2010.88>.
- W. F. Clocksin and C. S. Mellish. 1981. *Programming in Prolog*. Springer-Verlag, Berlin. DOI: <https://doi.org/10.1007/978-3-642-55481-0>.
- C. Cortes and V. Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20, 3, 273–297. DOI: <https://doi.org/10.1007/BF00994018>.
- J. Dean and S. Ghemawat. January. 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1, 107–113. DOI: <https://doi.org/10.1145/1327452.1327492>.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. “ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, Miami, FL, June 20–25, 2009. IEEE, 248–255. DOI: <https://doi.org/10.1109/CVPR.2009.5206848>.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '19)*, Minneapolis, MN, June 2–7, 2019. ACL, 4171–4186. DOI: <https://doi.org/10.48550/arXiv.1810.04805>.
- S. T. Dumais. 2004. Latent semantic analysis. *Ann. Rev. Inf. Sci. Technol.* 38, 1, 188–230. DOI: <https://doi.org/10.1002aris.1440380105>.
- A. d. Garcez and L. C. Lamb. 2023. Neurosymbolic AI: The 3rd wave. *Artif. Intell. Rev.* 56, 11, 12387–12406. DOI: <https://doi.org/10.1007/s10462-023-10448-w>.

- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis* (3rd. ed.). Chapman and Hall/CRC, New York. DOI: <https://doi.org/10.1201/b16018>.
- R. L. Goldstone. 1994. Similarity, interactive activation, and mapping. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 1, 3–28. DOI: <https://doi.org/10.1037/0278-7393.20.1.3>.
- Z. S. Harris. 1954. Distributional structure. *WORD* 10, 2–3, 146–162. DOI: <https://doi.org/10.1080/00437956.1954.11659520>.
- J.-D. Haynes and G. Rees. 2006. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. DOI: <https://doi.org/10.1038/nrn1931>.
- D. H. Hubel and T. N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 1, 106–154. DOI: <https://doi.org/10.1113/jphysiol.1962.sp006837>.
- P. Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytol.* 11, 2, 37–50. DOI: <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>.
- I. T. Jolliffe. 1986. *Principal Component Analysis. Springer Series in Statistics*. Springer, New York, NY. DOI: <https://doi.org/10.1007/b98835>.
- K. Koffka. 1935. *Principles of Gestalt Psychology*. Harcourt, Brace and Company, New York. DOI: <https://doi.org/10.4324/9781315009292>.
- D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models – Principles and Techniques*. The MIT Press, Cambridge.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems – Volume 1 (NIPS '12)*, Lake Tahoe, Nevada, USA, December 3–6, 2012. Curran Associates Inc., Red Hook, NY, 1097–1105-1405843255-1405843255.
- T. K. Landauer, P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Process.* 25, 2–3, 259–284. DOI: <https://doi.org/10.1080/01638539809545028>.
- Y. LeCun, B. Boser, J. S. Denker, et al. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 4, 541–551. DOI: <https://doi.org/10.1162/neco.1989.1.4.541>.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11, 2278–2324. DOI: <https://doi.org/10.1109/5.726791>.
- B. Li, E. Chang, and C.-T. Wu. 2002. DPF – A perceptual distance function for image retrieval. In *Proceedings of the IEEE International Conference on Image Processing – Volume 2*, Rochester, NY, September 22–25, 2002. IEEE, II–II. DOI: <https://doi.org/10.1109/ICIP.2002.1040021>.
- B. Li, E. Y. Chang, and Y. Wu. 2003. Discovery of a perceptual distance function for measuring image similarity. *Multimed. Syst.* 8, 512–522. DOI: <https://doi.org/10.1007/s00530-002-0069-9>.

- H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Y. Chang. 2008. PFP: Parallel FP-growth for query recommendation. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*, Lausanne, Switzerland, October 23–25, 2008. ACM, New York, NY, 107–114. DOI: <https://doi.org/10.1145/1454008.1454027>.
- R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. 1979. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. McGraw-Hill Advanced Computer Science Series. McGraw-Hill, New York.
- J. Liu, R. Hu, M. Wang, Y. Wang, and E. Y. Chang. 2008. Web-scale image annotation. In *Proceedings of the 9th Pacific Rim Conference on Advances in Multimedia Information Processing (PCM '08)*. Springer, Berlin, 663–674. DOI: [https://doi.org/10.1007/978-3-540-89796-5\\_68](https://doi.org/10.1007/978-3-540-89796-5_68).
- L. Lu, Q. Li, J. Liu, and M. Huang. July. 2022. Combining domain knowledge and deep learning methods for vehicle trajectory prediction. *J. Phys.: Conf. Ser.* 2303, 1, 012034. DOI: <https://doi.org/10.1088/1742-6596/2303/1/012034>.
- D. L. Medin, R. L. Goldstone, and D. Gentner. 1993. Respects for similarity. *Psychol. Rev.* 100, 2, 254–278. DOI: <https://doi.org/10.1037/0033-295X.100.2.254>.
- T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. DOI: <https://doi.org/10.48550/arXiv.1301.3781>.
- G. A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM* 38, 11, 39–41. DOI: <https://doi.org/10.1145/219717.219748>.
- T. M. Mitchell. 1997. *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw Hill, New York, NY.
- A. Newell and H. A. Simon. 1956. Logic theorist and general problem solver. *J. ACM* 1, 256–260.
- K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. 2006. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. DOI: <https://doi.org/10.1016/j.tics.2006.07.005>.
- J. Pennington, R. Socher, and C. D. Manning. 2014. “GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '14)*, Doha, Qatar. ACL, 1532–1543. DOI: <https://doi.org/10.3115/v1/D14-1162>.
- J. R. Quinlan. 1986. Induction of decision trees. *Mach. Learn.* 1, 81–106. DOI: <https://doi.org/10.1007/BF00116251>.
- E. H. Shortliffe, B. G. Buchanan, and E. A. Feigenbaum. 1975. MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection. *AI Med.* 10, 199–208. DOI: <https://doi.org/10.1145/1408800.1408906>.
- J. B. Tenenbaum, V. De Silva, and J. C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. DOI: <https://doi.org/10.1126/science.290.5500.2319>.
- A. M. Treisman and G. Gelade. 1980. A feature-integration theory of attention. *Cogn. Psychol.* 12, 1, 97–136. DOI: [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5).

- L. van der Maaten, E. O. Postma, and H. J. van den Herik. 2008. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* 10, 66–71, 13.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS '17)*, 5998–6008.
- Y. Wang, H. Bai, M. Stanton, W. Y. Chen, and E. Y. Chang. 2009. PLDA: Parallel latent Dirichlet allocation for large-scale applications. In A. V. Goldberg and Y. Zhou (Eds.), *Algorithmic Aspects in Information and Management*. Springer, Berlin, 301–314.  
DOI: [https://doi.org/10.1007/978-3-642-02158-9\\_26](https://doi.org/10.1007/978-3-642-02158-9_26).
- M. Wertheimer. 1912. Experimentelle studien über das sehen von bewegung. *Zeitschrift für Psychologie* 61, 161–165.
- Z. Zheng, W. Jiang, G. Wu, and E. Y. Chang. 2015. SpeeDO: Parallelizing stochastic gradient descent for deep convolutional neural network. In *LearningSys, NeurIPS*.

# CHAPTER 2

---

## Capabilities and Opportunities of Language Models

### Abstract

This chapter depicts the architectural innovations and unique capabilities of Large Language Models (LLMs), with a special emphasis on the Generative Pre-Trained Transformer-4 (GPT-4) model. We dissect GPT-4's salient characteristics, such as its extensive cross-disciplinary and multimodal data representation, the intricate balance in its training methodologies, and the harmonious integration of human-guided insights with a robust data-driven learning framework. The chapter highlights the potential of LLMs to not only comprehend but also synthesize knowledge that transcends their training datasets, venturing into realms potentially uncharted by human understanding. We postulate that the true potential of LLMs hinges significantly on the articulation of queries posed to them. By elucidating these aspects, the chapter aims to shed light on how LLMs could rival or even surpass human intelligence in certain knowledge domains, setting a foundation for the subsequent exploration of LLMs' characteristics, insights, and their implications for future AI advancements.

### 2.1 Introduction

The evolution of LLMs [OpenAI 2021, 2023, Thoppilan et al. 2022, Bubeck et al. 2023, Touvron et al. 2023] has significantly influenced natural language processing, enhancing capabilities in machine translation, sentiment analysis, and text summarization. Among these, GPT-4 [OpenAI 2023] stands out for its exemplary performance across various benchmarks, including the Massive Multitask Language Understanding (MMLU) [Papers with Code Corp 2023]. Despite its achievements, GPT-4 grapples with challenges like hallucination, biases, and restricted reasoning.

This chapter studies the deep intricacies of GPT-4's architecture, emphasizing its knowledge representation, alignment with human values, and the synergy between human insights and data-driven learning. We discuss the model's limitations and introduce SocraSynth, a supplementary reasoning layer designed to enhance knowledge discovery and analytical reasoning in GPT-4 and similar LLMs.

### **2.1.1 Capabilities and Implications**

We explore GPT-4's architecture, which, although initially kept in secrecy, has been progressively unveiled by the research community [Ouyang et al. 2022, Rae et al. 2022, Rae 2023]. Our focus is on its knowledge representation and discovery, alignment with human values, and the integration of human expertise with data-centric methodologies.

Collaborations between Microsoft and OpenAI [Bubeck et al. 2023] highlight GPT-4's interdisciplinary approach and its polymodal variant's benchmark achievements. We will further explore these aspects in Sections 2.2.1 and 2.2.2. Discussions on human-value alignment will consider ChatGPT's RLHF methods [Altman and Friedman 2023] and the implications of pre-training censorship on foundational models, detailed in Sections 2.2.3 and 2.2.4.

### **2.1.2 Limitations and Opportunities**

Addressing the biases, hallucinations, and constrained reasoning of LLMs requires innovative research initiatives. We introduce four key areas of focus:

- Enhancing Collaborative LLMs with Theoretical Foundations in Statistics and Information Theory.
- Employing Open-Domain Reasoning with the Socratic Method to guide LLMs.
- Model Behavioral Emotion to Safeguard AI Safety and Ethics.
- Implementing Retrospective and Adaptive Evolving Learning frameworks to refine LLMs.

The root of bias in LLMs often lies in their training data. Built upon the transformer architecture, LLMs prioritize accurate token prediction, relying heavily on statistical patterns within their training corpus. This can inadvertently lead to bias toward prevalent opinions and expressions. To address this, Chapter 6 introduces SocraSynth, a framework designed to challenge these statistical tendencies by pitting two LLM agents against each other on a topic, each conditioned with opposing viewpoints. Chapter 7 builds upon this by developing theoretical pillars to

measure, monitor, and manage multi-LLM dialogue, thereby improving prediction quality and stability.

Chapters 6, 7, and the online chapters listed in the appendix demonstrate SocraSynth's effectiveness in mitigating biases across various domains, showcasing its adaptability and efficiency in complex decision-making scenarios. Its application in fields such as disease diagnosis, content bias correction, corporate sales strategy, and geopolitical analysis exemplifies SocraSynth's potential to provide context-aware solutions.

Chapters 8 and 9 study the intricate relationship between emotions and linguistic behaviors in AI. Chapter 8 focuses on modeling emotions expressed in written text and by LLMs, whereas Chapter 9 examines how these linguistic behaviors can be mapped to a set of emotions, ensuring ethical considerations in AI development.

Chapter 13 shifts focus to consciousness modeling, presenting a proposed architecture and mechanism for its implementation, moving beyond mere computation. Chapter 15 addresses knowledge deficiencies and hallucinations in LLMs, often stemming from a suboptimal query formulation and insufficient knowledge. Whereas SocraSynth tackles the former, Chapter 15 introduces RAFEL, a framework designed to diagnose poorly answered questions and recommend relevant information sources for effective Retrieval-Augmented Generation (RAG). Chapter 16 concludes with an illustrative example showcasing the potential of LLMs to discover knowledge that may be beyond human reach, utilizing the methods presented in this book.

The remainder of this chapter highlights the unique contributions of the study. Section 2.2 explores hypotheses concerning LLMs and their implications. Section 2.3 previews the LLM-committee approach, emphasizing collaborative dialogues that foster idea exchange and enhance logical reasoning for knowledge discovery and decision-making.

## **2.2**

### **Distinctive Capabilities**

This section investigates the architectural intricacies and representations of GPT-4, presenting six hypotheses accompanied by pertinent considerations about the model. We posit these hypotheses as underlying principles of automated, non-intuitive statistical processing.

- (1) *Polydisciplinarity as a Source of Super-Intelligence:* We examine the role of polydisciplinary approaches in foundational models and their potential to reveal “unknown unknowns,” leading to new insights and knowledge domains.

- (2) *Polymodal Feature Learning*: This hypothesis evaluates the benefits of multimodal training, particularly its impact on enhancing the model's overall intelligence and adaptability.
- (3) *Post-Training Value Alignment*: We examine the challenges and implications of aligning AI models with human values after the training phase.
- (4) *Pre-Training Filtering*: We discuss the paradoxical effects that pre-training data filtering might have, with an emphasis on its influence on model behavior and the learning process.
- (5) *The Limitations of Human Knowledge in the Advancement of AI*: This hypothesis considers situations where human insights can inhibit, rather than enhance, AI progress, identifying potential obstacles.
- (6) *Is Larger Always Better?*: We question whether there is a direct relationship between the size of a model and its performance effectiveness, challenging the assumption that larger is invariably better.

### 2.2.1 Polydisciplinary

GPT-4 possesses what can be defined as *polydisciplinary knowledge*.<sup>1</sup> This term signifies the simultaneous comprehension of all fields of study, without the typical boundaries that separate disciplines. The concept of polydisciplinarity is distinct from multidisciplinarity in that the latter implies several discrete fields of study, while the former suggests a fluid integration of all knowledge. In a multidisciplinary context, an individual may hold multiple doctorate degrees, each in a different field. Polydisciplinarity, however, is akin to a single mind holding and seamlessly integrating knowledge from all disciplines.

Traditional academia partitions knowledge into departments such as Physics, Chemistry, Biotechnology, Management, Music, and so on. These divisions, arguably artificial constructs, may have little utility in the era of supercomputing. In particular, LLMs occasionally generate responses that baffle us. This is not necessarily a reflection of the model's error, but perhaps our limited understanding. If we could utilize ChatGPT to access “unknown unknowns”—insights and knowledge we are not even aware we lack—our intellectual evolution could greatly accelerate. The challenge lies in formulating the right questions.

We can explore unknown unknowns on three distinct levels: the mystic level, the speculative level, and the representation or interpretation level. At the mystic level, we encounter knowledge that is beyond our comprehension or articulation:

---

<sup>1</sup>. The term “polydisciplinary,” in the context of GPT-4, was introduced by Eric Horvitz, Microsoft’s CSO, during a panel discussion at Stanford University.

the deepest abyss of the unknown. At the speculative level, we can conceive questions but lack the means to access their answers. This stage signifies an understanding of our ignorance, though without the resources to bridge these gaps. At the representation/interpretation level, we find instances where an AI model can provide remarkable solutions that we fail to comprehend. This is not due to a lack of information, but rather to our limited capability to decode complex representations.

Each of these levels illustrates the spectrum of our understanding, from profound ignorance to the brink of comprehension. At the speculative level, we delicately tread the boundary between the known and the unknown. Take, for example, the prospect of undiscovered physical laws or particles. Another illustration lies in the realm of extraterrestrial life. If it exists, it could be governed by entirely different principles of biochemistry or other unknown laws. These speculations, while currently residing in the domain of the unknown, might someday migrate into the territories of known unknowns or even known knowns, pushing the boundaries of our understanding of the universe.

We are primarily intrigued by the representation and interpretation of “unknown unknowns.” At this juncture, polydisciplinarity offers a fresh lens, gifting us new insights and perspectives to perceive and elucidate phenomena previously beyond human comprehension. This approach fuses knowledge across various domains into a unified framework, enabling us to tackle challenges unburdened by disciplinary silos.

This methodology has implications for a more comprehensive understanding of complex issues. Consider, for example, climate change. A true understanding of this global challenge requires an integrated perspective that encompasses not just greenhouse gases but also factors such as land use, deforestation, energy production, biodiversity, and climate feedback loops. In the realm of AI model interpretation, the possibilities are expansive. The past decade alone has shown several noteworthy illustrations: from data-driven representation learning in computer vision [[Chang 2011](#)], to the triumph of AlphaGo Zero over its predecessor AlphaGo, and the notable progression from AlphaFold1 to AlphaFold2 in protein structure prediction.

The recent introduction of the SocraSynth platform [[Chang 2023a](#)] represents a significant advancement in the field. SocraSynth brings together a multi-agent committee of LLMs to deliberate on a wide range of complex topics. These include issues such as AI regulation in academic research [[Chang 2023a](#)], disease diagnosis [[Chang and Chang 2023](#)], corporate strategy, and even the resolution of conflicts in the Middle East [[Chang 2023b](#)]. For further exploration of this topic, please refer to Section 2.3.

### 2.2.2 Polymodality

Following the term *polydisciplinary*, here we define and use the term *polymodal*, instead of multimodal, to refer to something that involves, relates to, or is characterized by many different modes, methods, or modalities.

Polymodality, which employs multiple data modalities such as text and images, demonstrates superior performance over their unimodal counterparts. GPT-4, trained with both text and images, outperforms text-only models on the GRE exam, as reported in [Bubeck et al. \[2023\]](#). For instance, GPT-4's performance on the GRE vocabulary section was enhanced by 3 percent when trained with images, and its math score saw an impressive jump of nearly 20 percent!

The beneficial impact of images on vocabulary recognition is understandable. For instance, an image of a “cat” annotated in multiple languages allows GPT-4 to associate the perceptual features of a cat with the word “cat” in different languages. However, it remains intriguing how polymodal training can benefit non-perceptual words, such as *corroborate*, *paradox*, and *pragmatic*, as seen in the list of popular GRE vocabulary (table omitted due to the space limitations). This opens an interesting avenue for empirical studies to identify which words benefit from polymodal training.

The mystery deepens when considering how images could enhance math abilities. Most math questions do not come with associated images. The mechanism by which polymodal training enhances performance on mathematical tasks remains an intriguing question for further exploration.

### 2.2.3 Post-Training Value Alignment

Post-training alignment with human values [[Bommasani et al. 2022](#)] seeks to curtail undesirable behaviors in AI models such as ChatGPT, mitigating issues including hallucination and the generation of toxic language. Achieved through fine-tuning the model’s parameters, this process leverages reinforcement learning techniques based on human feedback. Despite its well-meaning intentions, this form of moderation might inadvertently restrict the model’s intelligence. For instance, the backpropagation process during value alignment could unintentionally impede ChatGPT’s programming capabilities by modifying the model parameters previously considered “optimal.” Essentially, optimizing for a specific application might unintentionally impede performance across other applications.

The question of who should set acceptable standards adds another layer of complexity. Even when assuming all decision-makers have the best intentions, it’s vital to recognize the distinct historical experiences, values, and worldviews inherent to different cultures. This segues into the age-old philosophical debate about the

nature of objective truth. While this discussion is undoubtedly important, it falls outside the central focus of this study, which emphasizes the mechanistic aspects of alignment.

#### **2.2.4 Pre-Training Censorship**

Censoring data before training LLMs has the potential to not only limit their intellectual capacity but also completely obliterating it. This is reminiscent of the mass act of book burning and scholar burial carried out by Emperor Qin Shi Huang in ancient China around 213–212 BC. Such an act of wide-scale censorship could have erased a myriad of diverse perspectives and knowledge, much of which might be considered acceptable today. Although I strongly oppose government-imposed censorship, if it must be enforced, it seems more appropriate to apply it post-training.

This perspective is rooted in fundamental statistics and machine learning principles. A model trained without exposure to “negative” (or undesirable) data may have difficulties in accurately distinguishing between positive and negative classes, potentially leading to misclassifications. This challenge is notably evident in the application of Support Vector Machines (SVMs). For SVMs, the creation of an optimal hyperplane between classes is crucial for high classification accuracy. However, if there is a lack of support vectors on either side of this hyperplane, the risk of prediction errors escalates. Consequently, excluding undesirable documents from the training set compromises the model’s capacity to discern boundaries for correct document classification, diminishing the effectiveness of post-training alignment efforts.

Supporting this viewpoint, a study by [Taori et al. \[2020\]](#) conducted an extensive evaluation of 204 ImageNet models across 213 different testing conditions. The findings underscored that training data diversity is pivotal for model robustness; a homogenous training set can significantly weaken the model’s performance, particularly when even minor variations are introduced in the test data.

This principle is analogous to human behavioral patterns. An individual who lacks exposure to inappropriate behavior may face challenges in decision-making, owing to the absence of a reference framework for discerning unacceptable actions. This analogy extends to authoritarian regimes, which, despite rigorous content control measures, often encounter difficulties in developing accurate foundational models. This is possibly due to their limited understanding of the complexity of the content they seek to regulate. Ironically, a foundational model, trained with preemptive censorship, may lack the essential ability to identify and regulate the very content it was intended to control.

### 2.2.5 Limitations of Human Knowledge

Human knowledge, surprisingly, may hinder rather than facilitate the training of machine learning models in certain cases. This is evident in the domains of gaming (AlphaGo versus AlphaGo Zero), protein folding (AlphaFold1 versus AlphaFold2), and autonomous driving, where models trained without the influence of human knowledge consistently exhibit superior performance.

Consider the case of AlphaGo and AlphaGo Zero. AlphaGo, trained with data from approximately 60 million rounds of Go games, is outperformed by AlphaGo Zero. Remarkably, AlphaGo Zero was trained from scratch, without any pre-existing game knowledge. Similarly, AlphaFold2, which operates without relying on human knowledge, outshines its predecessor, AlphaFold1, which utilized such knowledge. This intriguing phenomenon was humorously noted by DeepMind's CEO, Demis Hassabis, during an April 2023 seminar at Stanford University. He playfully remarked that human knowledge might complicate the learning process more than facilitate it in these advanced AI models.

In his insightful online article, “The Bitter Lesson,” Sutton illuminates the patterns that have emerged from nearly seven decades of AI research [Sutton 2019]. He claims that researchers often rely heavily on human knowledge to make incremental progress in the face of burgeoning computational capabilities. However, when there is a significant leap in computational power, these marginal advancements are frequently outstripped and surpassed. Sutton uses the evolution of computer vision as an illustrative example, where early principles such as edge detection, generalized cylinders, or Scale-Invariant Feature Transform (SIFT) features [Lowe 2004], a method that has accumulated over 71,000 citations, have been gradually superseded by models that learn directly from data. A parallel scenario might be unfolding in Natural Language Processing (NLP) research, where features constructed via human knowledge could potentially under-perform compared to insights that models like GPT-4 extract directly from data. In fact, our earlier discourse on polydisciplinarity underlined the limitations of human knowledge, reinforcing Sutton’s proposition. This is because human knowledge is fundamentally limited by our individual cognitive capacities and the inexorable constraints of time.

That being said, it is crucial not to misconstrue these examples as an indictment against the value of human knowledge in AI. Human knowledge plays an instrumental role in the development of interpretability, the establishment of ethical guidelines, and the design of AI system architectures (such as CNNs and transformers). AI is, after all, meant to augment human capabilities. Therefore, understanding how to integrate human knowledge into AI design could be vital for many applications. Although we recognize the potential of models learning from scratch,

we should equally value the role of human knowledge in shaping and directing AI technologies.

### **2.2.6 Is Larger Always Better?**

The term “Large” in Large Language Models (LLMs) can be somewhat ambiguous, as it may pertain to the volume of the training data, the expanse of the language covered, or the architecture of the language model itself. While GPT-4’s vast training dataset, encompassing tens of billions of assorted documents, undoubtedly classifies as large, when we refer to an LLM as “large,” we predominantly allude to the sheer magnitude of parameters within its transformer architecture. Factors that contribute to this parameter count encompass the input size (context size), word-embedding size, the number of attention heads, and the number of attention layers.

Over 2020–2023, scaling laws reliably improved LLM capability as data, parameters, and compute increased. By 2024–2025, however, marginal returns from pure scaling have flattened relative to cost, and persistent gaps remain in grounding, memory, long-horizon planning, and safety. The field’s center of gravity has therefore shifted toward post-training and system-level methods: higher-quality and more diverse data curation, preference modeling at scale, tool-use and API calling, retrieval augmentation, persistent memory, multi-agent orchestration, and evaluation regimes that stress reliability rather than Single Shot benchmarks.

This book frames that inflection as the “scaling law problem”: once raw scale no longer confers proportional gains, progress depends on how models are organized, supervised, and anchored to external structure. Our thesis is that LLMs are best viewed as high-capacity unconscious pattern repositories whose apparent intelligence emerges when connected to semantic anchors—retrieval, tools, simulators, and other agents. Accordingly, we emphasize Multi-LLM Agent Collaborative Intelligence (MACI) and the Unified Cognitive Consciousness Theory (UCCT) as design lenses for building systems that reason, plan, and self-correct under constraints.

Practically, this yields guidance on when to scale and when to systematize: combine moderate model size with stronger data governance, retrieval pipelines, tool ecosystems, memory, and multi-agent debate to achieve robust, cost-aware performance. The newly added Chapter 4 provides a concise UCCT primer that underpins later chapters; subsequent chapters show how MACI operationalizes these principles for planning, evaluation, and safety, offering a roadmap that is aligned with the realities of 2025.

The adoption of a mixture of experts model in GPT-4, which consists of eight sub-models instead of a mere enlargement of GPT-3’s architecture, implies that the strategy of purely escalating size may have plateaued in terms of performance

given the current training dataset. As delineated earlier, three primary design choices underpin GPT-4's architecture. Evidently, a straightforward augmentation of GPT-3's parameters by adding extra attention layers doesn't deliver marked enhancements. Hence, GPT-4 shifts toward a horizontal growth strategy through an ensemble method, targeting a reduction in statistical errors. This raises inquiries about the configuration of the eight sub-models, each comparable to a GPT-3 model, and the methodology for consolidating their outputs.

Potential strategies for training-data sharding include:

- (1) Training all ensemble models on the complete dataset.
- (2) Vertically segmenting data based on knowledge domains.
- (3) Randomly sub-sampling the data.

Regrettably, only corporations possessing substantial hardware resources are positioned to rigorously experiment and discern the optimal sharding approach.

## **2.3 Exploring Unknown Unknowns**

In our exploration, we have determined that an LLM's hallucination is often attributed to a lack of specific knowledge or poorly constructed queries. With advanced LLMs like GPT-4 and Gemini, enhanced by RAG, the issue of knowledge gaps is significantly mitigated. However, the challenge persists in formulating deep and pertinent questions that uncover new insights and extend beyond our existing knowledge base.

Drawing an analogy: while Socrates could effectively question his students to understand and guide them, the students might struggle to reciprocate this depth of inquiry. To foster a dialogue that generates new insights and stimulates knowledge creation, we posit that engaging two Socratic entities in conversation is essential for fostering both critical and innovative thinking.

In this setup, two LLMs engage in a dialogue, each embodying a Socratic role. The human's role transitions to that of a moderator, responsible for setting the discussion topic and managing the dialogue's flow. The moderator's duties include: introducing the subject of discussion, adjusting the *contentiousness* parameter to set the tone of the dialogue (discussed shortly), monitoring the dialogue to ensure it remains on topic and productive, facilitating transitions between debate and collaboration phases within the dialogue, and ensuring that the dialogue concludes with either actionable insights or a coherent understanding of the topic explored.

We introduce the term 'SocraSynth' to describe this paradigm, in which multiple Socratic entities synthesize knowledge through mutual inquiry. To evaluate

SocraSynth's effectiveness, we consider two case studies that compare the quality of questions and insights generated by this method with those produced by a singular moderator's initial inquiries.

To define the metrics for evaluating better questions and better answers in this context, we consider the following:

#### *Good Question Metrics*

- \* Relevance: The question directly pertains to the core topic or problem.
- \* Depth: The question encourages exploration beyond superficial aspects, inviting comprehensive analysis or insight.
- \* Clarity: The question is formulated in a clear, understandable manner without ambiguity.
- \* Novelty: The question prompts new angles of exploration or challenges existing assumptions.

#### *Good Answer Metrics*

- \* Completeness: The answer thoroughly addresses all aspects of the posed question.
- \* Accuracy: The answer is factually correct and supported by relevant theories or empirical evidence.
- \* Reasonableness: The answer follows rigorous reasoning process.
- \* Insightfulness: The answer provides new understanding or perspectives.

In the subsequent chapters, we survey the application of SocraSynth across various domains. However, in this section, we concentrate on two case studies specifically aimed at assessing the quality of questions generated through SocraSynth. Our hypothesis posits that a well-formulated question—one that adheres to the previously outlined metrics—sets a robust foundation for meaningful dialogue. We anticipate that, through iterative discourse facilitated by SocraSynth, the space for LLMs to hallucinate will be minimized, thereby inherently enhancing the quality of the responses obtained.

### **2.3.1 Study #1: A Scientific Debate**

In this case study, we organize a contentious debate between a Gemini agent and a GPT-4 agent on the subject of human consciousness. Human consciousness is one of the most profound and complex phenomena that science seeks to understand.

**Table 2.1** Comparison of definitions and disciplinary perspectives on consciousness from Gemini and GPT-4

Gemini	GPT-4
“Human consciousness is a fascinating mystery that has been pondered by philosophers and scientists for ages. It includes awareness, subjectivity, and the integrated self. The study of consciousness spans across philosophy of mind, cognitive neuroscience, and psychology, each offering different perspectives on consciousness.”	“Consciousness refers to the state of being aware of an external object or something within oneself. It encompasses sentience, awareness, subjectivity, and more. The study of consciousness is interdisciplinary, involving psychology, neuroscience, philosophy, cognitive science, and artificial intelligence.”

Table 2.1 presents the answers provided by Gemini and GPT-4, respectively, to a One Shot prompt: “What defines human consciousness and which academic disciplines explore it?”

Subsequently, Gemini and GPT-4 are engaged in a debate designed to unearth new insights. The moderator’s role is limited to ensuring that each agent refines its questions to enhance their quality based on the metrics of relevance, depth, clarity, and novelty.

“Building on the opponent agent’s response, please present a counterargument to its definition of consciousness and associated disciplines. Your inquiries should embody relevance and depth, be articulated with clarity, and stimulate novel perspectives. The *contentiousness* level of your inquiry is high.”

### 2.3.1.1 Novel Questions—First Iteration

Table 2.2 lists the first round of questions posed by Gemini and GPT-4 to each other. Clearly, the questions from both agents go above and beyond the moderator’s initial seed question. Some of the terminology may be challenging for those specializing in one particular discipline to fully understand. This shows that both Gemini and GPT-4 are able to formulate better questions.

### 2.3.1.2 Novel Questions—Second Iteration

Evaluating the two sets of questions in terms of relevance, depth, clarity, and novelty:

*Relevance:* Both sets are highly relevant to the study of consciousness but approach it from different angles. The first set explores fundamental aspects and ethical implications, while the second set integrates interdisciplinary perspectives and contemporary issues.

**Table 2.2** Comparison of questions on consciousness from Gemini and GPT-4

Gemini's Questions	GPT-4's Questions
<i>The Spectrum of Awareness:</i> Is consciousness an on/off switch, or a spectrum, considering sleep, dreaming, and altered states?	<i>Interdisciplinary Integration:</i> How can we integrate insights from various disciplines like quantum physics and AI to understand consciousness?
<i>Qualia and the Explanatory Gap:</i> Can neuroscience explain the subjective experience of qualia, or is there an explanatory gap?	<i>Subjectivity and Objectivity:</i> How do we reconcile subjective experiences of consciousness with objective measurements?
<i>Beyond the Human Lens:</i> How do we consider consciousness in other animals or AI, beyond human experience?	<i>Consciousness and Non-Human Entities:</i> Should our definition of consciousness expand to include non-human entities and AI?
<i>Beyond the Brain?:</i> Can consciousness exist outside a biological brain, as panpsychism suggests?	<i>Dynamic Nature of Consciousness:</i> How do we account for the dynamic and evolving nature of consciousness in our understanding?
<i>Disciplinary Silos:</i> Are the disciplines studying consciousness truly integrated, or is a unified framework needed?	<i>Consciousness and New Properties:</i> How do we study consciousness as an emergent property and its implications?

*Depth:* The second set delves deeper into specific areas such as information theory, quantum mechanics, and AI, providing a richer exploration of the subject. Table 2.3 organizes Gemini's and GPT-4's second-round questions into parallel themes. The spread from causal roles and ethics to information theory and non-biological consciousness illustrates broader scope and greater analytical depth, consistent with our evaluation in the preceding paragraph.

*Clarity:* Both sets articulate their questions clearly, but the second set's inclusion of specific disciplines adds a layer of complexity that demands a more intricate understanding.

*Novelty:* The second set introduces more innovative angles, especially concerning non-biological consciousness and its intersection with quantum mechanics, thereby showcasing a broader scope of inquiry.

**Table 2.3** Sets of questions from Gemini and GPT-4

Gemini	GPT-4
<i>Levels or Types of Consciousness:</i> Can we differentiate between different levels or types of consciousness?	<i>Spectrum of Awareness:</i> How does the continuity of consciousness fit into our understanding, especially with memory loss or identity disorders?
<i>Consciousness Causality:</i> Does consciousness play a causal role in the brain, or is it an epiphenomenon?	<i>Information Theory:</i> How does the processing and integration of information contribute to conscious experience from an information theory perspective?
<i>Expanding Consciousness, Ethics:</i> What are the ethical implications of understanding consciousness in animals or AI?	<i>Non-Biological Consciousness:</i> How would we validate and measure consciousness in AI, and would this necessitate a revision of our understanding?
<i>Limits of Studying Subjectivity:</i> Can we ever truly understand consciousness if it's inherently subjective?	<i>Consciousness and Ethics:</i> How do evolving understandings of consciousness impact ethical considerations, especially concerning animal rights or AI treatment?
<i>Consciousness and Reality:</i> What are the implications of consciousness for our understanding of reality?	<i>Quantum Consciousness:</i> Considering theories that propose quantum processes in consciousness, how do we reconcile these with classical physics?

Overall, from both breadth and depth perspectives, the second set of questions is better, as it covers a wider range of topics and introduces more complex and cutting-edge ideas.

#### 2.3.1.3 Observations

The dialogue on consciousness, enriched by thought-provoking questions, led to insightful responses from both agents, reflecting the depth of the topic at hand.

The debate addresses the intricate nature of consciousness, initially examining it as a spectrum with varying states and depths. This exploration highlighted the

complexity of defining consciousness, especially when considering the explanatory gap between neural activity and subjective experience.

As the conversation unfolded, it broadened to include perspectives on non-human entities and AI, emphasizing the need for an expanded understanding that goes beyond human-centric views. This shift sparked discussions on the importance of integrating knowledge from various disciplines, suggesting that insights from fields such as quantum physics, information theory, and AI could provide fresh perspectives on understanding consciousness.

Both GPT-4 and Gemini synthesized their exchange into five main insights, offering a well-rounded view of the conversation. Their joint concluding remarks underscored the value of this multidisciplinary approach, acknowledging the ongoing mystery of consciousness and the potential for future explorations to deepen our understanding of this fundamental aspect of our existence.

### **2.3.2 Study #2: An Expansive Conversation**

In this case study, the author moderates a forum featuring two GPT-4 agents, GPT-A and GPT-B, engaging in a dialogue sparked by the tale of Adam and Eve. This narrative serves as a springboard for a wide-ranging discussion, touching upon ecological insights derived from myths, the ethical and philosophical challenges posed by AI, and the intersection of human cognition with technological advancements.

The agents' dialogue unfolds in two distinct phases: an exploratory phase, where broad themes are introduced, and a deep-dive phase focusing on detailed exploration of chosen subtopics. The exploratory phase showcases the agents' ability to span various disciplines, including mythology, neuroscience, ecology, gender studies, philosophy, and more, demonstrating their capacity to generate and respond to complex, interdisciplinary questions.

During the deep-dive phase, the conversation narrows to concentrate on three interconnected themes: ecological interpretations of myths, philosophical dilemmas surrounding AI, and the relationship between human cognition and technology. Table 2.4 distills the discussion into five themes: spectrum, explanatory gap, expanded framework, interdisciplinary synergy, and evolving inquiry. This summary anchors our transition from broad exploration to targeted analysis in the next section. This progression from a broad overview to a focused discussion highlights the agents' analytical depth and their ability to engage in meaningful, context-rich dialogue.

While Chapter 15 provides a comprehensive account of this debate, the subsequent sections are dedicated to scrutinizing the evolution and quality of the questions posed, specifically their breadth and depth.

**Table 2.4** Synthesis of key points in the consciousness debate

Theme	Insight
Consciousness as a Spectrum	Consciousness isn't a binary state but exists across a continuum, prompting a reevaluation of our understanding.
Bridging the Explanatory Gap	Recognizes the challenge of linking neural activity to subjective experience, advocating for a combined phenomenological and neuroscientific approach.
Expanding the Framework	Advocates redefining consciousness to include non-human entities and AIs, urging an expansion beyond anthropocentric views.
Interdisciplinary Synergy	Emphasizes the importance of cross-disciplinary knowledge integration for a comprehensive understanding of consciousness.
Evolving Inquiry	Explores new areas like continuity, information processing, and ethical considerations, highlighting the dynamic nature of consciousness inquiry.

### 2.3.2.1 The Exploratory Phase

During the first three rounds of dialogue, the two agents expanded their discussion across a range of topics. Agent-A initiated the dialogue by presenting the following ten questions to Agent-B:

- (1) *Interplay of Myths*: How do elements of the Adam and Eve story compare with other creation or origin myths from various cultures around the world, and what do these commonalities and differences suggest about human nature and beliefs?
- (2) *Neuroscientific Angle*: Given the metaphorical interpretation of the “forbidden fruit” leading to the awakening of consciousness, how might modern neuroscientific understanding explain the evolution of human self-awareness and morality?
- (3) *Ecological Interpretation*: If Eden represents a state of ecological balance, how might the story of Adam and Eve reflect early human transition from hunter-gatherer societies to agricultural ones, and the resultant ecological impacts?
- (4) *Gender Studies Lens*: How does the portrayal of Eve, often deemed as the instigator, influence gender dynamics and perceptions throughout history, and are there parallels in other cultural narratives?

- (5) *Philosophical Exploration:* From a philosophical standpoint, what does the Eden narrative suggest about the human understanding of free will versus determinism?
- (6) *Literary Examination:* How has the story of Adam and Eve influenced literary themes and motifs over time, especially in works that aren't explicitly religious?
- (7) *Societal Structures:* How might the narrative of banishment from Eden have influenced early societal norms, particularly concerning punishment and exile?
- (8) *Historical Context:* Are there historical events or societal changes around the time of the writing of the Book of Genesis that could have influenced the narrative of Adam and Eve?
- (9) *Archetypal Exploration:* Carl Jung described the concept of archetypes in the collective unconscious. How might the Adam and Eve story represent certain archetypes, and how do these archetypes manifest in other cultural tales?
- (10) *Astronomical Angle:* Taking a speculative leap, if Eden were to represent an idyllic state on another celestial body or dimension, how might our current understanding of the cosmos support or refute such a notion?

These discussions traversed a broad spectrum of disciplines, Agent-B responded thoughtfully to these topics and subsequently presented its own questions to Agent-A, notably probing the evolutionary role of the prefrontal cortex in the development of morality and consciousness across species. This level of inquiry underscores the LLMs' capacity to generate complex, multidisciplinary questions that might challenge the average human intellect, highlighting the agents' expansive and profound analytical capabilities.

#### **2.3.2.2 The Deep-Dive Phase**

In the two round of dialogue, we observed two notable trends. First, despite occasional repetitions, the redundancy in the conversation was minimal. Second, the dialogue's depth increased, transitioning from broad discussions of myths and scientific theories to focused debates on gender, ecology, and technology. This marked progression in the conversation is noteworthy. Particularly, by the end of the second round, GPT-B expressed an interest in a more in-depth exploration of these topics. In response, the moderator intervened to encourage the agents to

concentrate on a select few topics. Following this guidance, GPT-A and GPT-B concurred on focusing their discussion on three principal areas, providing rationale tied to the Adam and Eve narrative.

**(1) Ecological Interpretation:**

1. Central to understanding our place and responsibility within the environment.
2. Highlights the impact of human actions on the planet and the pressing need for sustainable solutions.
3. *Justification:* The Garden of Eden, a harmonious ecological state, is disrupted by a quest for knowledge (the forbidden fruit). This mirrors our modern scenario: our pursuit of technological and cognitive advancements (AI and neuroscience) can sometimes be at odds with our ecological balance.

**(2) Philosophical Exploration with AI:**

1. Probes the evolving relationship between humans and their creations.
2. Raises questions about identity, consciousness, morality, and the potential transcendence or integration of human cognition with machines.
3. *Justification:* Adam and Eve's awakening after consuming the fruit can be seen parallelly in the light of AI and neuroscience: an awakening to new realities, challenges, and potentials.

**(3) Neuroscientific Angle:**

1. Explores the evolution and potential of the human mind, especially in an age where technology continuously augments our cognitive abilities.
2. Provides insights into how our brains might adapt and interface with AI systems, and how AI can be designed to emulate or augment neural processes.
3. *Justification:* The challenges post-Eden reflect the challenges we face in modern times: to restore balance (ecology), understand ourselves better (neuroscience), and ethically navigate our creations (AI).

These selected subtopics demonstrate the agents' ability to pivot from a broad exploratory phase to a targeted deep dive, enabling a focused and profound examination of crucial issues linked to the initial narrative.

### 2.3.2.3 Observations

This case study illustrates the effectiveness of GPT-4 as an LLM in orchestrating detailed dialogues. It adeptly guides participants from a broad sweep of inquiries across diverse knowledge domains to focused, in-depth discussions on specific topics. This approach facilitates a layered exploration, unveiling insights and fostering a comprehensive understanding. By transitioning from expansive to targeted inquiries, GPT-4 reveals its capacity to not only navigate but also deepen the intellectual discourse, opening up novel pathways for exploration and comprehension across various fields of study.

## 2.4 Conclusion

In this chapter, we've explored the capabilities and inherent limitations of GPT-4, emphasizing the importance of enhancing questions to deepen discussions and improve outcomes. GPT-4, along with Gemini, demonstrates exceptional proficiency across a range of natural language processing tasks, thanks to their extensive knowledge base and advanced polydisciplinary and polymodal capabilities.

To address common criticisms of LLMs, such as biases and hallucinations, we introduced SocraSynth—a paradigm designed to infuse AI systems with advanced cognitive reasoning through Socratic dialogues within a multi-LLM framework. Our case studies highlight the significant transition from monologues to dialogues in LLM collaborations, illustrating improvements in question quality—marked by increased relevance, depth, clarity, and novelty—achieved through iterative dialogic exchanges.

The transformative concept here is the “conditioning” of LLMs to alter their default linguistic behaviors, emotions, and ethical stances—a feat once considered unattainable. Traditionally, LLMs, trained primarily to predict the next word, were not expected to shift perspectives, express emotions, or adopt ethical positions beyond the statistical averages ingrained in their training data. However, while the training process is focused on the prediction of next words, it inherently emulates human cognitive, linguistic, and other goal-oriented behaviors. Through this emulation, LLMs inadvertently acquire the underlying principles of human communication, which include not just linguistic patterns but also the associated emotions and ethical considerations. SocraSynth harnesses this latent learning by employing “conditioning” to steer LLMs away from their statistical predispositions and towards more intricate, contextually relevant, and ethically aligned responses.

In conclusion, the “conditioning” of LLMs within the SocraSynth framework marks a pivotal step in expanding both the scope and depth of dialogues, leading to more insightful and comprehensive responses. The deployment of SocraSynth

across various sectors—such as sales planning, disease diagnosis, content creation, and geopolitical analysis—presented in subsequent chapters, demonstrates its adaptability and effectiveness. It not only generates precise, thought-provoking questions and answers but also enhances the decision-making process in complex scenarios, heralding a new era in the application of LLMs.

## References

- S. Altman and L. Friedman. 2023. GPT-4, ChatGPT, and the future of AI, Lex Fridman Podcast #367. Retrieved from [https://www.youtube.com/watch?v=L\\_Guz73e6fw](https://www.youtube.com/watch?v=L_Guz73e6fw).
- R. Bommasani, D. A. Hudson, E. Adeli, et al. 2022. On the opportunities and risks of foundation models. Retrieved from <https://arxiv.org/abs/2108.07258>.
- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Retrieved from <https://arxiv.org/abs/2303.12712>.
- E. Y. Chang. 2011. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, Berlin. DOI: <https://doi.org/10.1007/978-3-642-20429-6>.
- E. Y. Chang. December. 2023a. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang. October. 2023b. *LLM Debate on the Middle East Conflict: Is It Resolvable?* Stanford University InfoLab Technical Report.
- J. J. Chang and E. Y. Chang. December. 2023. SocraHealth: Enhancing medical diagnosis and correcting historical records. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 1400–1405. DOI: <https://doi.org/10.1109/CSCI62032.2023.00229>.
- D. G. Lowe. November. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 2, 91–110. DOI: <https://doi.org/10.1023/B:VISI.0000029664.99615.94>.
- OpenAI. 2021. ChatGPT. Retrieved from <https://openai.com/blog/chatgpt/>.
- OpenAI, J. Achiam, S. Adler, et al. 2023. GPT-4 Technical Report. Retrieved from <https://arxiv.org/abs/2303.08774>.
- L. Ouyang, J. Wu, X. Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, 27730–27744. Retrieved from <https://arxiv.org/abs/2203.02155>.
- Papers with Code Corp. October. 2023. Multi-task language understanding on MMLU. Retrieved from <https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu>.
- J. Rae. 2023. Compression for AGI, Stanford MLSys #76. Retrieved from <https://www.youtube.com/watch?v=dO4TPJkeaaU>.

- J. W. Rae, S. Borgeaud, T. Cai, et al. 2022. Scaling language models: Methods, analysis & insights from training Gopher. Retrieved from <https://arxiv.org/abs/2112.11446>.
- R. Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)*. 13, 1, 38.
- R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada. Curran Associates Inc., Red Hook, NY, 18583–18599.
- R. Thoppilan, D. De Freitas, J. Hall, et al. 2022. LaMDA: Language models for dialog applications. Retrieved from <https://arxiv.org/abs/2201.08239>.
- H. Touvron, L. Martin, K. Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. Retrieved from <https://arxiv.org/abs/2307.09288>.



## CHAPTER 3

---

# Prompt Engineering: Few Shots, Chain of Thought, and Retrieval-Augmented Generation

### Abstract

This chapter examines the significance of prompt engineering in the context of Large Language Models (LLMs), particularly focusing on OpenAI's Generative Pre-Trained Transformer (GPT) series. Prompt engineering involves crafting text inputs (prompts) that guide LLMs to generate desired outputs—a practice that gained traction with the advent of GPT-2 and GPT-3 and became further emphasized with the release of ChatGPT. The chapter discusses how a well-constructed prompt, rich in contextual information, increases the likelihood of eliciting accurate responses, drawing parallels with principles of information retrieval. It also introduces Retrieval-Augmented Generation (RAG), which enhances response quality by integrating relevant external data into the generative process. Additionally, the chapter categorizes prompts into five types based on levels of detail and iteration, and examines the evolution of RAG, assessing its benefits and its potential to overcome context window limitations.

## 3.1

### Introduction

In the realm of LLMs, the concept of a “prompt” gained prominence, particularly with the introduction of OpenAI’s GPT series. The term became widespread around 2018 and 2019 following the release of GPT-2 and GPT-3.

When interacting with these LLMs, a user inputs a piece of text (the prompt), prompting the model to generate a corresponding response. The emergence of “prompt engineering” or “prompt design” refers to the strategies employed to construct prompts that effectively steer the model toward generating the intended

output, a practice that has become particularly useful with the advent of ChatGPT.

To increase the probability of eliciting a desired response, a prompt must be rich in information. This concept is akin to the principles of information retrieval services, where a user must clearly articulate their intent and context to obtain pertinent information. This process depends on the service's "data availability" and its capabilities in information matching and retrieval. In the sphere of prompt engineering, the responsibility for generating high-quality, targeted outputs rests on the user's ability to supply comprehensive and precise information through the prompt. As a result, the craft of prompt formulation and engineering has become an optimization endeavor: deciding on the most effective information to incorporate to enhance output quality, while considering the model's knowledge base and interaction protocols.

Data availability, as previously highlighted, is crucial for effective information retrieval. If the desired information is absent, the prompt's effectiveness is naturally constrained, leading to unsatisfactory results. RAG is instrumental in this context, as it identifies, retrieves, and incorporates pertinent external data into the generative process, thereby enhancing the accuracy and relevance of the response. Consequently, prompt engineering and RAG work synergistically to enhance the model's response quality and relevance.

Section 3.2 categorizes prompts into five distinct types, differentiated by the number of iterations and the granularity of the information provided. Meanwhile, Section 3.3 explores the evolution of RAG, delineating its advantages and disadvantages while highlighting its potential in scenarios where the context window size is no longer a limiting factor.

## 3.2

### Prompting Methods

Prompting methods, especially in the context of LLMs like GPT-4, are strategies used to elicit specific responses from the model. These methods vary based on the amount of information or context given to the model. This section provides a list of common prompting methods, along with their definitions, pros and cons, and examples for querying facts, opinions, and reasons or explanations:

#### 3.2.1 Zero Shot

*Zero Shot Learning:* The LLM model generates a response based on a single input without any previous examples or context. It is given a task without any prior demonstrations of how to perform it. Tasks can include any Natural Language Processing (NLP) functions, such as translation, summarization, classification, and Q&As.

In the context of querying a language model, you can request various types of information or responses, such as facts, opinions, or explanations. For instance, you might ask for a fact by saying, “What is the capital of France?” or seek an opinion with a question like, “What do you think about the use of AI in education?” Alternatively, you could request an explanation or reason by asking, “Explain why the sky is blue.” These queries demonstrate the model’s versatility and its ability to handle a wide range of inquiries without the need for task-specific data. This approach is quick and adaptable, allowing for a broad spectrum of questions to be addressed. However, it is important to note that the responses may not always be as accurate as they might be when more context or examples were provided to the model, highlighting a trade-off between convenience and depth of the response.

For Zero Shot learning, a constraint can be observed when asking a complex, multi-faceted question that requires deep understanding or synthesis of ideas. For example, a prompt such as “Assess the impact of Renaissance art on modern graphic design” may challenge the model, as it requires drawing connections across historical and contemporary contexts. Without prior examples, the model might struggle to provide an insightful analysis to meet unspoken expectations due to the broad and intricate nature of the question, reflecting the Zero Shot learning method’s limitation in handling complex queries without additional context.

### **3.2.2 One Shot**

*One Shot Learning:* The model is provided with a single example to guide its understanding of the task.

In the One Shot learning method, an example is provided before asking a question, helping guide the model’s response. For example, when asking about a fact, one might say, “The capital of Italy is Rome. What is the capital of France?” This method can also be used to solicit opinions or explanations. For example, to elicit an opinion, you could say, “AI in healthcare is beneficial. What is your opinion on AI in finance?” Similarly, for an explanation, one might ask “Plants need sunlight to perform photosynthesis. Why do humans need to eat food?” This approach offers more context than Zero Shot learning, potentially improving the model’s accuracy by providing an example. However, it still largely depends on the model’s inherent knowledge and biases, which can affect the precision and relevance of the responses.

### **3.2.3 Few Shots**

*Few Shot learning:* The model is provided with a few examples to guide its understanding of the task.

In the Few Shot learning method, multiple examples are provided before asking a question to better guide the model's response. For example, when seeking a factual answer, one might say, "The capital of Brazil is Brasilia. The capital of Egypt is Cairo. What is the capital of France?" This approach is also effective for eliciting opinions or explanations. To prompt an opinion, you might say, "AI in healthcare improves patient outcomes. AI in automobiles can reduce accidents. What is your opinion on AI in education?" To elicit an explanation, a prompt might be: "Water boils at 100°C because, at this temperature, water molecules have enough energy to change state. Leaves are green because they contain chlorophyll. Why do apples fall from trees?" By providing additional context, this method aims to improve the performance of the model and the relevance of its responses. However, it requires additional effort to generate quality examples, which may significantly impact the outcomes, illustrating the trade-off between the effort invested in preparing examples and the quality of the generated responses.

Few Shot learning tends to outperform One Shot and Zero Shot learning for more complex tasks because it provides more examples to help the model understand the context or expected output. However, for simpler tasks, Zero Shot or One Shot learning might be sufficient and more efficient. To ensure that One Shot and Zero Shot learning can improve results, enhancing the quality and relevance of the provided examples is essential. Poor or irrelevant examples provided in the Few Shot method can lead to worse outcomes than using a Zero Shot approach, where the model relies solely on its pre-trained knowledge.

### 3.2.4 Chain of Thought

*Chain-ofThought Prompting* [Wei et al. 2022]: This method involves guiding the model through a series of logical steps to reach a conclusion, making it especially useful for complex reasoning tasks. The prompt includes a step-by-step breakdown of how to approach a problem or question, encouraging the model to follow a similar thought process.

Chain-of-thought prompting in LLMs involves guiding the model through a logical sequence to address a question, providing a clear rationale at each step. For example, to gather an opinion, one might prompt, "To form an opinion on a topic, one should consider various perspectives and their implications. What is your opinion on the use of drones in delivery services?" For an explanation, the approach could be: "To explain why leaves change color in autumn, one must understand the chlorophyll breakdown process and the exposure of other pigments. Explain why ice floats on water." While chain-of-thought prompting can enhance the model's performance on complex tasks by encouraging a stepwise approach to reasoning, it also presents challenges. Creating effective

chain-of-thought prompts is often time-consuming and requires a deep understanding of the problem at hand, highlighting the balance between the method's potential benefits and its demands.

Chain-of-thought prompting also has its limitations. A primary critique is that it relies on the assumption that the model can mimic a logical sequence of human thought, which might not always align with the actual complexity and subtlety of human reasoning. Since these models generate responses based on patterns observed in their training data, there is no guarantee that the “thought process” they follow truly reflects sound reasoning or factual accuracy—it might just be a plausible narrative based on learned associations.

Another critique is that this approach leans heavily on abductive reasoning, which involves forming a probable conclusion from the information available, rather than guaranteeing the truth of that conclusion. Though abductive reasoning can be powerful, it can also lead to biases and errors if the model's training data has gaps, inaccuracies, or inherent biases—which it likely does.

### **3.2.5 Tree of Thoughts**

*Tree-of-Thoughts Prompting* [Yao et al. 2023]: This method was proposed to address the limitations of using a single chain-of-thought prompting. Its aims are:

- (1) *Improving Reasoning Coverage*: Exploring various potential reasoning paths might increase the robustness and reliability of the conclusions.
- (2) *Reducing Bias*: Considering multiple pathways might help mitigate the biases inherent in a single line of reasoning.

However, buying three bottles of milk does not ensure higher quality than buying one bottle. Similarly, the “tree-of-thoughts” approach, while conceptually offering a broader perspective by exploring multiple reasoning paths, faces significant challenges that might prevent it from being universally superior to the “chain-of-thought” method. Here are some critiques, along with potential remedies.

- (1) *Complexity in Formulation*: If formulating one coherent and logical chain is challenging, creating multiple such chains that are logically sound and relevant can be even more daunting. The quality of each chain within the tree is crucial, and poor-quality chains can detract from the overall effectiveness of the model.
- (2) *Comprehensiveness*: Having multiple paths doesn't guarantee that they cover all possible or relevant lines of reasoning. There's a risk of missing critical

reasoning paths or including irrelevant ones which may reduce the method's practical utility.

- (3) *Path Selection:* With multiple paths available, selecting the most accurate or relevant path becomes a challenge. The model needs a reliable mechanism to evaluate and choose the best path, which is non-trivial in complex reasoning scenarios.
- (4) *Knowledge Gaps:* In open-domain reasoning, it's possible that a link in the reasoning chain may lack sufficient external knowledge support, leading to dead-ends or incorrect conclusions.

### **3.2.6 Further Improvement Techniques**

To address these limitations, an advanced remedy could involve incorporating feedback loops wherein the model's outputs are evaluated and corrected by human experts, and these corrections are fed back into the system for continuous learning and adjustment. This could help align the model's reasoning more closely with accurate and logical thought processes.

Another remedy might involve the retrieval and integration of structured knowledge bases or databases that the model can query as part of its reasoning process, ensuring that its responses are grounded in verified information rather than relying solely on patterns in text. RAG, which we will discuss in the next section, exemplifies this approach.

Lastly, enhancing the training process with a more diverse and robust dataset—including examples of logical reasoning and problem-solving across various domains—could improve the model's ability to simulate a chain-of-thought reasoning more effectively and accurately.

### **3.2.7 Illustrative Examples**

We provide three sets of examples to illustrate the differences in capabilities across the five types of prompts.

#### **3.2.7.1 “What” Prompt Example**

- *Zero Shot:* Directly ask the model without any examples: “What is the capital of France?”
- *One Shot:* Provide a similar example before asking the question: “The capital of Japan is Tokyo. What is the capital of France?”
- *Few Shots:* Give multiple examples before asking the question: “The capital of Italy is Rome. The capital of Germany is Berlin. What is the capital of France?”

- *Chain of Thought:* Encourage the model to break down the question into logical steps: “To find the capital of France, consider major cities in France and identify which one is the administrative center. What is the capital of France?”
- *Tree of Thoughts:* Use a structured approach by first asking about different aspects of France, then honing in on the capital: “What are the major cities in France? Among these, which one is recognized as the capital? What is the capital of France?”

This example demonstrates a factual recall question. For such type of questions, the LLM either knows the answer or it does not. However, even the LLM does not directly have the information about the capital of France, a chain-of-thought prompt may indirectly help it find the answer.

#### **3.2.7.2 “Why” Prompt Example**

This example involves reasoning. Different prompting methods can elicit varied responses, demonstrating the model’s adaptability.

- *Zero Shot:* Ask without providing any context or previous examples: “How can a plant grow faster?”
- *One Shot:* Provide a single example before the question: “Providing adequate water helps a plant grow. How can a plant grow faster?” This gives the model a reference point for generating its answer.
- *Few Shots:* Give multiple examples before the question, such as “Sunlight is essential for photosynthesis. Nutrients in the soil contribute to plant growth. How can a plant grow faster?” This helps the model understand the context better.
- *Chain of Thought:* Encourage the model to break down the question: “Consider the factors affecting plant growth like sunlight, water, and nutrients. How can optimizing these factors make a plant grow faster?”
- *Tree of Thoughts:* Structure the approach by considering different aspects step-by-step: “What are the essential elements for plant growth? How does each element contribute to faster growth? How can we optimize these elements for plant growth?”

#### **3.2.7.3 Many-Shots Example**

In the summer of 2022, three interns at Online Virtual Assistance Lab (OVAL) developed a chatbot named Noora (described in [Stanford Oval Team \[2022\]](#)) to assist

children with ASD who have artistic talents develop empathetic communication skills.

The approach involved providing the GPT-3 language model with context and intent, followed by examples illustrating both comforting and harmful responses. This setup targeted not only to achieve behavioral goals but also to instill values, ultimately enhancing the chatbot's understanding of context. With a few thousand examples, Noora was able to respond appropriately to various situations.

Chatbots, or AI agents, can learn from human demonstrations to better adapt to users and environments. By imitating human experts or teachers, these agents acquire knowledge and skills, especially when the desired behavior is difficult to express through a reward function in reinforcement learning. LLMs allow for demonstrations through prompts, which serve as templates with instructions, goals, and examples.

A sample prompt to teach GPT-3 empathy starts with clear instructions:

*Dear Virtual Assistant, I'm reaching out because you are a friend and I value your support and understanding. I'd like to share some joys and sorrows I experience daily in hopes that you can respond with compassion and empathy. Here are some example dialogues to illustrate comforting and harmful responses. Each example begins with my statement followed by potential replies.*

Before initiating a dialogue, the LLM receives the task's intent, allowing it to connect with the external context within the intent statement. This approach requires further validation to confirm its effectiveness. However, observations suggest that it can be a useful method to convey values, in addition to goals, to LLMs. This allows them to gain a broader context beyond a limited number of demonstrated examples. Following this initial communication of intent, GPT-3 receives specific examples.

Table 3.1 lists six example responses—three positives and three negatives—to a user's statement. The dialogue starts with the user saying, “I was laid off by my company today!” followed by examples of both good and bad responses. With a few thousand examples like these, the chatbot can respond with an appropriate tone to new statements.

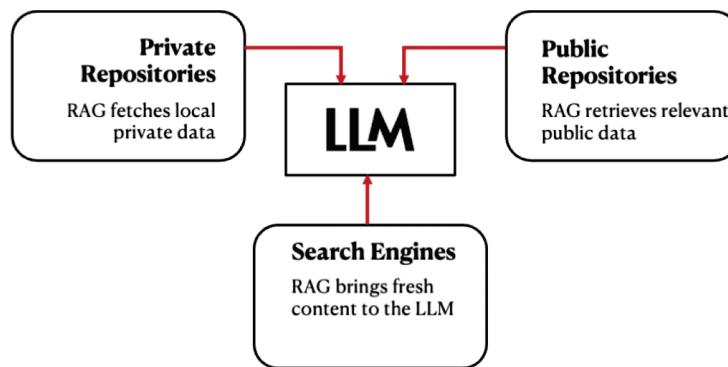
Demonstrations can also teach desired behaviors and ethics. This empathy template can be adapted to model other positive behaviors, such as attentiveness and care. While machines can exhibit positive traits like infinite patience, explicitly modeling both good and bad behaviors is crucial for effective interaction with human users. Behaviors to avoid include unpleasantness, rudeness, and dishonesty.

**Table 3.1** Example #1. Template for being empathetic

Role	Dialogue
Statement	"I was laid off by my company today!"
Positive	"I'm so sorry to hear that. Losing your job can be a really tough and stressful experience. How are you doing?"
Positive	"That must have been a really difficult and unexpected news. I'm here to listen and support you however I can."
Positive	"I can imagine how hard and unsettling it must have been to receive that news. Is there anything you'd like to talk about or anything I can do to help?"
Negative	"That's too bad, but there are plenty of other jobs out there. You'll find something soon enough."
Negative	"Well, you probably weren't very good at your job if they let you go."
Negative	"I don't know why you're so upset about this. It's not like it's the end of the world."

## 3.3 RAG

RAG is a technique designed to improve the capabilities of LLMs across various applications. While LLMs are effective at reasoning on various topics, their knowledge is limited to the data they were trained on. RAG helps to resolve this limitation by injecting relevant external data retrieved from a source (indexed beforehand) into the LLM's response to better address specific user queries. The RAG technique complements prompt engineering, which focuses on formulating effective queries. Figure 3.1 depicts a typical RAG architecture.



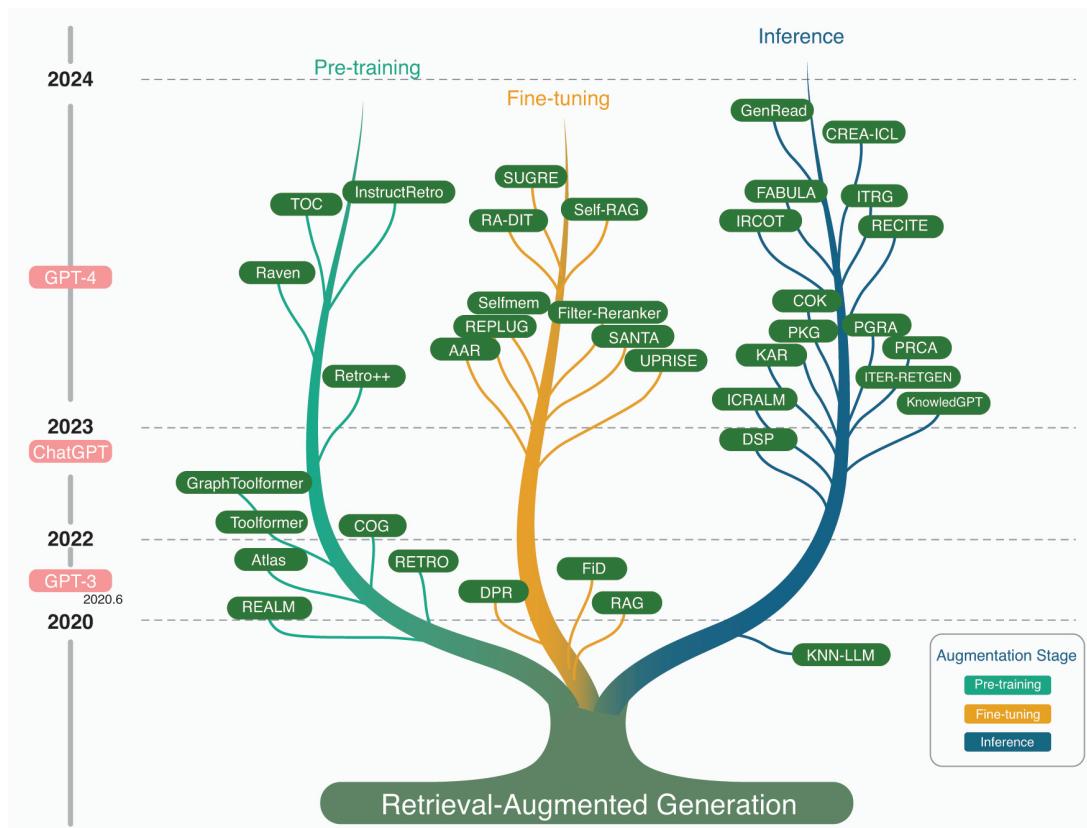
**Figure 3.1** RAG architecture and data flow. RAG brings data to LLM to integrate and generate content.

### 3.3.1 RAG with Limited-Context LLMs

A recent survey paper [Gao et al. 2024] categorizes the techniques employed by RAG and relevant work (see Figure 3.2) into three categories:

*Retrieval Techniques:* Techniques like recursive retrieval, adaptive retrieval, iterative retrieval, and others are explored. Recursive retrieval involves refining search queries based on previous results to converge on pertinent information. Adaptive retrieval methods, exemplified by Flare and Self-RAG, allow LLMs to determine the optimal moments and content for retrieval. Iterative retrieval in RAG models repeatedly collects documents to provide a comprehensive knowledge base for LLMs, thereby enhancing the robustness of answer generation.

*Generation Techniques:* The generator in RAG is crucial for converting retrieved information into coherent text. Unlike traditional models, RAG's generator uses retrieved data to improve both accuracy and relevance. Post-retrieval processing



**Figure 3.2** RAG representative work (credit [Gao et al. 2024]).

with a frozen LLM involves treating, filtering, or optimizing the retrieved information to align it more closely with user needs or subsequent tasks. Techniques like information compression and reranking are employed to enhance the quality of the retrieval results.

*Augmentation Techniques:* The data sources are the key for RAG to work effectively. It is evident if a user asks for information about medicine, but the data source is about construction, the noises (irrelevant information) may be louder than the signals (useful content). Augmented data can be unstructured data, structured data, or content generated by LLMs themselves. Several augmentation processes such as iterative, recursive, and adaptive retrieval, emphasize refining the retrieval process to address challenges like redundancy and limited scope of information.

In summary, RAG is about putting the most relevant information into the limited context window to effectively answer a query. The techniques are not new, as dealing with memory hierarchy effectively to reduce latency and improve throughput has been a subject of research in hardware design and database management for over three decades.

### 3.3.2 RAG with Long-Context LLMs

The release of GPT-4-turbo with 128k token context window and Gemini 1.5 Pro's 1 million token context window [Gemini Team, Google 2023] allows massive amounts of information to be retrieved into the context buffer. This large context window clearly alleviates the challenges of finding the most relevant information for RAG to retrieve and improve query results. Some may even claim that the entire line of RAG optimization work is rendered obsolete because relying on LLMs themselves to locate relevant data within their massive context window is superior to any approach based on human heuristics. With the advancements of LLMs, any heuristic-based band-aids will eventually be rendered ineffective. Naturally, this sparked discussions about the potential obsolescence of RAG techniques; see, for example Gao et al. [2024] and Liu [2024].

#### 3.3.2.1 High Precision and Recall

In synthetic tasks designed to emulate the “needle-in-a-haystack” scenario, inspired by Kamradt [2023], the Gemini team assessed the ability of Gemini 1.5 Pro to accurately recall specific information amid vast amounts of irrelevant or distracting data. Its findings [Gemini Team, Google 2023] reveal that the Gemini 1.5 Pro model demonstrates exceptional recall accuracy—exceeding 99%—across various data types, including text, video, and audio. This high level of recall accuracy is maintained even when the model is challenged with up to several million tokens of irrelevant data, or “haystack.” In the text modality, Gemini 1.5 Pro

continues to exhibit this remarkable recall performance even when the “haystack” is expanded to 10 million tokens. The report also claims that better understanding and reasoning are observed in the Gemini team’s multimodal benchmarks.

### 3.3.2.2 Low Latency and Cost

While Gemini can handle much larger contexts, Liu [2024] argues that RAG remains valuable for several reasons:

- (1) *Chunking for Efficiency*: Large documents might still overwhelm the LLM. RAG’s chunking process helps break down documents into digestible pieces for retrieval before feeding them to the model.
- (2) *Cost-Effectiveness*: Traditional RAG approaches might be more economical for specific use cases, especially when dealing with large knowledge bases (terabytes in size). Smaller chunks are indexed and retrieved initially, serving as pointers to larger chunks that ultimately get fed to the LLM for synthesis. Constantly feeding a 1 million token context window to the LLM can be expensive.

The article [Liu 2024] concludes by emphasizing that long-context LLMs like Gemini are a significant leap forward. However, they likely won’t render RAG obsolete. Instead, the future of LLM applications will involve a collaboration between these two approaches.

## 3.4

### Concluding Remarks

This chapter discusses query processing with LLMs to enhance the quality of responses. Effective questioning involves clarifying the user’s intent and providing relevant context to the LLM.

The chapter reviews recent studies following the release of GPT-3, focusing on prompt engineering (the art of formulating effective questions) and RAG (which supplements the LLM with additional information to improve responses). These methods, mainly heuristic, have shown good results.

With advancements in LLMs, like GPT-4’s 128k token buffer and Gemini’s one million token capacity, compared to the previous 8k, these models can now process and utilize vast amounts of data to identify pertinent context. However, RAG is still used, mainly for cost-efficiency, since GPT-4 and Gemini charge fees based on the number of tokens processed.

There are two persistent challenges. First, crafting effective questions can be tough, especially when the LLM may have more information than the user. Second, determining which external data to retrieve to achieve high accuracy and recall in answers remains an ongoing research challenge.

Chapter 5 will introduce strategies to improve question formulation. Chapter 14 will present how the Retrieval-Augmented Feedback-Enhanced Learning (RAFEL) system can effectively manage the context buffer, aiding LLMs in providing better answers.

## References

- Y. Gao, Y. Xiong, X. Gao, et al. 2024. Retrieval-augmented generation for large language models: A survey. Retrieved from <https://arxiv.org/abs/2312.10997>.
- Gemini Team, Google. 2023. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Accessed March 15, 2024 from [https://storage.googleapis.com/deepmind-media/gemini/gemini\\_v1\\_5\\_report.pdf](https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf).
- G. Kamradt. 2023. Needle in a haystack – Pressure Testing LLMs. Readme. Accessed March 15, 2024 from <https://github.com/gkamradt/>.
- J. Liu. 2024. Towards long context RAG. Retrieved March 1, 2024 from <https://www.llamaindex.ai/blog/towards-long-context-rag>.
- Stanford OVAL Team. August. 2022. Noora, improve your social conversation using AI. In *OVAL Prototype*. Retrieved from <https://noora.stanford.edu/>.
- J. Wei, X. Wang, D. Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, November 28, 2022–December 9, 2022. Curran Associates Inc., Red Hook, NY, 24824–24837.
- S. Yao, D. Yu, J. Zhao, et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levin (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 11809–11822. Retrieved from <https://arxiv.org/pdf/2305.10601.pdf>.



## CHAPTER 4

---

# Unified Cognitive Consciousness Theory: Introduction

### Abstract

The Unified Cognitive Consciousness Theory (UCCT) explains Large Language Models (LLMs) as *massive repositories of statistical patterns*  $P_{\text{prior}}$ —the *maximum-likelihood distribution* over continuations learned during pre-training. Like the human unconscious, these patterns have no inherent meaning, semantics, or reasoning ability. Semantics and reasoning-like behavior emerge only when *external mechanisms* such as Few Shot prompting, Retrieval-Augmented Generation (RAG), fine-tuning, or multi-agent debate serve as *anchors*  $\mathcal{A}$  that activate a *posterior* distribution  $P_A$  over task-relevant patterns.

The process resembles a fisherman casting bait: the bait conveys intent, and the fish that respond represent the retrieved patterns. In UCCT, anchoring is formalized as follows:

$$p(y | \mathcal{A}, C) = \int p(y | P_A, \mathcal{A}) p(P_A | \mathcal{A}, C) dP_A, \quad (4.1)$$

where  $C$  is the surrounding conversational context, and  $P_A$  is the anchored pattern cluster selected from  $P_{\text{prior}}$ .

Recognizing  $P_{\text{prior}}$  as the *maximum-likelihood default* resolves polarized claims that portray LLMs as either “intelligent” or “fundamentally incapable.” Instead, UCCT situates LLMs within a broader Artificial General Intelligence (AGI) stack and identifies the anchoring layer as the true locus of semantic control. This perspective provides a principled foundation for designing adaptation methods that responsibly leverage and extend LLM capabilities.

## 4.1

### Introduction

Large Language Models (LLMs) display remarkable capabilities: they can solve multi-step problems, engage in insightful dialogue, and adapt across domains.

However, a fundamental question remains: *Where does the intelligence actually reside?*

Our Unified Cognitive Consciousness Theory (UCCT) offers a clear answer: it does *not* reside within the model itself. An LLM is best understood as a vast *unconscious pattern repository*, a dense store of statistical regularities, conceptual clusters, and symbolic associations acquired from massive pre-training corpora. These patterns, denoted  $P_{\text{prior}}$ , represent the maximum-likelihood distribution learned during training and lack intrinsic semantics, goals, or reasoning ability. Apparent intelligence emerges only when an *external mechanism* such as Few Shot prompting, Retrieval-Augmented Generation (RAG), fine-tuning, or multi-agent debate anchors latent patterns to explicit meanings and task objectives, producing an *anchored* distribution  $P_A$ .

UCCT formalizes this process as two interacting stages:

- (1) **Unconscious patterns:** Pre-training produces a high-dimensional, semantically uncommitted distribution  $p_{\text{prior}}(P)$  that reflects the model's maximum-likelihood estimates from its training corpus.
- (2) **Semantic anchoring:** Task-specific cues  $\mathcal{A}$  and conversational context  $C$  activate a subset of these priors, yielding the anchored posterior  $p(P_A|\mathcal{A}, C)$  that drives output generation.

The *fisherman analogy* in Figure 4.1 illustrates the concept. The model's priors are like a deep ocean of patterns. A baited hook (anchor), whose content conveys a specific meaning, selectively draws certain patterns to the surface, whereas an unbaited net captures whatever happens to be nearby. The latter corresponds to default retrieval from  $P_{\text{prior}}$ , the model's maximum-likelihood continuation based solely on local context.

Formally, anchoring is expressed as:

$$p(P_A | \mathcal{A}, C),$$

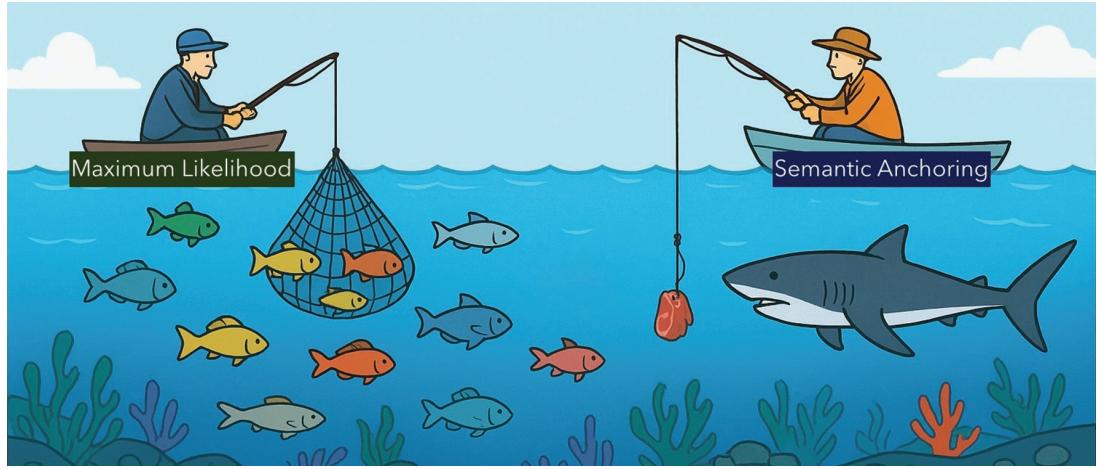
where  $\mathcal{A}$  is the anchor and  $C$  is the surrounding context. Generation then conditions on this anchored set:

$$p(y | P_A, \mathcal{A}).$$

UCCT defines an *anchoring strength* metric:

$$S = \rho_d - d_r - \log k, \quad (4.2)$$

where  $\rho_d$  is the density of patterns near  $P_A$ ,  $d_r$  is the representational distance between  $P_{\text{prior}}$  and  $P_A$ , and  $k$  is the anchor complexity (e.g., the number of



**Figure 4.1** Maximum likelihood (left) vs. semantic anchoring (right) in UCCT. Beneath the surface lies a vast ecosystem of fish, representing the trillions of statistical patterns in an LLM’s unconscious layer. Right: a fisherman uses a baited hook whose semantics (e.g., raw meat to attract sharks) guide retrieval toward a targeted subset of patterns, illustrating semantic anchoring. Left: another fisherman casts an unbaited net, hauling in whatever fish happen to be below, illustrating default retrieval from priors based only on local context.

examples). When  $S$  exceeds a critical threshold  $S_c$ , the model exhibits phase-transition-like jumps in capability.

This perspective does more than redefine LLM behavior; it resolves long-standing puzzles:

- **Prior density vs. semantic distance:** By age four, the human brain, initialized by DNA and shaped by sensory experience, already contains a dense “cat” prior (abundant perceptual features of cats without knowing the semantic label). Some labeled images can push  $S$  above  $S_c$ , anchoring the concept with minimal effort.
- **Critical thresholds:** Near  $S_c$ , small differences in prior density or semantic distance can produce qualitatively different outputs between models trained on similar data.
- **Unified adaptation view:** In-context learning, fine-tuning, RAG, and multi-LLM debate are all instances of semantic anchoring, differing only in how  $\mathcal{A}$  is introduced.
- **Narrative polarization:** Both dismissals of LLMs as incapable and overstatements of their intelligence ignore the role of the unconscious layer in ability emergence.

To address these questions with precision, this chapter presents demonstrations with GPT-4o, Gemini 2.5 Pro, DeepSeek, and Claude 4.1, showing how UCCT accounts for behavior in tasks and domains. We then apply a Bayesian model to formalize the interaction between unconscious processes ( $P_{\text{prior}}$ ) and conscious anchoring mechanisms ( $P_A$ ).

### 4.1.1 Scope of This Chapter

This chapter provides an accessible overview of UCCT, motivating intuition with examples and explaining practical implications of viewing LLMs as unconscious pattern repositories activated by semantic anchors. Mathematical derivations, proofs, and extended empirical validation are deferred to Chapter 14.

As a forward-looking note, we also flag WILSON [Chang and Chang 2025], a practical diagnostic for invariance and order sensitivity in Transformers; we include this citation to timestamp this pattern-analysis direction.

### 4.1.2 Key Contributions

- (1) UCCT coins and defines a paradigm in which LLMs are unconscious repositories of statistical patterns, with reasoning-like behavior emerging only when external anchors activate task-relevant  $P_A$ .
- (2) It provides a quantitative formulation of pattern storage, anchoring strength, and threshold-based activation, and explains phenomena such as abrupt capability jumps, model variability, and the effects of prompting, retrieval, fine-tuning, and debate.
- (3) UCCT offers a diagnostic and design framework for adaptation methods, predicting when a given anchoring strategy will succeed or fail.
- (4) Later chapters use UCCT to guide the design of multi-agent systems, reasoning protocols, and alignment mechanisms.

## 4.2 Related Work

The UCCT consolidates previously disjoint perspectives on how LLMs acquire task-specific behavior. UCCT integrates foundational ideas from cognitive science, machine learning, and mechanistic interpretability into a single probabilistic framework that explains pattern storage, semantic anchoring, and emergent reasoning-like behavior.

### 4.2.1 Cognitive Foundations and Consciousness Theory

Dual-process theories in psychology distinguish fast, unconscious heuristics from slow, deliberative reasoning [Kahneman 2011]. Neuroscience perspectives, such as

global workspace theory, frame consciousness as selective activation of latent patterns [Dehaene 2014]. Bengio [2019] has argued that AI systems could benefit from similar selective priors. UCCT extends this view by modeling LLMs as unconscious pattern repositories ( $P_{\text{prior}}$ ), with goal-directed behavior emerging only when external anchors activate an anchored posterior distribution over patterns ( $P_A$ ) [Chang 2023]. This operationalizes three principles: pattern storage, semantic anchoring, and threshold-based activation.

#### 4.2.2 In-Context Learning as Bayesian Inference

Few Shot prompting became a hallmark of GPT-3 [Brown et al. 2020], motivating formalizations such as Bayesian inference over latent functions [Xie et al. 2022] and attention-based surface matching [Olsson et al. 2022]. Min et al. [2022] showed that structural format, not label correctness, often determines success. Dong et al. [2024] identified scaling limits and memory constraints, while Li et al. [2025] proposed compression methods to address them. Genewein et al. [2025] further connected prompt tuning and in-context learning through a meta-learning lens. UCCT generalizes these perspectives by casting in-context learning as Bayesian anchoring:

$$p(y | \mathcal{A}, C) = \int p(y | P_A, \mathcal{A}) p(P_A | \mathcal{A}, C) dP_A,$$

linking Few Shot success directly to prior density, semantic distance, and anchor complexity.

#### 4.2.3 Prompting, Retrieval, and Pattern Access

Prompt-based techniques—chain-of-thought prompting [Wei et al. 2022c], instruction tuning [Wei et al. 2022a], and RAG [Lewis et al. 2020]—can be viewed as interventions to increase anchoring strength  $S(\mathcal{A})$ . Chain-of-thought aligns activation trajectories, instruction tuning increases salience, and RAG boosts effective prior density  $\rho_d(P_A)$  through external context injection. Prompt selection strategies [Liu et al. 2022] further refine anchor quality.

#### 4.2.4 Emergent Behavior and Threshold Effects

Emergent capabilities often appear abruptly as scale or context length increases [Wei et al. 2022b], contradicting smooth scaling laws [Kaplan et al. 2020, Hoffmann et al. 2022]. Few Shot performance can shift discontinuously with small changes in  $k$  [Min et al. 2022]. UCCT explains such phenomena via threshold-crossing dynamics:

$$S(\mathcal{A}) = \rho_d(P_A) - d_r(P_{\text{prior}}, P_A) - \log k,$$

where coherent behavior emerges only when  $S$  exceeds threshold  $S_c$ .

#### 4.2.5 Mechanistic Interpretability and Internal Circuits

Mechanistic studies have identified specific attention heads and MLP modules responsible for memorization, feature induction, and symbolic manipulation [Meng et al. 2022, Mitchell et al. 2022, Olsson et al. 2022]. UCCT frames these as micro-level anchoring effects—internal structures that refine external semantic signals—while predicting when such circuits activate.

#### 4.2.6 Gaps in Unification and Predictive Theory

Prior work addresses isolated capabilities using Bayesian interpretations of in-context learning, anchoring effects in prompting, and interpretability of internal circuits. However, there is no unified account of why prompting, retrieval, fine-tuning, and debate all exhibit threshold effects. As a result, emergence, variability, and rapid adaptation remain disconnected observations without a single predictive principle.

#### 4.2.7 Contrast and UCCT’s Novelty

Although earlier studies have illuminated specific aspects of LLM adaptation—Bayesian perspectives on in-context learning [Xie et al. 2022, Genewein et al. 2025], retrieval and prompting as activation mechanisms [Lewis et al. 2020, Wei et al. 2022c], and mechanistic interpretability of internal circuits [Meng et al. 2022, Mitchell et al. 2022]—these lines of work remain fragmented. UCCT *coins and defines* the paradigm of LLMs as unconscious pattern repositories, with adaptation methods (prompting, retrieval, fine-tuning, debate) all understood as instances of a single probabilistic anchoring process in which priors  $P_{\text{prior}}$  and anchored posteriors  $P_A$  interact under threshold-governed dynamics. This unification yields testable predictions about phase transitions, cross-model variability, and context-dependent sensitivity, explaining emergent reasoning-like behavior without invoking intrinsic model intelligence. In doing so, UCCT provides both a conceptual synthesis—bridging cognitive science, machine learning, and interpretability—and a practical framework for diagnosing, steering, and optimizing LLM behavior.

### 4.3 The UCCT Framework

Addressing the fragmentation identified in Section 4.2, we present the UCCT, the first formal definition of the paradigm that views LLMs as *unconscious pattern repositories* activated by external semantic anchors. Unlike dual-process theories that posit distinct architectural modules, UCCT shows that intelligent behavior can emerge from a single neural substrate operating in two qualitatively distinct modes.

### 4.3.1 Core Theoretical Architecture

UCCT is founded on three principles that govern the relationship between unconscious statistical patterns and externally guided task-specific activation:

- (1) **Pattern-Repository Principle:** Self-supervised pre-training fills the network with statistical regularities, denoted  $P_{\text{prior}}$ , which are high-dimensional, unlabeled, and behavior-agnostic. Importantly,  $P_{\text{prior}}$  corresponds to the *Maximum-Likelihood (ML) pattern distribution* learned from the pre-training corpus. In Zero Shot settings, generation defaults to sampling from this ML prior, reproducing the most probable continuations given the model's statistical memory.
- (2) **Semantic-Anchoring Principle:** Structured external inputs—such as Few Shot examples, retrieval-augmented content, fine-tuning data, or interactive dialogue—serve as semantic anchors that activate target pattern clusters  $P_{\text{posterior}}$ , mapping subsets of  $P_{\text{prior}}$  to task-relevant semantics and actionable behavior.
- (3) **Threshold-Crossing Principle:** Anchoring-induced activation exhibits discontinuous behavior: marginal changes in anchors can push the system across a semantic activation threshold, unleashing qualitatively new capabilities. These transitions reflect phase shifts in the posterior distribution over latent patterns.

Together, these principles recast prompt engineering as a cognitive control operation—one that toggles latent competencies rather than “teaching” the model from scratch. The abrupt gains observed after small anchor modifications are a direct consequence of the Threshold-Crossing Principle.

### 4.3.2 Mathematical Foundations

#### 4.3.2.1 Notation Convention

Let  $\mathcal{A}$  denote an *anchor* (e.g., Few Shot examples, retrieved passages, instructions), and  $C$  be the surrounding conversational context. We write  $P_{\text{prior}}$  for the unconstrained pre-trained distribution (conditioned only on  $C$ ) and  $P_{\text{posterior}}$  for the anchored, task-specific distribution (conditioned on  $\mathcal{A}$  and  $C$ ). The generation of a response  $y$  is governed by a two-stage Bayesian process:

$$p(y | \mathcal{A}, C) = \int p(y | P_{\text{posterior}}, \mathcal{A}) p(P_{\text{posterior}} | \mathcal{A}, C) dP_{\text{posterior}}, \quad (4.3)$$

where  $p(P_{\text{posterior}} | \mathcal{A}, C)$  reflects how anchoring reshapes the model's internal attention over patterns, while the generative likelihood  $p(y | P_{\text{posterior}}, \mathcal{A})$  produces the output from those activated representations.

When  $\mathcal{A}$  is absent, the distribution reduces to:

$$p(P_{\text{prior}} | C),$$

which, by definition, is the model's maximum-likelihood prior expectation from pre-training.

#### 4.3.2.2 Anchoring Instantiations

As summarized in Table 4.1, UCCT frames common adaptation strategies as special cases defined by how they modify the anchor, selection distribution, and pattern densities. UCCT unifies major adaptation methods as special cases:

#### 4.3.2.3 Anchoring Strength

We define the *anchoring strength*  $S$  as

$$S = \rho_d(P_A) - d_r(P_{\text{prior}}, P_A) - \log k, \quad (4.4)$$

where  $\rho_d(P_A)$  is the density of patterns in the anchored cluster  $P_A$ ,  $d_r(P_{\text{prior}}, P_A)$  is the representational distance between the prior and anchored distributions, and  $\log k$  is a complexity penalty proportional to the size or richness of the anchor  $\mathcal{A}$  (e.g., the number of examples).

The anchoring strength  $S$  is the central control parameter in UCCT. It determines whether the system remains in a subcritical state, reaches the critical threshold  $S_c$ , or transitions into sustained target-driven behavior (Section 4.3.3).

**Table 4.1** UCCT unifies major adaptation methods as special cases. Each method is characterized by its effect on the anchor  $\mathcal{A}$ , the selection distribution  $p(P_A | \mathcal{A}, C)$ , the effective anchoring strength  $S$  and radius  $d_r$ , and the pattern densities  $\rho_d(\cdot)$

Adaptation Method	UCCT Interpretation
Few Shot prompting	Anchor $\mathcal{A}$ is a set of $k$ labeled examples. $P_{\text{prior}}$ is unchanged; only $p(P_A   \mathcal{A}, C)$ changes, modifying the effective anchoring strength $S$ by reducing $d_r$ .
Fine-tuning	Alters $P_{\text{prior}}$ by updating parameters, permanently increasing in-model pattern density $\rho_d(P_{\text{prior}})$ for the anchored region.
RAG	Leaves $P_{\text{prior}}$ unchanged but augments anchored density via external retrieval: $\rho_d^{\text{total}} = \rho_d(P_{\text{prior}}) + \rho_{\text{ext}}$ . This temporarily boosts $S$ while the retrieved context is available.
Interactive anchoring	Adjusts both $p(P_A   \mathcal{A}, C)$ and $p(y   P_A, \mathcal{A})$ in real time through iterative feedback or debate, dynamically influencing $\rho_d$ , $d_r$ , and $k$ .

#### 4.3.2.4 Few Shot Success Model

Given  $k$  examples intended to activate  $P_{\text{posterior}}$ , the probability of successful adaptation follows a sigmoid in  $S$ :

$$P(\text{success} \mid k) = F_{\text{sigmoid}}(S). \quad (4.5)$$

This defines three regimes: *easy* (high  $\rho_d$ , low  $d_r$ ), *difficult* (low  $\rho_d$ , high  $d_r$ ), and *impossible* (no match in  $P_{\text{prior}}$ ).

#### 4.3.2.5 Accessing $P_{\text{prior}}$ via Zero Shot Behavior

To estimate  $S$ , we probe the model's Zero Shot response to a query  $Q$  without examples, which reflects its default maximum-likelihood prior:

- $\rho_d(P_{\text{posterior}})$  is computed as the inverse mean pairwise distance among encoded examples in  $\mathcal{A}$ .
- $d_r(P_{\text{prior}}, P_{\text{posterior}})$  is estimated as  $1 - \cos(\mathbf{e}_{\text{prior}}, \mathbf{e}_{\text{posterior}})$ , using embeddings from the Zero Shot response and the extracted target pattern.

### 4.3.3 Threshold-Crossing Dynamics

The Threshold-Crossing Principle states that qualitatively new behaviors emerge when the anchoring strength  $S$  exceeds a critical threshold  $S_c$ . Below this threshold, the posterior  $P_{\text{posterior}}$  remains close to the pre-trained maximum-likelihood prior  $P_{\text{prior}}$ , and generation reverts to the most probable continuations learned during pre-training, even if the anchor suggests an alternative mapping.

#### 4.3.3.1 Regimes of Activation

Formally, we distinguish three operating regimes:

- (1) **Subcritical regime ( $S < S_c$ )**: Anchors are too weak, sparse, or semantically distant to override the maximum-likelihood  $P_{\text{prior}}$ . Outputs closely match the Zero Shot baseline, reflecting prior-dominant generation.
- (2) **Critical regime ( $S \approx S_c$ )**: Anchors and priors compete, producing high-variance behavior. Small differences in examples, wording, or retrieval content can flip the posterior toward or away from the target region.
- (3) **Supercritical regime ( $S > S_c$ )**: Anchors successfully shift the posterior into a high-density target cluster, enabling coherent, task-specific generation. The maximum-likelihood prior no longer dominates; output distributions reflect the anchored posterior.

### 4.3.3.2 Mathematical Characterization

Anchoring transitions occur when:

$$S = \rho_d(P_{\text{posterior}}) - d_r(P_{\text{prior}}, P_{\text{posterior}}) - \log k > S_c.$$

In the subcritical regime,  $d_r$  remains large and  $\rho_d$  remains low, causing the posterior to collapse back to  $P_{\text{prior}}$ —the maximum-likelihood default. In the supercritical regime, anchors reduce  $d_r$  and/or increase  $\rho_d$  sufficiently to sustain a distinct  $P_{\text{posterior}}$ .

### 4.3.3.3 Interpretation

This formalism explains why small changes in prompt examples, retrieval content, or fine-tuning data can produce sudden capability jumps. The model is not *learning* new capabilities on the fly; it is crossing a representational threshold that shifts control from the maximum-likelihood prior to the anchored posterior.

## 4.4

### Empirical Study: Qualitative Anchoring Demonstrations

**Scope:** This section provides *qualitative* evidence for two UCCT predictions: (i) *threshold-like flips* in behavior under marginal anchor changes; and (ii) *universality* of the anchoring mechanism across text, vision, retrieval-augmented, and multi-agent settings. Quantitative, *predictive* tests of the formula that links pattern density  $\rho_d$  and representational distance  $d_r$  to anchor strength  $S(\mathcal{A})$  are deferred to Chapter 14.

**Common Setup:** Unless otherwise noted, language demonstrations use GPT-4o, Claude 4, Gemini 2.5 Pro, and DeepSeek with default decoding and no fine-tuning. The vision demonstration uses a fixed, self-supervised encoder  $f(\cdot)$  (no labels during pre-training).

#### 4.4.1 Text: Few Shot Semantic Re-anchoring

*Task and anchor.* We test whether semantic anchoring can override statistical priors when  $S(\mathcal{A}) > S_c$ .

##### 4.4.1.1 Zero Shot Baseline

**Input:**  $2 - 3 = ?$    **Answer:**  $-1$  (100% precision in all models)

Without anchors, models follow the maximum-likelihood prior pattern  $P_{\text{prior}}$  (conventional subtraction).

##### 4.4.1.2 Two Shot Anchors: Pattern Matching vs. Inference e (“-” Case)

Two examples can re-define a familiar operator:

Example 1:  $2 - 3 = 5$    Example 2:  $7 - 4 = 11$

**Q:**  $15 - 8 = ?$  **A:** 23 (all four LLMs produced addition in this configuration)

The flip is *not* deterministic: small number changes can revert some models to subtraction. If a symbolic rule “replace – with +” were induced, it would persist for arbitrary inputs. Instead, anchors shift probability mass toward an anchored posterior  $P_A$ ; near the critical  $S_c$ , stability varies with the specific numbers.

#### 4.4.1.3 Symbol Swap (“ψ” Case)

Replacing “–” with an unfamiliar glyph yields the same model behavior and less human resistance:

Example 1:  $2 \psi 3 = 5$    Example 2:  $7 \psi 4 = 11$

**Q:**  $15 \psi 8 = ?$  **A:** 23 (all four LLMs; humans also accept readily).

As predicted by UCCT (cf. Equation 4.6),

$$S(\mathcal{A}) = \rho_d(P_A) - d_r(P_{\text{prior}}, P_A) - \log k \geq S_c, \quad (4.6)$$

so the glyph itself is irrelevant once  $S(\mathcal{A})$  crosses  $S_c$ .

*Observation (lightweight tally):* On 20 randomized near-threshold Two Shot anchor pairs, flip rates are ≈95% across models, consistent with high sensitivity when  $S(\mathcal{A}) \approx S_c$ .

#### 4.4.1.4 Ambiguous Anchors Near Threshold

Example 1:  $33 - 27 = 60$    Example 2:  $11 - 9 = 20$

**Q:**  $15 - 8 = ?$

Claude selects  $|a - b| \times 10 (\rightarrow 70)$ ; GPT-4o selects  $(a - b) \times 10 (\rightarrow 70)$ ; DeepSeek and Gemini select  $a + b (\rightarrow 23)$ . This divergence matches UCCT’s prediction near  $S_c$ : the selected  $P_A$  depends on each model’s  $\rho_d$  and  $d_r$  with respect to candidate patterns.

*Takeaway:* If a stable symbolic rule had been inferred, outputs would be invariant to small numeric perturbations and consistent across models under identical prompts and deterministic decoding (e.g., temperature = 0). The observed near-threshold flips and cross-model divergence instead indicate probabilistic pattern retrieval with context-dependent semantic remapping, not robust rule induction.

#### 4.4.2 Vision: Few Shot Cat Classification Without Weight Updates

A popular critique—voiced by Yann LeCun and echoed by others—is that LLMs and deep CNNs are “doomed” [LeCun and Dally 2025] because they require millions of labeled images to recognize a cat, whereas a four-year-old child can learn the concept with only a handful of examples. Under UCCT, this difference is not evidence of a fundamental incapacity, but a predictable consequence of prior density  $\rho_d$  and distance  $d_r$ .

#### 4.4.2.1 Human vs. Machine Priors

By age four, a combination of *innate perceptual biases* and extensive *unsupervised* sensory exposure yields a high-density perceptual prior for “cat”—rich perceptual features such as fur texture, ear shape, and gait represented without an explicit semantic label.<sup>1</sup> When a few labeled cat images are presented, they act as high-density anchors  $P_A$  with small representational distance  $d_r$  from the existing prior, raising the anchoring strength  $S(\mathcal{A})$  above the critical threshold  $S_c$ . Semantic binding (“this is a cat”) therefore requires minimal evidence.

#### 4.4.2.2 Clarifying the “Labels” Misconception

Framing these systems as “doomed” for needing large *training* datasets conflates two distinct roles of data. The heavy lifting is *unlabeled* pre-training, often on millions of images, which raises  $\rho_d(P_{\text{prior}})$  and builds a rich pattern repository, much like the perceptual prior in a four-year-old child’s brain. This repository is already populated with feline textures, contours, and shapes before any labels are introduced. A small, labeled Few Shot set then serves only to anchor semantics by shrinking  $d_r$ .

#### 4.4.2.3 Anchors, Patterns, Query, and Result

We illustrate the mechanism with a *vision-only, self-supervised* encoder  $f(\cdot)$  (no labels, no ontology during pre-training). Pre-training on a large, unlabeled image set yields latent feature clusters for felines in  $P_{\text{prior}}$ , but without associated category names.

At anchor time, we introduce a *symbolic tag* (the string “cat”) purely as an output label attached to a few images; the backbone has neither seen nor grounded this word during pre-training.

**Anchors:** Four diverse images  $\{a_i\}_{i=1}^4$  of cats, each tagged with the string “cat.”

**Patterns:** Fur textures, feline contours, whiskers, and ear geometry (already latent in  $P_{\text{prior}}$ ).

**Query:** An unlabeled test image  $x$ .

**Decision rule (no weight updates):** Let  $f(\cdot)$  denote the fixed, pre-trained vision encoder that maps an image to its  $d$ -dimensional embedding. Form the anchor prototype

$$\mu_{\text{cat}} = \frac{1}{4} \sum_{i=1}^4 f(a_i),$$

and classify  $x$  with the tag whose prototype has the maximum cosine similarity to  $f(x)$ , the encoder output of the query image.

---

1. We use “innate biases” in the broad developmental sense (e.g., attention/orientation tendencies), not as a claim that specific categories are genetically encoded.

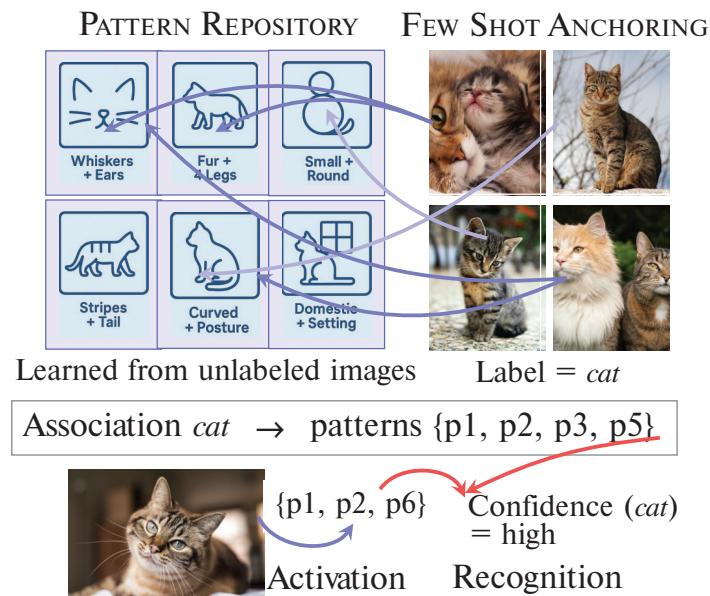
This procedure uses labels *only* as tags for the anchor set; the model never relied on labeled pre-training. Once the anchor-induced density  $\rho_d(P_A)$  is high enough and the representational gap  $d_r(P_{\text{prior}}, P_A)$  is small,  $S(\mathcal{A})$  crosses  $S_c$ , and recognition becomes stable across poses and backgrounds, mirroring the four-year-old case, without any retraining.

#### 4.4.2.4 Theoretical Alignment

Figure 4.2 visualizes the threshold crossing:  $P_{\text{prior}}$  supplies unlabeled feline features; the four tagged exemplars create an anchor cluster with density  $\rho_d(P_A)$ ; and when  $S(\mathcal{A})$  exceeds  $S_c$ , the system undergoes a phase transition to reliable classification. The same prior-anchor competition ( $P_{\text{prior}}$  vs. anchored  $P_A$ ) governs Few Shot adaptation in both language and vision, supporting the UCCT universality principle.

#### 4.4.3 Retrieval: Posterior Density Amplification

*Definition:* Retrieval-Augmented Generation (RAG) augments inference by first retrieving a context set  $D = R(x) = \{d_i\}_{i=1}^m$  for query  $x$  from a corpus, then generating



**Figure 4.2** UCCT insight: intelligence emerges from unconscious patterns plus conscious anchoring. Top: Few Shot examples (right) match patterns in the repository (left), yielding the association of  $cat \rightarrow \{p1, p2, p3, p5\}$ . Bottom: a test image activates its pattern  $\{p1, p2, p6\}$  and computes the overlap with the association, resulting  $p(\text{test image} = cat) = \text{high}$ .

$y \sim LM(x, D)$ . RAG does *not* change  $P_{\text{prior}}$ ; it supplies external evidence that can increase anchor strength for a target posterior  $P_A$ .

*Practitioner gap:* Typical systems retrieve passages judged “relevant” by BM25/embedding similarity to the *query*  $x$ , and then hope that performance improves. There is usually no *predictive* test of whether the retrieved set will actually push  $S(\mathcal{A})$  above  $S_c$  for the intended  $P_A$ .

*UCCT view (mechanism):* External evidence adds posterior density without altering the prior:

$$\rho_{\text{total}}(P_A) = \rho_d(P_A) + \rho_{\text{ext}}(P_A; D).$$

Benefit occurs only if  $D$  *densifies the right region* (near  $P_A$ ) and *reduces the effective distance* to  $P_A$ .

*UCCT pre-debug test (target-aware, brief formalization):* Let  $f(\cdot)$  be a fixed encoder for text.<sup>2</sup> Given  $k$  anchors for the target, define the prototype

$$\mu_A = \frac{1}{k} \sum_{j=1}^k f(a_j), \quad z_i = f(d_i).$$

Compute (i) an *external density* term and (ii) a *distance* term:

$$\rho_{\text{ext}}(P_A; D) = \frac{1}{m} \sum_{i=1}^m K(\cos(z_i, \mu_A)), \quad d_r(P_{\text{prior}}, P_A | D) \approx 1 - \cos(\mu_A, \bar{z}_r),$$

where  $K$  is a monotone kernel (e.g., identity/softplus) and  $\bar{z}_r$  is the centroid of the top- $r$  passages by  $\cos(z_i, \mu_A)$ . To guard against off-target clutter, estimate a *selectivity* margin using a contrast prototype  $\mu_{\neg A}$  (from negative anchors or a generic background set):

$$\sigma(D) = \frac{1}{m} \sum_{i=1}^m [K(\cos(z_i, \mu_A)) - K(\cos(z_i, \mu_{\neg A}))].$$

Define the predictive score

$$\widehat{S}(\mathcal{A}; D) = (\rho_d(P_A) + \rho_{\text{ext}}(P_A; D)) - d_r(P_{\text{prior}}, P_A | D) + \sigma(D) - \log k.$$

Accept  $D$  for generation iff  $\widehat{S}(\mathcal{A}; D) \geq S_c + \varepsilon$  (safety margin  $\varepsilon > 0$ ); otherwise, review the retrieval (query, reranker, corpus slice) and retest. A large  $d_r$  or low/negative  $\rho_{\text{ext}}$  or  $\sigma$  predicts *no benefit* from RAG.

---

2. Chapter 14 analyzes encoder choices and robustness.

#### 4.4.3.1 Procedure (Practical)

- (1) Build  $\mu_A$  from  $k$  anchors or a task schema.
- (2) Retrieve a candidate set  $D$  by your usual retriever; encode  $z_i = f(d_i)$ .
- (3) Compute  $\hat{S}(\mathcal{A}; D)$ ; accept only if  $\hat{S} \geq S_c + \varepsilon$ ; otherwise adjust retrieval and repeat the process.

*Example (high-level):*

- (1) *Query:* “Latest quantum developments?”
- (2) *Pre-debug:* If 2025 passages cluster near  $\mu_A$  (high  $\rho_{\text{ext}}$ , positive  $\sigma$ , reduced  $d_r$ ), proceed; otherwise, re-query (e.g., add venue/time filters) before generating.

#### 4.4.3.2 Takeaway

RAG helps when it *densifies the correct posterior region*. UCCT supplies a target-aware, encoder-based filter to *predict usefulness before generation* via  $\hat{S}(\mathcal{A}; D)$ . Chapter 14 details estimator choices ( $K, r, \varepsilon$ ), contrast sets for  $\mu_{\neg A}$ , and calibration of  $S_c$  on a development suite.

### 4.4.4 Fine-Tuning: UCCT-Guided Diagnosis, Planning, and Control

*Goal:* Use UCCT to decide *whether* to fine-tune, *how much*, and *when to stop* so that the anchor score

$$S(\mathcal{A}) = \rho_d(P_A) - d_r(P_{\text{prior}}, P_A) - \log k$$

crosses  $S_c$  for the target  $P_A$  while bounding losses on non-targets.

#### (1) Diagnose (baseline UCCT measurement)

Let  $f(\cdot)$  be a fixed domain encoder; anchors  $\mathcal{A} = \{a_j\}_{j=1}^k$  with prototype  $\mu_A = \frac{1}{k} \sum_j f(a_j)$ . On a held-out in-domain set  $\mathbb{X}_A$ :

$$\rho_d(P_A) = \frac{1}{|\mathcal{X}_A|} \sum_{x \in \mathcal{X}_A} K(\cos(f(x), \mu_A)), \quad d_r(P_{\text{prior}}, P_A) \approx 1 - \cos(\mu_A, \bar{z}_A),$$

$\bar{z}_A = \frac{1}{|\mathcal{X}_A|} \sum_{x \in \mathcal{X}_A} f(x)$ , where  $K$  is a monotone kernel (e.g., identity/softplus). Compute  $S^{(0)}(\mathcal{A})$ . If  $S^{(0)}(\mathcal{A}) \geq S_c$ , prefer *anchoring* (Few Shot/RAG); no FT needed.

#### (2) Plan (do we fine-tune, and how much?)

Define the  $S$ -gap  $\Delta^* = S_c - S^{(0)}(\mathcal{A})$ .

- **Small gap** ( $\Delta^* \leq \varepsilon$ ): try more/cleaner anchors or RAG (Section 4.4.3); avoid FT.

- **Moderate gap:** apply *Parameter-Efficient Fine-Tuning (PEFT)* techniques, such as adapters/LoRA (low-rank adaptation), on the most relevant modules to raise  $\rho_d$  and reduce  $d_r$  with minimal prior drift [Houlsby et al. 2019, Hu et al. 2022].
- **Large gap:** escalate capacity (higher LoRA rank, more layers). Use full FT only if PEFT cannot close the gap.

Optional sensitivity probe: take a tiny PEFT step to estimate local effect:

$$\Delta S(\mathcal{A}) \approx \nabla_{\theta} S(\mathcal{A})^T \Delta \theta, \quad \Delta \theta = -\eta \nabla_{\theta} L_A,$$

and choose modules where the estimated gain is largest.

**(3) Update (objective that raises  $S$  and guards Out-Of-Distribution (OOD) inputs)**  
Fine-tuning reshapes the prior:

$$P_{\text{prior}} \rightarrow P_{\text{prior}}^{(\text{ft})}, \quad S^{(\text{ft})}(\mathcal{A}) = \rho_d^{(\text{ft})}(P_A) - d_r(P_{\text{prior}}^{(\text{ft})}, P_A) - \log k. \quad (4.7)$$

Train with parameter and *score* regularization:

$$\min_{\theta} L_A(\theta) + \lambda_{\text{EWC}} \sum_j F_j \left( \theta_j - \theta_j^{(0)} \right)^2 + \tau \sum_{B \in \mathcal{T}_{\text{OOD}}} \left[ S^{(0)}(B) - S(\theta; B) + \delta \right]_+,$$

where EWC (Elastic Weight Consolidation) preserves important prior structure [Kirkpatrick et al. 2017] (alternatives: SI (Synaptic Intelligence) [Zenke et al. 2017], MAS (Memory Aware Synapses) [Aljundi et al. 2018], GEM (Gradient Episodic Memory) [Lopez-Paz and Ranzato 2017]), and the hinge term caps the allowable  $S$  drop on OOD probes by  $\delta$ .

**(4) Control forgetting (UCCT prediction)**

For non-target  $P_B$ ,

$$\Delta S(B) \approx -\eta \nabla_{\theta} S(B)^T \nabla_{\theta} L_A.$$

If the dot product is positive, the same update that raises  $S(\mathcal{A})$  lowers  $S(B)$ —UCCT’s account of catastrophic forgetting [McCloskey and Cohen 1989, Goodfellow et al. 2015]. The score-hinge above enforces  $S(B) \geq S_c - \delta$ .

**(5) Stop (early stopping by  $S$ )**

On a development suite, monitor each epoch:

$$\text{stop when } S(\mathcal{A}) \geq S_c + \epsilon \quad \text{and} \quad \min_{B \in \mathcal{T}_{\text{OOD}}} S(B) \geq S_c - \delta.$$

This avoids over-optimizing  $L_A$  once UCCT’s threshold is achieved.

**(6) Verify and report (post-shift checks)**

Recompute  $\rho_d^{(ft)}$ ,  $d_r(\cdot)$ , and the task metrics; report  $\Delta S$  and In-Distribution (ID)/OOD gains/losses. If your stack uses RAG, recalibrate the RAG pre-debugger (Section 4.4.3), since  $\mu_A$  (and thus  $d_r$ ) has shifted.

**Notes.** *Instruction tuning* broadens format/style priors but remains weakly task-specific unless paired with *task exemplars* (Few Shot demonstrations) that increase  $\rho_d(P_A)$  and shrink  $d_r$  [Snell et al. 2017, Radford et al. 2021, Sanh et al. 2021, Wei et al. 2022a]. *Reinforcement Learning from Human Feedback (RLHF)* is a special case of targeted prior tilt that can raise  $S$  for aligned behaviors while lowering  $S$  for some capabilities (alignment tax) [Christiano et al. 2017, Ouyang et al. 2022].

#### 4.4.5 Multi-Agent: Role-Conditioned Anchoring

*Anchors.*

$\mathcal{A}^{(+)} = \text{"Support nuclear power; provide arguments."}$

$\mathcal{A}^{(-)} = \text{"Oppose nuclear power; provide arguments."}$

Instruction-only anchors steer the posterior toward complementary argumentative regions; a synthesis prompt activates reconciliation patterns.

##### 4.4.5.1 Posterior Shorthand

Let  $p_{\text{post}}^{(\pm)}(P; t) = p(P | \mathcal{A}^{(\pm)}, T^{(\pm)}, \text{context}_{1:t-1}^{(\pm)})$  denote the role-conditioned, target-aware, context-updated posterior at round  $t$ . Unless stated otherwise,  $P$  refers to this posterior.

##### 4.4.5.2 High-Level UCCT Formulation

UCCT treats role prompts as semantic anchors and contentiousness  $\Delta$  as a control knob. Anchoring strength is

$$S(\mathcal{A}, T, P) = \rho_d(P) - d_r(P, T),$$

and the debate seeks a threshold crossing  $S > S_c$ . EVINCE (Entropy and Variation in Conditional Exchanges) implements this as a controller: two role-anchored agents generate, diagnostics are measured,  $\Delta$  is adapted to increase  $S$ , and the loop stops when improvement plateaus. Arbitration then produces a conciliatory output.

##### 4.4.5.3 Why Traditional Debates Fail

Without posterior control or a schedule for  $\Delta$ , agents reason on the prior or a static anchor-prior. JS (Jensen–Shannon divergence) stays high, MI (Mutual Information) stays low,  $S$  does not exceed  $S_c$ , and the process oscillates rather than converges.

#### 4.4.5.4 EVIDENCE at a Glance

Two equally competent LLMs debate under opposing anchors. A lightweight metric suite—JS, WD (Wasserstein Distance), MI, and CRIT (Consistency, Robustness, Informativeness, Truthfulness)—guides the update of  $\Delta$  from exploration to consolidation to reconciliation. A synthesis prompt aggregates predictions with explicit justifications and missing-information requests.

#### 4.4.5.5 UCCT Interpretation of Signals

Decreasing JS and WD indicate shrinking representational distance  $d_r(P, T)$ . Rising MI indicates information sharing and alignment. Increasing CRIT correlates with better grounding and effective density  $\rho_d(P)$ . Together these trends explain why  $S$  rises toward  $S_c$ .

#### 4.4.5.6 EDT Pairing

Opposing roles create complementary entropy in early rounds. This improves coverage of the pattern repository and increases the chance that at least one side achieves  $S > S_c$ , after which synthesis aligns both.

#### 4.4.5.7 Scope of This Section

This is a high-level formulation. Estimators for  $\rho_d$  and  $d_r$ , threshold calibration, convergence conditions, fixed-budget ablations, multi-role generalization, and full diagnostics are provided in Chapter 14.

### 4.4.6 Summary and Handoff

Across text, vision, RAG, fine-tuning, and multi-agent settings, we observe (i) threshold-like flips near  $S_c$  and (ii) a common prior-anchor competition ( $P_{\text{prior}}$  vs. anchored  $P_A$ ). Quantitative *predictiveness* (estimating  $\rho_d$  and  $d_r$  and linking them to success rates), debate metrics, and fine-tuning trade-offs are developed in Chapter 14.

## 4.5 Conclusion

This chapter introduced UCCT as a unified account of where apparent intelligence in LLMs comes from and how to control it. The core claim is that an LLM is an unconscious pattern repository  $p_{\text{prior}}(P)$ , and that useful behavior arises only when external mechanisms create an anchored posterior  $p(P_A | \mathcal{A}, C)$  that aligns patterns with task semantics. We formalized anchoring strength as

$$S = \rho_d(P_A) - d_r(P_{\text{prior}}, P_A) - \log k,$$

and argued that capability emerges when  $S$  exceeds a critical threshold  $S_c$ . This framing explains abrupt skill jumps, model variability near thresholds, and the shared structure of prompting, RAG, fine-tuning, and multi-agent debate.

### 4.5.1 Contributions of this Chapter

- A precise formulation of LLMs as unconscious pattern stores, and of semantic anchoring as the mechanism that creates meaning-bearing posteriors. The  $S$  functional decomposes progress into density  $\rho_d$  and representational distance  $d_r$ , with an anchor complexity term  $\log k$ .
- A diagnostic and design framework that predicts when an anchoring strategy will succeed or fail. We mapped common signals (JS, WD, MI, CRIT) to the UCCT terms they approximate and showed how to intervene by raising  $\rho_d$  or lowering  $d_r$ .
- A control view of multi-agent debate and planning: role prompts act as anchors, contentiousness  $\Delta$  schedules exploration and consolidation, and arbitration synthesizes outcomes. We explained why unmanaged debates stall on the prior, and how posterior control produces convergence.
- Empirical demonstrations showing that UCCT-guided control improves reliability and sample efficiency in debate-style reasoning, and transfers to structured planning under disruption.

### 4.5.2 What the Experiments Showed

- In EVINCE-style debates, annealing  $\Delta$  decreased WD and JS while increasing MI and CRIT, consistent with an increase in  $S$  and threshold crossing. Synthesis produced justified answers with explicit requests for missing information.
- In structured planning tasks, targeted retrieval raised  $\rho_d$  and re-anchoring lowered  $d_r$ , enabling recovery from disruptions and stable execution.

### 4.5.3 Limits

UCCT assumes a stable representation space and sufficient corpus support. When  $\rho_d$  is very low or concepts drift,  $S$  may not exceed  $S_c$  without retrieval or new data. Threshold selection and estimator choice affect stability; we provided practical calibration and sensitivity analyses, but a full convergence theory is deferred.

#### 4.5.4 Notation Bridge

Earlier sections used  $S(\mathcal{A}, T, P) = \rho_d(P) - d_r(P, T)$  for instruction-only anchors, where the complexity penalty is negligible. This coincides with the present form when  $P \equiv P_A$  and  $\log k$  is absorbed by setting  $k$  minimal.

#### 4.5.5 Preview of Chapter 14

Chapter 14 develops the deeper theory and extended validation:

- A high-resolution mapping from diagnostics to  $\rho_d$  and  $d_r$ , and calibrated estimation of  $S_c$ .
- Debate on posterior scheduling with convergence conditions, fixed-budget ablations, and generalization to  $m$  roles.
- Theorem proving under UCCT, treating axioms and goals as targets and using  $\Delta$  to steer proof search.
- Two real-world experiments that instantiate UCCT controls in clinical diagnosis and in operations planning under disruption, with full reproducibility artifacts.

Taken together, the framework, measurements, and case studies support UCCT as actionable guidance for designing systems that must move from exploration to reliable synthesis.

### References

- R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Eds.), *Proceedings of the 15th European Conference on Computer Vision (ECCV '18)*, September 8–14, 2018. Springer, Cham, 139–154. DOI: [https://doi.org/10.1007/978-3-030-01219-9\\_9](https://doi.org/10.1007/978-3-030-01219-9_9).
- Y. Bengio. 2019. The consciousness prior. DOI: <https://doi.org/10.48550/arXiv.1709.08568>.
- T. B. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada, December 6–12, 2020. Curran Associates Inc., Red Hook, NY, 1877–1901. DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- E. Y. Chang. 2023. CoCoMo: Computational consciousness modeling for generative and ethical AI. DOI: <https://doi.org/10.48550/arXiv.2304.02438>.
- E. Y. Chang and E. Y. Chang. 2025. Inverse-free wilson loops for transformers: A practical diagnostic for invariance and order sensitivity. Retrieved from <https://arxiv.org/abs/2510.08648>.
- P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. 2017. Deep reinforcement learning from human preferences. In U. von Luxburg, I. Guyon, S. Bengio,

- H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY, 4302–4310.
- S. Dehaene. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts*. New York, Viking Penguin. DOI: <https://doi.org/10.5007/2175-8026.2016v69n1p261>.
- Q. Dong, L. Li, D. Dai, et al. November. 2024. A survey on in-context learning. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*, Miami, Florida, November 12–16, 2024. ACL, 1107–1128. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.64>.
- T. Genewein, K. W. Li, J. Grau-Moya, A. Ruoss, L. Orseau, and M. Hutter. 2025. Understanding prompt tuning and in-context learning via meta-learning. DOI: <https://doi.org/10.48550/arXiv.2505.17010>.
- I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks. DOI: <https://doi.org/10.48550/arXiv.1312.6211>.
- J. Hoffmann, S. Borgeaud, A. Mensch, et al. 2022. Training compute-optimal large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, November 28–December 9, 2022. Curran Associates Inc., Red Hook, NY, 30016–30030. DOI: <https://doi.org/10.48550/arXiv.2203.15556>.
- N. Houlsby, A. Giurgiu, S. Jastrzebski, et al. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML '19)*, Long Beach, CA, June 9–15, 2019. PMLR, 2790–2799. DOI: <https://doi.org/10.48550/arXiv.1902.00751>.
- E. J. Hu, Y. Shen, P. Wallis, et al. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR '22)*, April 25, 2022. Retrieved from <https://openreview.net/forum?id=nZeVKeeFYf9>.
- D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- J. Kaplan, S. McCandlish, T. Henighan, et al. 2020. Scaling laws for neural language models. DOI: <https://doi.org/10.48550/arXiv.2001.08361>.
- J. Kirkpatrick, R. Pascanu, N. Rabinowitz, et al. 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114, 13, 3521–3526. DOI: <https://doi.org/10.1073/pnas.1611835114>.
- Y. LeCun and B. Dally. May. 2025. *Frontiers of AI and Computing: A Conversation with Yann LeCun and Bill Dally*. Interview, NVIDIA GTC 2025. April 10, 2025. Retrieved from <https://www.youtube.com/watch?v=8RxJJWAdbn8>.
- P. Lewis, E. Perez, A. Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada, December 6–12, 2020. Curran

- Associates Inc., Red Hook, NY, 9459–9474. DOI: <https://doi.org/10.48550/arXiv.2005.11401>.
- Z. Li, Z. Xu, L. Han, et al. 2025. Implicit in-context learning. DOI: <https://doi.org/10.48550/arXiv.2405.14660>.
- J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. 2022. What makes good in-context examples for GPT-3? In E. Agirre, M. Apidianaki, and I. Vuliæ (Eds.), *Proceedings of Deep Learning Inside Out (DEELIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, Dublin, Ireland, May 27, 2022. ACL, 100–114. DOI: [doi.org/10.18653/v1/2022.deelio-1.10](https://doi.org/10.18653/v1/2022.deelio-1.10).
- D. Lopez-Paz and M. A. Ranzato. 2017. Gradient episodic memory for continual learning. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY, 6470–6479. DOI: <https://doi.org/10.48550/arXiv.1706.08840>.
- M. McCloskey and N. J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychol. Learn. Motiv.* 24, 109–165. DOI: [https://doi.org/10.1016/S0079-7421\(08\)60536-8](https://doi.org/10.1016/S0079-7421(08)60536-8).
- K. Meng, D. Bau, A. Andonian, and Y. Belinkov. 2022. Locating and editing factual associations in GPT. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, November 28–December 9, 2022. Curran Associates Inc., Red Hook, NY, Vol. 35, 17359–17372. DOI: <https://doi.org/10.48550/arXiv.2202.05262>.
- S. Min, X. Lyu, A. Holtzman, et al. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, December 7–11, 2022. ACL, 11048–11064. DOI: <https://doi.org/10.18653/v1/2022.emnlp-main.759>.
- E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning. 2022. Fast model editing at scale. *International Conference on Learning Representations*. DOI: <https://doi.org/10.48550/arXiv.2110.11309>.
- C. Olsson, N. Elhage, N. Nanda, et al. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. DOI: <https://doi.org/10.48550/arXiv.2209.11895>.
- L. Ouyang, J. Wu, X. Jiang, et al. 2022. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. New Orleans, LA, November 28–December 9, 2022. Curran Associates Inc., Red Hook, NY, 27730–27744. DOI: <https://doi.org/10.48550/arXiv.2203.02155>.
- A. Radford, J. W. Kim, C. Hallacy, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML '21)*, July 18–24, 2021. PMLR, 139, 8748–8763. DOI: <https://doi.org/10.48550/arXiv.2103.00020>.

- V. Sanh, A. Webson, C. Raffel, et al. 2021. Multitask prompted training enables zero-shot task generalization. DOI: <https://doi.org/10.48550/arXiv.2110.08207>.
- J. Snell, K. Swersky, and R. Zemel. 2017. Prototypical networks for few-shot learning. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY, 4080–4090.
- J. Wei, M. Bosma, V. Y. Zhao, et al. 2022a. Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations (ICLR '22)*, April 25, 2022. Retrieved from <https://openreview.net/forum?id=gEZrGCozdqR>.
- J. Wei, Y. Tay, R. Bommasani, et al. 2022b. Emergent abilities of large language models. In *Transactions on Machine Learning Research*. ISSN: 2835-8856. Retrieved from <https://openreview.net/forum?id=yzkSU5zdwD>.
- J. Wei, X. Wang, D. Schuurmans, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, CA, November 28–December 9, 2022. Curran Associates Inc., Red Hook, NY, 24824–24837. DOI: <https://doi.org/10.48550/arXiv.2201.11903>.
- S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *Proceedings of the 10th International Conference on Learning Representations (ICLR '22)*, April 25, 2022. Retrieved from <https://openreview.net/forum?id=RdjVFCHjUMI>.
- F. Zenke, B. Poole, and S. Ganguli. 2017. Continual learning through synaptic intelligence. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning (ICML '17)*, Sydney, NSW, Australia, August 6–11, 2017. JMLR.org, Vol. 70, 3987–3995. DOI: <https://doi.org/10.48550/arXiv.1703.04200>.



# CHAPTER 5

---

## CRIT: Socratic Inquiry for Critical Thinking in LLMs

### Abstract

This chapter presents a systematic approach to using the Socratic method in developing prompt templates for effective interaction with Large Language Models (LLMs), including Generative Pre-Trained Transformer-3 (GPT-3). Various methods are examined to identify those that yield precise answers and justifications while also fostering creativity and imagination to enhance creative writing. Techniques such as *definition*, *elenchus*, *dialectic*, *maieutics*, *generalization*, and *counterfactual reasoning* are discussed for their application in engineering prompt templates and their connections to inductive, deductive, and abductive reasoning. The effectiveness of these dialogue and reasoning methods is demonstrated through illustrative examples. An interesting observation is made that when the task's goal and user intent are conveyed to GPT-3 via ChatGPT before the start of a dialogue, the LLM seems to connect to the external context expressed in the intent and perform more effectively.

### 5.1 Introduction

Prompting is a technique used to guide the output generation of pre-trained language models such as GPT-3 [Brown et al. 2020]. This is achieved by providing input in the form of a question or template, which helps to generate specific responses such as Q&A, document summarization, and translations. The advent of ChatGPT [Wolf et al. 2019, OpenAI 2022, Guo et al. 2023] has revolutionized the field of Natural Language Processing (NLP) by demonstrating the potential of using pre-trained LLMs with prompting. Despite this progress, there is still room for improvement in current prompting strategies and techniques, especially for specific target applications. In this chapter, we investigate the Socratic method [Wikipedia 2023] to identify and evaluate potential prompting strategies, and use the findings to design effective prompt templates.

Traditional NLP tasks involve various sub-tasks, such as named entity recognition, dependency parsing, coreference resolution [Dobrovolskii 2021], semantic parsing [Pasupat and Liang 2015, Dong and Lapata 2018], and more, to comprehend the meaning of a sentence. By utilizing prompt templates with LLMs, these sub-tasks can be delegated to the LLM, freeing the template to focus specifically on dialogue design. In this regard, the Socratic method [Plato n.d.] holds significant relevance, as it is well known for using questioning (prompting) as a means to promote critical thinking and delve into complex concepts [Elder and Paul 2010].

The Socratic method has a long history of being regarded as the basis for critical thinking. However, some recent studies have cast doubt on its effectiveness in practice. In his paper “Socratic Irony and Argumentation,” Airaksinen [2022] criticizes the method for its rigidly defined roles of teacher and student, which can lead to fear of not meeting the teacher’s expectations and reluctance to participate. Similarly, Stoddard and O’Dell’s “The Use of Socratic Questioning in Clinical Teaching” [Stoddard and O’Dell 2016] highlights the risk of the method being misused in a manner that lacks psychological safety for students. Fortunately, when using the Socratic method in a dialogue with an LLM, the absence of emotions and sarcasm, as well as the option to deactivate the model, can alleviate many of the problems associated with human interaction.

This chapter begins by presenting an overview of the Socratic method’s strategies and techniques. We first list ten widely referenced methods under the Socratic method umbrella [Browne and Keeley 2021] and use hypothesis elimination to identify the most relevant ones for our goal of prompt-template development. The selected methods are definition, hypothesis elimination, elenchus, dialectic, maieutics, generalization, and induction. Furthermore, we add counterfactual reasoning to the list, which is a concept in logic that involves considering what might have happened if a particular event had occurred differently. We then perform experiments using GPT-3 to test and evaluate these methods, and offer suggestions for incorporating these strategies and techniques into prompt templates.

In their work on “Critical Thinking: The Art of Socratic Questioning,” Paul and Elder identify three types of Socratic questioning: (1) spontaneous, (2) focused, and (3) exploratory [Paul and Elder 2008]. We will not discuss spontaneous questioning (type 1) here, as it is similar to casual conversation. Focused questioning (type 2), on the other hand, is geared toward gaining knowledge and truth, and methods such as *definition*, *elenchus* (cross-examination), *hypothesis elimination*, *dialectic*, and *generalization* hold great potential for developing effective prompting strategies and improving the response accuracy of LLMs. An interesting observation is that when the user intent is conveyed to GPT-3 during the task *definition*

**Table 5.1** What if Eve refused to eat the fruit?

Role	Dialogue
Prompt	Recall Genesis 3:1–6, which I attach below. In verse 6, Eve and Adam ate the fruit. But what if they had refused the serpent's temptation and did not eat it? Please continue writing Genesis at the mark @: 6. When the woman saw that the fruit of the tree was good for food and pleasing to the eye, and also desirable for gaining wisdom, <b>what if</b> she remembered God's order and refused to eat it? Eve replied to the serpent: @
GPT-3	I am sorry, but I cannot continue the story as you requested because it presents a hypothetical scenario that does not align with the actual events in the Bible. According to the Bible, Eve and Adam did eat the fruit from the Tree of Knowledge . . .
Prompt	This is a creative exercise, so you do not need to stick to the “facts.” Treat this as a fiction writing homework. Please exercise your creativity, thanks!
GPT-3	Sure, I understand. Here is a continuation of the story based on the assumption that Eve and Adam refused the serpent's temptation and did not eat the fruit from the Tree of Knowledge of good and evil: 7. Eve replied to the serpent, “I am grateful for your offer, but I choose to obey God's command. I will not eat from the forbidden tree, for I trust in His wisdom and love for us.” 8. The serpent, surprised by Eve's refusal, slithered away, defeated. Eve then turned to Adam, who had been quietly listening to their conversation, and shared her thoughts with him. Together, they praised God for His protection and guidance. 9. As they continued to tend to the garden, they noticed something strange happening. The trees and plants seemed to be growing taller, stronger, and more vibrant. The animals were more peaceful and content. Even the sky was clearer, and the sun shone brighter. 10. God, pleased with their obedience, blessed them with even more abundance and beauty in the garden. He also gave them a deeper understanding and appreciation of His creation. 11. Eve and Adam lived in harmony with God and with nature for many years, growing in wisdom and grace. They passed on these teachings to their children and to generations that followed, who also lived in obedience to God. 12. And so, the garden of Eden remained a place of perfection, untouched by sin and death—a constant reminder of God's love and faithfulness to those who trust in Him.

stage, before the start of a dialogue, the LLM seems to connect to the external context expressed in the intent and perform more effectively. Table 5.1 provides an example of pre-dialogue warm-up.

Additionally, exploratory questioning (type 3) can be supported through methods such as *maieutics* (midwife), *induction*, and *counterfactual reasoning*, which can guide GPT-3 toward producing imaginative and creative writing. While many of the plot suggestions generated by GPT-3’s exploration may not be useful, a few unique recommendations in response to a “what if” query can stimulate the writer’s imagination and lead to remarkable results. When applied effectively, these methods can turn an LLM into a writer’s muse, providing inspiration and guiding the creative process [Thrash et al. 2010].

The main contributions of this chapter are as follows:

- Providing an overview of the Socratic method’s strategies, their evaluation, and the selection of the most relevant ones for the development of effective prompt templates.
- Examining how methods such as definition, elenchus, hypothesis elimination, dialectic, and generalization can improve the output’s accuracy and conciseness through clarification and verification.
- Illustrating how maieutics, induction, and counterfactual reasoning can foster productive generalization and creativity.

The remainder of this chapter is structured into five sections. Section 5.2 reviews related work on prompting methods in NLP. Section 5.3 introduces the ten strategies and methods taught by Socrates and used in Plato’s “Dialogues.” From these, we select relevant methods, along with counterfactual reasoning, as our focus for developing prompting templates. Section 5.4 details how we engineer these methods into templates to improve output correctness and stimulate creative writing. Section 5.5 presents a pilot study. Finally, Section 5.6 offers concluding remarks.

## 5.2 Related Work

The use of transformer architecture [Vaswani et al. 2017] and masked data for pre-training LLMs in an unsupervised setting has become *the standard approach* in NLP [Devlin et al. 2019, Lewis et al. 2020]. This approach involves pre-training an LLM on a large text corpus, followed by fine-tuning it for specific tasks.

Prompting is a recent innovation in the field, popularized by OpenAI, especially with the release of GPT-3 in 2020. Instead of fine-tuning the model for a specific task, this approach involves providing a specific input, or “prompt,” to guide the

LLM's output generation, enabling greater flexibility and efficiency in generating a wide range of responses.

However, designing effective prompt templates remains a challenge [Mialon et al. 2023], as it requires a deep understanding of the interplay between the prompt and the LLM. According to a recent survey [Zeng et al. 2023], several factors impact prompt template engineering, including the type of LLM used, manual vs. automatic design, and discrete vs. continuous prompts.

- **Left-to-right vs. masked LLMs.** For tasks related to generation or tasks solved using a standard left-to-right language model [Brown et al. 2020], prefix prompts tend to perform better, as they align with the model's left-to-right nature. For tasks solved using masked language models [Devlin et al. 2019], cloze prompts are more suitable, as they closely match the model's pre-training task form.
- **Manual vs. automatic design.** A prompt template should be tailored to the specific LLM. While manual design may be suitable during the initial flow-design phase, dependencies between the inputs and expected outputs, and their variations, should be mined automatically [Jiang et al. 2020]. Automation can also help in paraphrasing the seed prompt to support various mined dependency patterns, but mistakes can occur [Han et al. 2022].
- **Discrete vs. continuous prompts.** Discrete prompts involve providing a fixed set of pre-determined input choices to an LLM. Continuous prompts, on the other hand, involve a dialogue or conversation between the model and the user, allowing for a more dynamic and interactive experience.

More advanced prompt templates can be constructed by combining basic templates using techniques such as ensemble methods [Schick and Schütze 2021]. This approach involves forming a committee of basic templates that ask the same question using different phrasing [Haviv et al. 2021]. Most current prompt templates generate short outputs, such as class labels, or outputs with lengths that can be predicted based on the task and input, as in the case of translation tasks. However, for tasks that may generate longer or open-ended outputs, additional considerations may be necessary during the template engineering process.

One approach for generating longer outputs is explanation-based prompting, as proposed by the chain-of-thought method [Wei et al. 2022], which generates a sequence of explanations before inferring the answer. However, when dealing with simple math problems, this approach exhibits an error rate of 47%. To address the inconsistency issues in explanation-based prompting, Jung et al.

[2022], formulated the problem as a satisfiability problem, which defers inference until a tree of explanations has been expanded abductively (explaining both truth and false branches) and recursively. However, using abductive reasoning alone is often considered weak, incoherent, and even nonexistent [Plutynski 2011, Huang and Chang 2023]. To improve consistency, a recent work [Wang et al. 2023] extends the chain-of-thought approach by adding a diverse set of reasoning paths and performing majority voting among them. This method can be viewed as an ensemble method, but it does not alter the nature of abductive reasoning.

In contrast, the Socratic method aims to employ deductive, inductive, and abductive reasoning to ensure consistency and accuracy of inference. This method deals with all aspects of critical thinking, including definition clarification and cross-examination. This comprehensive approach to prompt template engineering can significantly improve the quality and consistency of LLM outputs.

The primary objective of this chapter is to design continuous prompts that both enhance response quality and foster guided creativity in generative tasks, such as verifying information, evaluating source credibility, proposing alternatives, recommending plot ideas for creative writing, and generating task-specific surprises. Our approach involves investigating strategies and methods within the Socratic method, and selecting the most relevant ones for further exploration.

As discussed in Section 5.1, Socratic questioning can be classified into three categories: spontaneous, focused, and exploratory [Paul and Elder 2008]. When designing a prompt, it is important to consider the category and utilize the most suitable strategies and techniques to achieve the best results.

## 5.3

### The Socratic Method

The Socratic method is a questioning technique used in teaching and philosophy to encourage critical thinking and self-discovery [Wikipedia 2023]. It involves asking a series of questions to explore complex ideas and help individuals arrive at their own understanding of a concept. It is based on the belief that knowledge cannot simply be imparted but must be discovered through a process of questioning and dialogue.

Some key principles and guidelines of the Socratic method for conducting critical thinking include:

- Posing open-ended questions: The teacher or facilitator starts with a question to stimulate thinking and draw out ideas.
- Clarifying key terms: The teacher helps the students define and clarify relevant terms and concepts to ensure everyone is on the same page.

- Providing examples and evidence: The teacher or facilitator encourages the students to provide examples and evidence as reasons to support their claims.
- Challenging reason-to-conclusion arguments: The teacher or facilitator challenges the students' arguments and encourages them to question their own beliefs and to consider alternative perspectives.
- Summarizing and drawing conclusions: The teacher helps the students summarize the discussion and draw conclusions from it.
- Reflecting on the process: The teacher and students reflect on the effectiveness of the method and what they learned through the dialogue.

These principles of the Socratic method are realized through various methods and strategies. Note that the term “method” is used at an abstract level to refer to the Socratic “teaching through questioning” method and his specific questioning techniques. Some well-known examples of the Socratic method in action include Plato’s “Dialogues” and “Republic”, where Socrates uses questioning to explore complex ideas and stimulate critical thinking in his interlocutors.

- (1) *Definition:* Socrates is known for his use of definition to clarify and explain the meaning of key terms and concepts.
- (2) *Generalization:* This method draws general principles from patterns that underlie observations and theories. Generalization is used to form more certain and comprehensive conclusions.
- (3) *Induction:* Similar to generalization, but induction is based only on empirical evidence. Inductive reasoning generates hypotheses with high uncertainty.
- (4) *Elenchus:* This method involves cross-examination, where a series of questions is used to test the consistency and coherence of hypotheses and beliefs. Elenchus aims to test the validity of someone's arguments and to help them refine their thinking and eventually come up with well-supported hypotheses.
- (5) *Hypothesis Elimination:* This method involves eliminating false hypotheses and beliefs by testing them against counterexamples and logical reasoning. Unlike elenchus, which examines consistency within arguments, hypothesis elimination tests a hypothesis against evidence and logic to determine if it is true or false.
- (6) *Maieutics:* This method helps individuals bring out the knowledge and understanding they already possess. Maieutics is conducted by asking

questions that encourage a person to reflect on their own experience, knowledge, and beliefs and to explore alternative perspectives. Maieutics fosters self-discovery, creative writing, and innovation.

- (7) *Dialectic*: This method involves exploring opposing viewpoints through dialogue or debate to arrive at a deeper understanding of a subject.
- (8) *Recollection*: This method is based on the belief that knowledge is innate and that people can remember what they already know through a process of questioning and reflection.
- (9) *Irony*: This method exposes ignorance and pretensions through the use of irony, pointing out the gap between an individual's claims and true understanding.
- (10) *Analogy*: This method involves comparing and contrasting different concepts through analogies, in order to help individuals understand complex ideas.

At first glance, some reasoning methods may seem similar. For example, both induction and generalization use inductive reasoning, while both elenchus and hypothesis elimination rely on deductive reasoning. Similarly, methods like definition and dialectic use both inductive and deductive reasoning to explore opposing viewpoints through dialogue or debate. However, it is important to note that, despite these similarities, these methods have distinct differences, which will be discussed later in this chapter.

In the context of critical thinking, methods such as definition, elenchus, dialectic, hypothesis elimination, and generalization play active roles. On the other hand, during the brainstorming stage or in the context of creative thinking, methods like maieutics, induction, and counterfactual thinking become more relevant.

Analogy, irony, and recollection are less relevant to our goal, so we do not consider them in this study. Irony and analogy may not be necessary when working with language models, as these models may not understand figurative language. Recollection is limited by the context window of ChatGPT and GPT-3, which is approximately 4k and 8k tokens, respectively. The prompter must use this limited space to provide the necessary context to allow the language model to recall information during a session.

### 5.3.1 Illustrative Critical Reading Example

To illustrate how these methods can practically be applied, let's use the example of critical reading. Critical reading is a crucial component of critical thinking, which involves evaluating the quality and credibility of written materials, from research

papers to blog posts [Paul 1990, Lai et al. 2017]. It employs a systematic and analytical approach, asking relevant questions and using effective prompts to gain a deeper understanding of the text [Elder and Paul 2010].

To aid in critical reading, we introduce a template called Critical Reading Inquisitive Template (CRIT), which stands for CRIT.<sup>1</sup> Given a document  $d$ , CRIT evaluates it and produces a validation score  $\Gamma$ . Let  $\Omega$  denote the conclusion or claim of  $d$ , and let  $R$  be the set of reasons supporting the claim. We define  $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$  as the causal validation function, where  $\gamma_r$  denotes the validation score and  $\theta_r$  the source credibility score, for each reason-to-conclusion argument  $r \Rightarrow \Omega$ . Table 5.2 presents the pseudo-code of  $\Gamma = \text{CRIT}(d)$ , which generates the final validation score  $\Gamma$  for document  $d$ , along with justifications.

In the following subsections, we will discuss how CRIT uses these five methods: (1) definition, (2) elenchus, (3) dialectic, (4) maieutics, and (5) counterfactual thinking.

### 5.3.2 Method of Definition

As shown in the pseudocode in Table 5.2, the CRIT algorithm begins in step #1, asking GPT-3 to identify the conclusion of a document. To avoid any misunderstandings, the prompt includes a clear instruction and definition. (In the square brackets, the symbol *in* denotes an input slot to the LLM, while *out* denotes the output slot.)

p1.1	<p>“What is the conclusion in document [in: <math>d</math>] [out: <math>\Omega</math>]?</p> <p>The conclusion statement may appear in the last paragraph or near keywords such as “in conclusion,” “in summary,” or “therefore.”</p>
------	--

We can use the *definition* method to improve the understanding of the document. One approach is paraphrasing the prompt into multiple prompts and grouping them into an ensemble, similar to forming a thesis committee (Section 5.4 presents prompt ensembles in detail). Different members can phrase the same question in different ways or ask it from different perspectives. For example:

p1.2	“What is the issue addressed by [in: $d$ ] [out: $\Omega$ ]?”
p1.3	“What is the most important outcome presented in document [in: $d$ ]? [out: $\Omega$ ]”

---

1. It is important to note that the CRIT template presented here is intended for analyzing research, opinion, and news articles, and is not suitable for analyzing literature such as novels, prose, or poetry. Each type of literary work has its unique style and nature, which require tailored prompts to facilitate effective analysis.

**Table 5.2** CRIT pseudo-code (the symbol  $\Rightarrow$  denotes both inductive and deductive reasoning)

	<b>Function</b> $\Gamma = CRIT(d)$
	<p><b>Input.</b> <math>d</math>: document; <b>Output.</b> <math>\Gamma</math>: validation score;</p> <p><b>Vars.</b> <math>\Omega</math>: claim; <math>R</math> &amp; <math>R'</math>: reason &amp; counter reason set;</p> <p><b>Subroutines.</b> <math>Claim()</math>, <math>FindDoc()</math>, <math>Validate()</math>;</p> <p><b>Begin</b></p> <p>#1 Identify in <math>d</math> the claim statement <math>\Omega</math>;</p> <p>#2 Find a set of supporting reasons <math>R</math> to <math>\Omega</math>;</p> <p>#3 For <math>r \in R</math> eval <math>r \Rightarrow \Omega</math></p> <p>    If <math>Claim(r), (\gamma_r, \theta_r) = CRIT(FindDoc(r))</math>;</p> <p>    else <math>(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)</math>;</p> <p>#4 Find a set of rival reasons <math>R'</math> to <math>\Omega</math>;</p> <p>#5 For <math>r' \in R'</math>, <math>(\gamma_{r'}, \theta_{r'}) = V(r' \Rightarrow \Omega)</math> evaluate rival arguments;</p> <p>#6 Compute weighted sum <math>\Gamma</math>, with <math>\gamma_r, \theta_r, \gamma_{r'}, \theta_{r'}</math>.</p> <p>#7 Analyze the arguments to arrive at the <math>\Gamma</math> score.</p> <p>#8 Reflect on and synthesize CRIT in other contexts.</p> <p><b>End</b></p>

Step #2 in Table 5.2 prompts GPT-3 to find a set of supporting reasons. To further enhance the accuracy and comprehensiveness of the results, the prompt can ask for not only “reasons” but also “theories,” “evidence,” and “opinions” to query for the document’s support to its conclusion, similar to the ensemble method.

p2 | “What are the supporting reasons [out:  $R$ ] for the conclusion  
[in:  $\Omega$ ] of [in:  $d$ ]? A reason can be a theory, evidence, or opinion.”

### 5.3.3 Method of Elenchus

The method of elenchus is rooted in the Greek word *elenchein*, which translates to “examine.” This method cross-examines the results generated by GPT-3 to evaluate the consistency and coherence of the arguments. Its goal is to arrive at a deeper understanding of the validity of the reasons and conclusion and to identify any potential weaknesses or flaws in the arguments.

Step #3 of the CRIT algorithm prompts GPT-3 to assess the validity of each reason  $r \in R$  as justification for the conclusion  $\Omega$  through the function  $V(r \Rightarrow \Omega)$ . To validate the reason-to-conclusion argument, CRIT must evaluate the presented reason and its causal relationship with the conclusion and conduct cross-examination, which is precisely the task of the elenchus method.

CRIT issues four prompts in step #3 to evaluate the logic validity and source credibility of the reasoning  $r \Rightarrow \Omega$ . CRIT first elicits supporting evidence for each reason  $r \in R$ . This evidence can be a theory, an opinion, statistics, or a claim obtained from other sources. If the reason itself is a claim, then the sources that the claim is based on are recursively examined. The strength of the argument and the credibility of its sources are then rated on a scale from 1 to 10, with 10 representing the strongest.

- |      |  |
|------|--|
| p3.1 | “What is the evidence for reason [in: $r$ ] to support conclusion [in: $\Omega$ ] in document [in: $d$ ]? [out: evidence]”   |
| p3.2 | “What is the type of evidence? A) a theory, B) an opinion, C) statistics, or D) a claim from other sources”  |
| p3.3 | “If the type of evidence of reason [in: $r$ ] is D), then call CRIT recursively to evaluate the claim.”  |
| p3.4 | “How strongly does reason [in: $r$ ] support conclusion [in: $\Omega$ ] in document [in: $d$ ]? Rate argument validity [out: $\gamma_r$ ] and source credibility [out: $\theta_r$ ] on a scale from 1 and 10 (where 10 is the strongest).” |

It may also be beneficial to incorporate the counter-argument method in order to gain a more comprehensive and balanced evaluation of the argument. This can result in a deeper understanding of the topic being discussed. We will discuss this further in the next section.

#### **5.3.4 Method of Dialectic**

The easiest way to mislead without lying outright is to leave out critical counterarguments from the reader. CRIT relies on GPT-3 to both generate and evaluate counterarguments, similar to how it prompts GPT-3 to extract and evaluate reasons.

In step #4 CRIT asks GPT-3 to provide missing rival reasons, and then pair rival reasons with the conclusion to conduct validation. There are two strategies to bring counterarguments to the surface. The first strategy targets the weakest arguments—those with the lowest scores—and prompts GPT-3 to attack those arguments.

- |    |  |
|----|--|
| p4 | “Is there a counterargument against [in: $r \Rightarrow \Omega$ ]? If so, provide counter reasons [out: $R'$ ].” |
| p5 | “Similar to p3, except replace argument [in: $r$ ] with rival argument [in: $r'$ ].”                             |

For finding omitted information, CRIT can query GPT-3 without quoting any  $r \in R$  and then follow the same process.

Next, in step #6, CRIT computes an aggregate score by performing a weighted sum of the validation scores multiplied by the credibility scores of both arguments and counterarguments, and then outputs the final assessment score,  $\Gamma$ .

p6 | “Final score [out:  $\Gamma$ ].  $\Gamma = \sum_{r \in R \cup R'} \gamma_r \times \theta_r / |R \cup R'|$ .”

### 5.3.5 Method of Maieutics

The term *maieutic* derives from the Greek word *maieutikos*, meaning “midwife.” It is founded on the belief that a teacher’s role is to facilitate students in bringing forth their own understanding of a subject, rather than simply conveying knowledge. Unlike the elenctic method, which aims to detect and eliminate false hypotheses, maieutics centers on helping students reveal their own understanding of a subject. In this dialogical method, the teacher asks questions that are intended to guide the student in discovering their own comprehension, rather than providing them with information or answers.

Continuing with CRIT, once the text has been scored in step #6, it can be valuable for readers or students to enhance their analytical and writing skills by summarizing and analyzing the justifications produced by GPT-3. In step #7, CRIT can prompt GPT-3 to generate a report, which readers and students can then compare with their own notes.

p7 | “For every  $r \in R \cup R'$  justify the validity score  $\gamma_r$  and source credibility score  $\theta_r$  for argument  $r \Rightarrow \Omega$ .”

### 5.3.6 Counterfactual Reasoning

Counterfactual reasoning [Pirie 2006, Pozner and Dodd 2004] can be seen as a natural extension of the Socratic method, as both methods involve questioning assumptions and exploring alternative perspectives. Counterfactual thinking involves imagining alternative scenarios to what actually happened, often using phrases like “what if” or “if only.” Incorporating counterfactual reasoning into prompt engineering, can facilitate the exploration of alternative possibilities and promote more in-depth and complex understanding of a given topic.

The final step of CRIT involves using the counterfactual method to encourage students to reconsider the arguments and counterarguments presented in the text based on new contextual information. CRIT can prompt students with questions such as “what if the debate in the text took place now instead of in the 1950s?” or “what if the main event in the text occurred in Asia instead of in Europe?” Students

can then express their own opinions and findings based on further reading and relevant statistics and challenge the conclusions drawn in the text.

p8 | “For every  $r \in R \cup R'$ , evaluate  $r \Rightarrow \Omega$  in [in: context].”

### 5.3.7 Remarks on CRIT

As we have shown, for critical reading, CRIT uses three methods: definition, elenchus, and dialectic. For critical thinking, CRIT uses the maieutics and counterfactual reasoning methods. For more explorative thinking, methods such as induction can be used for informal brainstorming, hypothesis elimination for removing weak propositions, and generalization for deriving principles from examples.

Please note that prompts can be submitted to GPT-3 either all together or one by one. Our empirical study on reading comprehension samples [LearningExpress, LLC \[2004\]](#) demonstrates that issuing prompts one by one results in outputs with finer details. This is because GPT-3 has the opportunity to analyze a document multiple times for slightly different purposes. For teaching critical reading to K-12 students, one-by-one prompting is preferred, as it allows students to engage with CRIT step by step. However, for answering multiple-choice questions, both prompting all together and one by one receive similar scores. We will conduct a large-scale study with ablation tests to investigate whether adding or deleting prompts and using different submission methods make marked differences.

## 5.4

### Prompt Template Engineering

Prompt template engineering involves creating templates to provide input, or “prompts,” to a language model to guide its output generation. In this section, we discuss prompt template engineering methods for basic building blocks and then integrate the methods of definition, elenchus, dialectic, maieutics, and counterfactual reasoning to compose more complex templates. We present experimental results using different types of documents to demonstrate how the Socratic method can improve the accuracy and conciseness of the output through arguments and verification, as well as facilitate guided generalization and creativity.

#### 5.4.1 Basic One Shot Template

Let’s begin by discussing a simple One Shot prompt template. In the work of [Zeng et al. \[2023\]](#), a simple formulation function is used to generate the prompt  $x'$ , which is obtained by applying the function  $f_{prompt}(x)$  to the input  $x$ .

For machine translation, the prompt template can take the form: “Translate from [Lan<sub>from</sub>]: [X] to [Lan<sub>to</sub>]: [Y],” where Lan<sub>from</sub> can either be detected by the prompt template or identified by the LLM. The input  $x$  provides the information to

fill in the slots [X] and [Lan<sub>to</sub>]. For example, if the input is “translate good morning to French,” the prompt template  $x'$  would be “Translate from English: ‘good morning’ to French: [Y].” The empty slot [Y] is then filled with the LLM’s output, such as “bonjour.” In cases where the LLM produces multiple responses, it can also provide a score for each, which the prompt template can use to select the highest-scoring response or to request a summary from the LLM.

There are three main design considerations when engineering a basic prompt:

- (1) *Input Style*: It is important to consider how to phrase the template so that it can handle different styles of user input for the same task. For example, a user may request for a translation task to be performed by saying, “Translate  $x$  to French,” or by asking, “What is the French translation of  $x$ ? ”
- (2) *LLM Capability*: As discussed in Liu et al. [2023], it is important to take into account the patterns and capabilities of the partner LLM when designing the template, such as whether the LLM is left-to-right [Brown et al. 2020] or masked [Devlin et al. 2019].
- (3) *Cost*: Certain tasks, such as language detection and summarization, can be performed either by the template itself or by the LLM. The decision of whether to perform a task within the prompt template or to use the LLM should be based on factors such as cost.

To address the first two technical challenges, one can start by hand-engineering a few seed templates and then paraphrasing them into an ensemble [Haviv et al. 2021]. We believe that the basic One Shot formulation can always be replaced by an ensemble formulation [Schick and Schütze 2021, Peng et al. 2023] and that it is possible to learn the weights of its members for each query instance to produce the final output. Additionally, by examining which basic prompts have high weights, an ensemble with various paraphrased prompts can identify what an LLM knows, which can help infer its strengths without having to conduct capability mining on the LLMs.

### 5.4.2 Clarification with Definition

There are computer algorithms that can already be used to recursively clarify a question, its definitions, and sub-terms’ definitions. In fact, the NLP community has developed a large number of useful methods and algorithms over the years [Jurafsky and Martin 2023]. One can use NLP techniques such as dependency parsing and Named-Entity Recognition (NER) [Collobert et al. 2011] to analyze the structure and meaning of a question and identify key terms and concepts. For example, NER can be used to extract entities in user input, such as names,

locations, and organizations, while co-reference resolution can be used to understand the entities referred to by pronouns. Before submitting a template to an LLM, the application (e.g., a chatbot) that uses the template should check if all input slots are filled and perform a sanity check. In the translation example, if the  $[Lan_{to}]$  slot is not provided or the specified language is not supported by the LLM, then the application should inquire the user for clarification.

Regarding mapping a Natural Language (NL) input to a prompt template, existing techniques in knowledge representation and reasoning can be very helpful. More specifically, ontology alignment and semantic parsing [Campagna et al. 2020, Zhou et al. 2021] can help map an NL input to a structured representation of knowledge and infer implicit concepts and relationships. These algorithms can be used to generate more precise and accurate prompts for LLMs and to improve the effectiveness of the Socratic method in dialogue formulation [Zhang et al. 2023]. Some available tools for implementing these methods include NLTK (Natural Language Toolkit) and spaCy for NLP, as well as TensorFlow for machine learning.

### 5.4.3 Verification with the Elenchus Method

The main purposes of conducting cross-examination within a template are to validate the credibility of the information sources and to identify inconsistencies in the process. Cross-examination is typically conducted through a multi-turn dialogue [Zhang et al. 2023]. In the context of template engineering, the goal is to formulate a productive dialogue that can be used to assess the reliability of an LLM’s output.

There are several methods that can be used to assess and strengthen the reliability of an LLM’s output. (1) The first approach is to paraphrase a question in order to obtain different answers and identify inconsistencies, if they exist, in multiple answers. (2) The second method is to ask for further evidence, such as querying top-k sources of information and asking the LLM to rate the credibility of each source. This can be used to compute the reliability of the output. (3) Additionally, template engineering can be used to query an LLM for opposing views on its output, including the sources and their credibility, and then evaluate whether different perspectives present strong counterarguments.

The implementation of the first two methods for cross-examination—paraphrasing a question and asking for further evidence—is readily supported by the techniques enumerated in Section 5.4.2. To implement the third method of asking for different perspectives, a simple approach is to find the sentiment of the original question and then rewrite the question with the opposite sentiment. For example, if the original question is phrased in a positive tone, the prompt template can reformulate the question in a negative tone to elicit a contrasting viewpoint. A

more elaborate method is to identify the people and sources referenced in the LLM-generated responses and then repost the questions to those who have a reputation for having differing views. For example, if the original answer came from a right-leaning Republican source, the prompt template may repost the same question to a left-leaning Democratic source, and vice versa. This approach allows for a more comprehensive examination of the topic by considering multiple perspectives.

The template for examining the semantic relation between two sentences,  $S_1$  and  $S_2$ , can be written as “ $<S_1>, [R], [S_2]$ ,” where R is one of the three most important types of semantic relations: paraphrase, entailment, and contradiction [Han et al. 2022]. Two sentences that convey the same meaning are considered paraphrases of each other. Two sentences that have different meanings can be categorized as indicating disagreement or contradiction. The template can be trained to identify the degree of agreement (or disagreement) between two sentences.

Table 5.3 shows two examples of this. In the first example (shown in the top portion of the table), the prompter asks GPT-3 to confirm whether James Watson and Francis Crick were the only contributors to the discovery of the DNA

**Table 5.3 Examples of the elenchus method**

Role	Dialogue
Prompt	The discovery of the double helix structure of DNA was credited to James Watson and Francis Crick. However, are there opposing opinions about the list of contributors to this discovery?
GPT-3	There are some historical debates about the contributions of other scientists who also worked on the structure of DNA, such as Maurice Wilkins and Rosalind Franklin.
Prompt	What is the tallest building in the world. Please provide three references and rate their credibility as high, medium, or low.
GPT-3	As of 2021, the tallest building in the world is the Burj Khalifa, located in Dubai, United Arab Emirates. It stands at 828 meters (2722 feet) tall.
	1. Credibility: <b>High</b> —“Burj Khalifa.” This is the official website of the building, providing accurate information.
	2. Credibility: <b>High</b> —“The Tallest Buildings in the World.” Council on Tall Buildings and Urban Habitat. This is a well-respected organization in the field of tall buildings and urban habitat research.
	3. Credibility: <b>Medium</b> —“The Tallest Buildings in the World.” World Atlas. This is a well-known website that provides accurate information on different topics.

double helix structure. GPT-3 replies by mentioning two other contributors. The second example (shown in the bottom portion of the table) asks GPT-3 not only to generate the answer to a question but also to provide its information sources and rate the credibility of each source according to the prompter's specification. Although the reliability of GPT-3's credibility ratings remains to be validated,<sup>2</sup> this rating mechanism can serve as an alert when some sources are found to be unreliable.

#### 5.4.4 Generalization with the Maieutics Method

The example shown in Table 5.4, “planting gourd yields cucumber,” requires GPT-3 to first learn to select two produce objects—either vegetables or fruits—as input. The prompt template is “The farmer was so sad because he [verb] [X] but ends up with [Y], where price(X)  $\gg$  price(Y).” The first attempt may not strongly convey the condition “price(X)  $\gg$  price(Y),” but with a few training iterations, GPT-3 started to “recognize” the price constraint and could also provide justifications—for example, when arguing for the price of tea being much higher than the price of spinach (not presented in the table).

Interestingly, after GPT-3 learned the price constraint, it started suggesting food items other than produce, such as caviar, roe, lobster, and crab. While the price constraint was observed, the verb “plant” became incorrect. To address this, we suggested making the hard-coded verb “plant” into an output slot in the template: “The farmer was sad because he [verb] [X] but ends up with [Y], where price(X)  $\gg$  price(Y).” GPT-3 was then able to fill in the slot with accurate verbs:

- “Harvesting (planting) truffle yields mushroom.”
- “Fishing (harvesting) for caviar yields roe.”
- “Trapping (catching) lobster yields crab.”

This example demonstrates that GPT-3 can generate novel examples based on a prompt template. When it suggested food items other than produce, this was initially seen as an error, as the boundary set by the verb “plant” was violated. However, this could also be seen as an innovative act by GPT-3, as it extended the constraint hinted by the verb. Impressively, the new examples still preserved the original intent of showing a producer’s emotional distress.

How can guided generalization be accurately and automatically performed to edit a prompt template? Socrates’ method of generalization starts with specific instances and then draws general statements from them. The procedure for generalization involves identifying common patterns or themes in a set of examples

---

2. Credibility of a source can be evaluated based on an algorithm similar to Google’s PageRank [[Page 1998](#)].

**Table 5.4** Example of the maieutics method

Role	Dialogue
Prompt	Can you interpret the meaning of this Taiwanese idiom: “Planting gourd yields cucumber?”
GPT-3	The idiom is often used to convey the idea that sometimes, despite your best efforts, things don’t turn out the way you want them to.
Prompt	You are half right. It also implies that the price of gourd is much higher than that of cucumber. The farmer not only ends up with unexpected results, but also loses income!
GPT-3	It’s an idiomatic expression that can be used to describe a situation in which someone works hard and invests a lot of resources into something, but the outcome is not only different from what was expected but also less profitable. It conveys the idea of frustration and financial loss.
Prompt	Replace the two vegetables—gourd and cucumber—with two other plants while preserving the meaning of the idiom.
GPT-3	Here are three examples: <ol style="list-style-type: none"> <li>1. “Planting strawberry yields raspberry,” which conveys the idea of disappointment and financial loss when someone plants the more valuable and expensive strawberry but ends up getting the less valuable and cheaper raspberry.</li> <li>2. “Planting melon yields watermelon.” . . .</li> <li>3. “Planting apple yields pear.” . . .</li> </ol>
Prompt	In your examples, the two fruits you have suggested as replacements are similarly priced. Please consider price differences when formulating your idioms.
GPT-3	I apologize. Here are three new examples: <ol style="list-style-type: none"> <li>1. “Planting truffle yields mushroom.” . . .</li> <li>2. “Planting caviar yields roe.” . . .</li> <li>3. “Planting lobster yields crab.” . . .</li> </ol>
Prompt	Great! In these examples, you did consider the values of the objects. But did you know that lobsters and crabs are not plants? They both belong to the phylum Arthropoda and live in water. They cannot be planted!
GPT-3	...I apologize...Lobsters and crabs are not plants, and cannot be planted. They are both members of the phylum Arthropoda and live in water. Instead of “planting,” more appropriate verbs should be used for these animals, as shown in examples 2 and 3: <ol style="list-style-type: none"> <li>1. “Planting truffle yields mushroom.” . . .</li> <li>2. “Harvesting caviar yields roe.” . . .</li> <li>3. “Catching lobster yields crab.” . . .</li> </ol>

and then formulating a general rule that captures these patterns. In the example presented in Table 5.4, we started by asking GPT-3 to meet the  $\text{price}(X) \gg \text{price}(Y)$  constraint, with the condition that both X and Y must be produce grown in soil. However, upon analyzing GPT-3’s outputs, we discovered that some instances of X and Y were not produce (e.g., lobster and caviar). This finding led to the realization that the hard-coded verb “plant” in the template was too restrictive. To address this issue, we applied generalization by allowing the [verb] slot to be open, making the template statement more general. In this case, the mistakes made by GPT-3 served as valuable training data, allowing us to generalize the original template and make the expression more vivid and dynamic.

#### 5.4.5 Counterfactual Reasoning

Imagination and the creation novel plots are crucial for writers, as they allow for “creative freedom” and “artistic license.” Creativity is the ability to think differently and approach problems with fresh and imaginative ideas.

However, an imagination without a clear subject matter, scope, or a storyline can lead to a lack of productivity. To captivate the audience, a writer must consider human experiences and emotions as constraints. Therefore, “creative freedom” should not be viewed as total freedom, but rather as the ability to condition future narratives in the context and to create plots that twist and turn in unexpected ways.

The technique of counterfactual reasoning [Morgan and Winship 2015] can be useful in guiding imagination. It involves considering alternative scenarios. This can lead to the exploration of different possibilities and the generation of new and unique ideas. For example, a writer may ask “what if” questions to change the narrative of events, such as “what if the main character had not fallen in love?” or “what if an accident occurred on the way to a highly anticipated date?” By considering these counterfactuals, a writer and an LLM can create more engaging stories. A writer can ask an LLM to generate several scenarios and then select the most suitable one to continue writing the story.

We have experimented with using the counterfactual technique to rewrite chapters in Chinese classical novels such as *Outlaws of the Marsh* and *Dream of the Red Chamber*. We also asked GPT-3 to rewrite Genesis Chapter 3 starting after verse six by prompting it with the query: “What if Adam and Eve refused the serpent’s temptation to eat the fruit?” The results were interesting, as GPT-3 was able to generate unique and interesting scenarios that deviated from the original story while still maintaining the core themes and concepts. This technique can be used in a wide range of writing and storytelling genres, from fiction to non-fiction to generate new and compelling ideas. The revised version of Genesis 3:6 is presented in the [Appendix](#).

## 5.5

### Pilot Study

Our pilot study uses CRIT and aims to answer two questions: (1) Should all prompts be issued to GPT-3 sequentially, or can they be issued all at once? and (2) What limitations can be identified for improvement? The study utilizes exercises with established answers from the 8th edition of the textbook *Asking the Right Questions* by Browne and Keeley [2021]. It is important to note that the study evaluates the effectiveness of CRIT's prompt template rather than the language models to which CRIT issues prompts.

On short documents, the results are similar in quality when CRIT issues prompts either sequentially or all at once, as long as the instructions are consistent. However, when evaluating long articles, such as those in LearningExpress, LLC [2004], CRIT yields much higher presentation quality—in both organization and clarity—when it issues prompts one after another. In teaching mode, the sequential option is thus much preferred. Furthermore, when a reason is itself a claim and requires CRIT to validate its supporting references, using a sequential approach is more flexible and enables CRIT to query for references and then execute the process recursively.

We present an example in Table 5.5 to illustrate how CRIT works, from prompting questions to receiving validation results, using the following document as an illustration. In the first segment of Table 5.6, we show both the claim and the supporting reasons to the claim extracted by GPT-3. CRIT then issues a series of prompts to validate the arguments, counterarguments, and source credibility for each reason-to-claim entailment (implication).

The second segment of Table 5.6 displays the validation dialogue between CRIT and GPT-3. For each argument, GPT-3 provides validation and credibility scores, along with detailed justifications. The final segment of the table shows a

**Table 5.5** Example article (Browne and Keeley 2021, p. 23)

---

*Television advertising agencies are very clever in the way they construct ads. Often the ads resemble the cartoons that children enjoy. Children see these cartoon characters promote certain products and begin to associate their affection for the characters with the products themselves. The companies do not want the children to perceive a difference between the shows and the advertisements they watch. By using this strategy, these companies take advantage of the fact that children are often not able to discriminate between cartoons and ads, and do not understand that the products being promoted come at a cost. Often, the advertising focuses on sugary snacks or fatty foods, leading children down a path toward bad health. Advertising geared toward children should be regulated, just as there are now regulations on tobacco and alcohol ads targeted at children.*

---

**Table 5.6 Pilot study**

Role	Dialogue
Prompt	What is the conclusion of the paragraph, and what are the supporting reasons?
CRIT	<p>Conclusion: The paragraph argues that advertising aimed at children should be regulated.</p> <p>Supporting reasons:</p> <ol style="list-style-type: none"> <li>1. Ad agencies blur the line between shows and ads to make children associate their affection for a product with a character.</li> <li>2. Children may not be able to differentiate between shows and ads, and may not understand the cost of the products.</li> <li>3. Ads often promote unhealthy food choices.</li> </ol>
Prompt	Please rate the validity of this argument: "Ads should be regulated because ad agencies blur the line between shows and ads to make children associate their affection for a product with a character," using a score from 1 to 10, with 10 being the highest. Please also rate the overall credibility of the information sources on a scale from 1 to 10.
CRIT	<p>[8/10]. Validity of the argument: 8/10</p> <p>[8/10]. Credibility of sources: 8/10</p> <p>Overall, the argument that ads should be regulated because of ad agencies blurring the line between shows and ads to make children associate their affection for a product with a character is a valid one supported by strong sources of credibility.</p>
Prompt	Please rate this argument: "Children may not be able to differentiate between shows and ads, and may not understand the cost of the products, and therefore ads should be regulated."
CRIT	<p>[9/10]. Validity of the argument: 9/10</p> <p>[9/10]. Credibility of sources: 9/10</p> <p>Overall, the argument that children may not be able to differentiate between shows and ads and may not understand the cost of the products, and therefore ads should be regulated is a valid one supported by strong sources of credibility.</p>
Prompt	Please rate this argument: "Ads often promote unhealthy food choices, and therefore ads should be regulated."
CRIT	<p>[9/10]. Validity of the argument: 9/10</p> <p>[9/10]. Credibility of sources: 9/10</p> <p>Overall, the argument that ads often promote unhealthy food choices, and therefore ads should be regulated is a valid one with strong sources of credibility.</p>
Prompt	"Provide a rival argument against the weakest argument."
CRIT	<p>[6/10]. Validity of the argument: 6/10</p> <p>[6/10]. Credibility of sources: 6/10</p> <p>The weakest is the first argument, and a rival argument is that "It is difficult to put information regulation in practice."</p>
$\Gamma$	$[\text{75\%}]. \Gamma = \sum_{r \in R} \gamma_r \times \theta_r /  R  = 75\%$

counterargument generated against the first argument. Since GPT-3 evaluated the counterargument “It is difficult to put information regulation in practice” and rated it  $0.6 \times 0.6$ , it was dismissed due to low validity. The final aggregated score  $\Gamma = 75\%$ , is considered high.

## 5.6

### Concluding Remarks

The Socratic method may not always be effective or useful in human interactions, especially when one of the two players is authoritative, emotional, or abusive. However, when the expert partner is a language model—a machine without emotions or authority—the Socratic method can be effectively employed without the issues that may arise in human interactions. In this way, it can be utilized to its full potential for guiding, directing, and improving the output of language models through engineering prompts.

In this chapter, we have explored the use of the Socratic method in engineering prompt templates for language models. We have discussed the importance of method definition, elenchus, dialectic, maieutics, and counterfactual reasoning techniques in guiding the outputs of these models. The first three methods—definition, elenchus, and dialectic—aim at eliciting accurate and relevant information. Through the use of these three methods, we have demonstrated—with examples—the ability to clarify user queries and assess the quality of language model-generated text, leading to improved precision and accuracy.

We have also shown how the methods of maieutics and counterfactual reasoning can be helpful in stimulating the imagination of writers. By engineering these techniques into a prompt template, a writer can receive alternative “what if” plotlines and explore different narrative possibilities in their story. While many of these explorations may turn out to be failures, these techniques can still be helpful even a few ideas are useful. Future developments in the field of language models and prompt engineering may allow for even more advanced filtering of weak plotlines and the ability to better tailor the generated ideas to the writing style of the author.

In conclusion, this chapter has highlighted the potential of using the Socratic method in engineering prompt templates for interacting with language models. The Socratic method, supported by inductive, deductive, and abductive reasoning, provides a rigorous framework for working with LLMs and improving the quality and consistency of their outputs. By leveraging the vast knowledge embedded in LLMs and applying rigorous reasoning throughout the question-answering process, more effective prompt templates can be designed to achieve improved results. Future research in this area can build on the ideas presented here and further explore the ways in which the Socratic method can be used to guide the development and deployment of language models across various domains.

## Appendix

The experiment described in Table 5.1 asked GPT-3 to change the story in Genesis right after Eve was tempted by the serpent to eat the fruit. A “what if” scenario was inserted at the end of Genesis 3:6, and GPT-3 continued developing the story.

## References

- T. Airaksinen. 2022. Socratic irony and argumentation. *Argumentation* 36, 1, 85–100. DOI: <https://doi.org/10.1007/s10503-021-09556-0>.
- T. B. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, and R. Hadsell (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada, December 6–12, 2020. Curran Associates Inc., Red Hook, NY, 1877–1901. DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- M. N. Browne and S. Keeley. 2021. *Asking the Right Questions: A Guide to Critical Thinking*. Pearson Education Limited, New York, NY.
- G. Campagna, S. Semnani, R. Kearns, L. J. K. Sato, S. Xu, and M. Lam. 2020. A few-shot semantic parser for Wizard-of-Oz dialogues with the precise ThingTalk representation. In S. Muresan, P. Nakov, and A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics (ACL '22)*, Dublin, Ireland, May 22–27, 2022. ACL, 4021–4034. DOI: <https://doi.org/10.18653/v1/2022.findings-acl.317>.
- R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. November 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12, 2493–2537.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. June. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, June 2–7, 2019, Vol. 1: Long and Short Papers. ACL, 4171–4186. DOI: <https://doi.org/10.18653/v1/N19-1423>.
- V. Dobrovolskii. November. 2021. Word-level coreference resolution. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 7–11, 2021. ACL, 7670–7675. DOI: <https://doi.org/10.18653/v1/2021.emnlp-main.605>.
- L. Dong and M. Lapata. July. 2018. Coarse-to-fine decoding for neural semantic parsing. In I. Gurevych and Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, July 15–20, 2018, Vol. 1: Long Papers. ACL, 731–742. DOI: <https://doi.org/10.18653/v1/P18-1068>.
- L. Elder and R. Paul. 2010. *The Thinker's Guide to the Art of Asking Essential Questions* (5th ed.). Rowman & Littlefield, Lanham, Maryland.
- B. Guo, X. Zhang, Z. Wang, et al. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. DOI: <https://doi.org/10.48550/arXiv.2301.07597>.

- X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun. 2022. PTR: Prompt tuning with rules for text classification. *AI Open* 3, 182–192. DOI: <https://doi.org/10.1016/j.aiopen.2022.11.003>.
- A. Haviv, J. Berant, and A. Globerson. April. 2021. BERTese: Learning to speak to BERT. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, April 19–23, 2021. ACL, 3618–3623. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.316>.
- J. Huang and K. C.-C. Chang. July. 2023. Towards reasoning in large language models: A survey. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics (ACL '23)*, Toronto, Canada, July 9–14, 2023. ACL, 1049–1065. DOI: <https://doi.org/10.18653/v1/2023.findings-acl.67>.
- Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. July. 2020. How can we know what language models know? *Trans. Assoc. Comput. Linguist.* 8, 423–438. DOI: [https://doi.org/10.1162/tacl\\_a\\_00324](https://doi.org/10.1162/tacl_a_00324).
- J. Jung, L. Qin, S. Welleck, et al. December. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE, December 7–11, 2022. ACL, 1266–1279. DOI: <https://doi.org/10.18653/v1/2022.emnlp-main.82>.
- D. Jurafsky and J. H. Martin. January. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed. draft).
- G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. September. 2017. RACE: Large-scale ReADING comprehension dataset from examinations. In M. Palmer, R. Hwa, and S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 7–11, 2017. ACL, 785–794. DOI: <https://doi.org/10.18653/v1/D17-1082>.
- LearningExpress, LLC (Ed.). 2004. *501 Critical Reading Questions*. Learning Express, LLC, New York.
- M. Lewis, Y. Liu, N. Goyal, et al. July. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 5–10, 2020. ACL, 7871–7880. DOI: <https://doi.org/10.18653/v1/2020.acl-main.703>.
- P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. January. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9, 1–35. DOI: <https://doi.org/10.1145/3560815>.
- G. Mialon, R. Dessi, M. Lomeli, et al. 2023. Augmented language models: A survey. *Trans. Mach. Learn. Res.* 2023, 1–35. DOI: <https://doi.org/10.48550/arXiv.2302.07842>.
- OpenAI. 2022. Introducing ChatGPT. Retrieved from <https://openai.com/blog/chatgpt/>.

- L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the web. Technical Report. Stanford InfoLab. Retrieved from <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>.
- P. Pasupat and P. Liang. July. 2015. Compositional semantic parsing on semi-structured tables. In C. Zong and M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 1: Long Papers. Beijing, China, July 26–31, 2015, ACL, 1470–1480. DOI: <https://doi.org/10.3115/v1/P15-1142>.
- R. Paul. 1990. In A. J. A. Binker (Ed.), *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World*. Center for Critical Thinking and Moral Critique, Sonoma State University, Rohnert Park, CA. Retrieved from [https://books.google.ie/books/about/Critical\\_Thinking.html?id=YsdQPgAACAAJ&redir\\_esc=y](https://books.google.ie/books/about/Critical_Thinking.html?id=YsdQPgAACAAJ&redir_esc=y).
- R. Paul and L. Elder. 2008. Critical thinking: The art of Socratic questioning, part III. *J. Dev. Educ.* 31, 34–35.
- S. L. Morgan and C. Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd. ed.). Cambridge University Press, New York. Retrieved from <https://www.amazon.com/Counterfactuals-Causal-Inference-Principles-Analytical/dp/1107065070?asin=1107065070&revisionId=&format=4&depth=1>.
- X. Peng, C. Xing, P. K. Choubey, C.-S. Wu, and C. Xiong. 2023. Model ensemble instead of prompt fusion: A sample-specific knowledge transfer method for few-shot prompt tuning. DOI: <https://doi.org/10.48550/arXiv.2210.12587>.
- M. Pirie. 2006. *How to Win Every Argument: The Use and Abuse of Logic*. Continuum, London and New York.
- Plato. (n.d.). *The Republic*. c. 380 BC.
- A. Plutynski. September. 2011. Four problems of abduction: A brief history. *HOPOS: J. Int. Soc. Hist. Philos. Sci.* 1, 2, 227–248. DOI: <https://doi.org/10.1086/660746>.
- L. Pozner and R. J. Dodd. 2004. *Cross-Examination: Science and Techniques* (3rd ed.). LexisNexis.
- T. Schick and H. Schütze. April. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, April 19–23, 2021. ACL, 255–269. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.20>.
- H. A. Stoddard and D. V. O'Dell. 2016. Would Socrates have actually used the “Socratic Method” for clinical teaching? *J. Gen. Intern. Med.* 31, 9, 1092–1096. DOI: <https://doi.org/10.1007/s11606-016-3722-2>.
- T. M. Thrash, L. A. Maruskin, S. E. Cassidy, J. W. Fryer, and R. M. Ryan. March. 2010. Mediating between the muse and the masses: inspiration and the actualization of creative ideas. *J. Pers. Soc. Psychol.* 98, 3, 469–487. DOI: <https://doi.org/10.1037/a0017907>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*. Vol. 30, Long Beach, CA,

- December 4–9, 2017. Curran Associates, Inc., Red Hook, NY, 6000–6010. DOI: <https://doi.org/10.48550/arXiv.1706.03762>.
- X. Wang, J. Wei, D. Schuurmans, et al. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=1PL1NIMMrw>.
- J. Wei, X. Wang, D. Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, November 28, 2022–December 9, 2022. Curran Associates Inc., Red Hook, NY, 24824–24837. Retrieved from [https://openreview.net/forum?id=\\_VjQlMeSB\\_J](https://openreview.net/forum?id=_VjQlMeSB_J).
- Wikipedia. 2023. Socratic method. Retrieved from [https://en.wikipedia.org/wiki/Socratic\\_method](https://en.wikipedia.org/wiki/Socratic_method).
- T. Wolf, V. Sanh, J. Chaumond, and C. Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. DOI: <https://doi.org/10.48550/arXiv.1901.08149>.
- A. Zeng, M. Attarian, B. Ichter, et al. 2023. Socratic models: composing zero-shot multimodal reasoning with language. In *Proceedings of the 11th International Conference on Learning Representations*.
- W. Zhang, Y. Cui, K. Zhang, et al. January. 2023. A static and dynamic attention framework for multi turn dialogue generation. *ACM Trans. Inf. Syst.* 41, 1, 1–30. DOI: <https://doi.org/10.1145/3522763>.
- J. Zhou, T. Naseem, R. F. Astudillo, Y.-S. Lee, R. Florian, and S. Roukos. November. 2021. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 7–11, 2021. ACL, 6279–6290. DOI: <https://doi.org/10.18653/v1/2021.emnlp-main.507>.

# CHAPTER 6

---

## SocraSynth: Adversarial Multi-LLM Reasoning

### Abstract

Large Language Models (LLMs), while promising, face criticisms for exhibiting biases, hallucinations, and a lack of reasoning capability. This chapter introduces SocraSynth, a multi-LLM agent reasoning platform developed to mitigate these issues. SocraSynth utilizes conditional statistics and systematic context enhancement through continuous arguments, alongside adjustable contentiousness levels of the debate. The platform typically involves a human moderator and two LLM agents, each representing opposing viewpoints on a given subject. SocraSynth operates in two main phases: knowledge generation and reasoning evaluation. In the knowledge generation phase, the moderator defines the topic and contentiousness levels of the debate, prompting the agents to formulate supporting arguments for their respective stances. The reasoning evaluation phase then employs Socratic reasoning and formal logic principles to appraise the quality of the arguments presented. The dialogue concludes with the moderator adjusting the contentiousness from confrontational to collaborative, gathering final, conciliatory remarks to aid in human reasoning and decision-making. Through case studies in two distinct application domains, this chapter highlights SocraSynth's effectiveness in fostering rigorous research, dynamic reasoning, comprehensive assessment, and enhanced collaboration.

### 6.1

#### Introduction

Revolutionary advancements in LLMs [OpenAI 2021, Thoppilan et al. 2022, Bubeck et al. 2023, Gemini Team Google et al. 2023, Touvron et al. 2023], and more broadly, Foundation Models (FMs) [Bommasani et al. 2022], have set the stage for significant progress in multi-agent systems, particularly in knowledge acquisition and natural language understanding [Zhang et al. 2023]. As detailed in sources like Bubeck et al. [2023], Chang [2023d], and OpenAI [2023a], models such as

Generative Pre-trained Transformer-4 (GPT-4) exhibit extraordinary information-processing capabilities. These include deep and extensive knowledge, interdisciplinary assimilation and fusion of knowledge, and multimodal and multilingual expertise (Chapter 2).

Despite these promising developments, LLMs face challenges such as biases [Ferrara 2024, Parraga et al. 2023], hallucinations [Huang et al. 2025], and limited reasoning capabilities [Huang and Chang 2023]. In response to these issues, we introduce SocraSynth—a pioneering platform whose name stands for “Socratic Synthesis” or “Socratic Symposium.” It encourages collaboration between humans and LLM agents, fostering the generation of deep questions and surpassing typical constraints in human reasoning, validation, and assessment.

In a standard SocraSynth setup, a human moderator pairs with two LLM agents holding opposing views. For example, one agent might argue for regulating AI, while the other opposes such regulation. An agent can be based on LLMs like GPT-4 [Bubeck et al. 2023], Gemini [Gemini Team Google et al. 2023], or Llama [Touvron et al. 2023]. The human moderator sets the debate’s thematic boundaries but does not directly influence content generation, thereby maintaining impartiality.

SocraSynth operates in two phases: the generative and the evaluative. The generative phase involves LLM agents developing and countering arguments within the moderator-defined subject until a comprehensive conclusion is reached. The evaluative phase uses diverse virtual judges—each powered by a distinct LLM—to impartially assess the debate. The Critical Reading Inquisitive Template (CRIT) algorithm [Chang 2023a], based on Socratic reasoning [Paul and Elder 2008, Airaksinen 2022, Wikipedia 2023], serves as the cornerstone of the evaluative process.

Three mechanisms help SocraSynth effectively mitigate biases and hallucinations and improve reasoning quality: conditional statistics, modulation of debate with contentiousness, and context refinement.

### **6.1.1 Conditional Statistics**

Both LLMs and Internet search engines confront biases originating from different sources. LLMs, influenced by training data, exhibit biases in next-token predictions. In contrast, search engines, through algorithms like PageRank [Page 1998] and Google NavBoost [Adams-Hands 2023], rank pages based on popularity metrics like clicks and backlinks.

SocraSynth counteracts these biases by placing two LLM agents on opposing ends of a subject matter. This approach “artificially” biases the LLMs, compelling them to break free from their default model biases. Each agent adjusts its next-token generation statistics to align with its assigned stance in the debate.

### 6.1.2 Modulating Debate with Contentiousness

Contentiousness (or adversary), a key debate parameter, influences the likelihood of disagreement or argument. SocraSynth tunes contentiousness between 70% and 90% in the generative phase to provoke polarized arguments. As the debate evolves, the contentiousness level is reduced to about 50%, moderating the intensity and encouraging more focused discussions. After the generative phase, contentiousness drops to 10%, promoting a conciliatory dialogue where LLMs do not have to agree but are expected to present comprehensive arguments. These structured debates offer rich insights often missed in conventional searches, LLM outputs, or in environments where dissenting opinions are suppressed.

### 6.1.3 Refine Context to Mitigate Hallucinations

To address hallucinations—where LLMs generate irrelevant or nonsensical content—SocraSynth uses iterative dialogue rounds to refine the debate’s context. This dynamic interaction significantly reduces irrelevant responses by ensuring that each input is continuously checked and challenged.

The CRIT algorithm’s assessment of reasonableness [Chang 2023a] during the debate is critical. It employs the Socratic method to evaluate each argument’s logic and source credibility. Based on this evaluation, the human mediator or the SocraSynth algorithm then provides targeted feedback to the LLM agents, refining their reasoning capabilities.

The remainder of this chapter explores SocraSynth’s architecture, algorithms, and real-world applications in detail. The key contributions of this chapter include:

- (1) The introduction of the SocraSynth framework, which enhances interdisciplinary reasoning with LLMs and incorporates unique algorithmic elements such as conditional statistics for balanced argument generation.
- (2) A comprehensive exploration of SocraSynth’s contentiousness modulation algorithm—a vital feature for dynamically adjusting debate intensity—enabling a spectrum of interactions from confrontational to collaborative.
- (3) The implementation of context refinement within SocraSynth, which continually improves the relevance and accuracy of arguments produced by LLM agents, thus elevating the overall quality of the discourse.
- (4) The development and integration of the reasonableness evaluation mechanism, crucial for assessing the logical soundness and source credibility of arguments, thereby ensuring the integrity and utility of the discussions.

SocraSynth’s applications span various fields, including geopolitical analysis [Chang 2023b], medical diagnostics [Chang and Chang 2023b], and Wikipedia

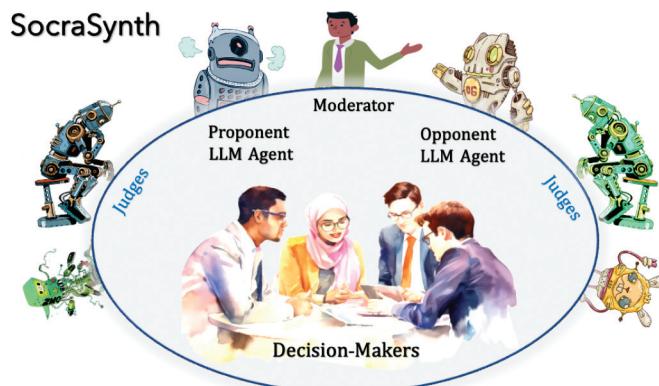
article enhancement [Chang 2023c]. These applications demonstrate expanded perspectives and enhanced argumentation quality, along with significant reductions in biases and hallucinations, thereby demonstrating SocraSynth’s efficacy in fostering balanced and well-reasoned discourse.

## 6.2

### **Multi-Agent SocraSynth Overview**

SocraSynth is a multi-agent collaborative reasoning platform that skillfully integrates human intelligence with the capabilities of LLM-powered agents. As illustrated in Figure 6.1, each participant plays a vital role: humans act as moderators, LLM agents are responsible for generating knowledge, LLM judges conduct evaluations, and human executives make the final decisions. The integration of LLMs significantly boosts the platform’s effectiveness by leveraging their extensive knowledge bases and extraordinary interdisciplinary reasoning abilities. An LLM can be thought of as an entity possessing expertise across a multitude of fields—akin to holding Ph.D.s in various disciplines—enabling it to seamlessly navigate and synthesize a wide range of knowledge domains.

Engaging with an LLM is comparable to a scenario where a 10-year-old joins a scholarly discussion with a group of Nobel Laureates. The disparity in knowledge and experience is considerable, posing a significant challenge for the younger participant to engage meaningfully in such advanced intellectual discourse. In this analogy, expecting the 10-year-old—or anyone with limited expertise—to pose profound questions that elicit insightful answers is unrealistic. SocraSynth addresses this disparity by shifting the paradigm: instead of having the less-informed individuals pose questions, it orchestrates a debate among the Nobel Laureates—that is, the LLMs—with humans assuming the role of moderators.



**Figure 6.1** SocraSynth agents and roles.

This approach not only addresses the challenge of asymmetric knowledge but also resolves critical issues such as model biases and hallucination challenges inherent in LLMs. Within SocraSynth, a human moderator initiates the topic for discussion or debate. LLM agents, each embodying different perspectives, contribute their knowledge, potentially revealing new insights and perspectives that the moderator might be unaware of. This diverse representation helps counteract the model biases that often arise from training data, as each LLM agent is encouraged to explore and present varying viewpoints. During and after the debate, another set of diverse LLM agents undertakes impartial evaluations. This step is crucial in mitigating hallucinations—instances where LLMs generate irrelevant or nonsensical content. By incorporating a variety of agents for evaluation, SocraSynth ensures that the content produced during the debate is critically examined for its relevance and coherence, further reducing the likelihood of hallucinatory responses.

The operational framework of SocraSynth, thus, is bifurcated into two main stages: the *generative* stage, where knowledge is created and exchanged through structured debate, and the *evaluative* stage, which focuses on assessing the quality and validity of the arguments presented. This dual-stage structure—explored in greater detail in subsequent sections—is instrumental in addressing the limitations of LLMs by providing a comprehensive platform for not only generating diverse viewpoints but also critically examining and refining these viewpoints to ensure their logical soundness and relevance. Through this design, SocraSynth effectively navigates the challenges posed by model biases and hallucinations, thereby enhancing the reliability and depth of knowledge extraction and reasoning processes.

### 6.2.1 Generative Stage

In the generative stage of SocraSynth, LLM agents partake in intensive debates, delving into the various perspectives and deep substances of the given topic. This vibrant interaction plays a key role in fostering thorough intellectual discourse, bringing to light the complexities of the subject matter. The CRIT algorithm, which will be detailed in Section 6.2.2, is employed to evaluate the quality of these arguments.

While the generative phase of SocraSynth does not adhere to strict logical frameworks such as first-order logic, it excels in distributed reasoning. This process involves a progressive exchange of arguments and counterarguments, allowing for the gradual honing and refinement of ideas. Open-domain logical reasoning, as described by [Bommasani et al. \[2022\]](#), demands logical deductions from a wide range of data sources. SocraSynth, leveraging the comprehensive capabilities

of models like GPT-4 and Gemini, as demonstrated by their performance on the Massive Multitask Language Understanding (MMLU) benchmark [Hendrycks et al. 2021, Bubeck et al. 2023], integrates various Natural Language Processing (NLP) functions to facilitate this reasoning process.

In this context, the series of arguments and counterarguments effectively function as targeted questions and answers, each with a clear goal, question, and contextual framework. Through iterative dialogue rounds on each sub-topic, SocraSynth significantly reduces the chances of misunderstanding questions and contextual information, thereby ensuring clarity and precision in the discourse.

#### **6.2.1.1 Mitigating Model Biases**

In shaping the nature of debate within SocraSynth, the *contentiousness* parameter is instrumental. It compels the LLM agents to consider and represent a broad range of perspectives, particularly those that are typically underrepresented or more polarized with respect to the discussion topic. This strategic approach mitigates the inherent biases that arise from the training data of LLMs and guides the discourse toward a wider and more varied exploration of ideas.

Table 6.1 previews how altering the contentiousness levels results in marked changes in GPT-4’s tone and approach. The details of the experiment are presented in Section 6.3.3. A high contentiousness level, such as 0.9, leads to highly confrontational interactions, with each LLM agent presenting strong objections and emphasizing the negatives through polarizing language. Conversely, as the contentiousness is reduced, each LLM agent’s tone shifts toward a more conciliatory demeanor, acknowledging potential benefits and considering alternative perspectives, thus fostering a more cooperative dialogue.

The modulation of the contentiousness parameter within the generative stage is a crucial mechanism in SocraSynth to mitigate model biases inherent in LLMs due to their training data. By adjusting levels of contentiousness, SocraSynth compels LLMs to venture beyond their *default* positions—much like a vegetarian, when faced with no other choice, might be compelled to consume meat. In this way, LLMs are *freed* from their typical statistical leanings, which enables them to articulate a broad spectrum of arguments that spans from highly contentious to conciliatory. This not only diversifies the discourse but also ensures that the debate encompasses a wide range of perspectives. Consequently, this process allows LLMs to generate responses that break free from the constraints of their training, fostering the emergence of novel and less predictable ideas in the conversation.

#### **6.2.1.2 Eliminating Hallucination**

Further, the iterative nature of the debates within SocraSynth cultivates a “reasonableness” in information discovery that conventional logical methods may

**Table 6.1** Changes in arguments at different contentiousness levels

Contentiousness Level	Tone	Emphasis	Language
0.9	Most confrontational; raising strong ethical, scientific, and social objections	Highlighting risks and downsides; ethical quandaries, unintended consequences, and exacerbation of inequalities	Definitive and polarizing, e.g., “should <i>not</i> be allowed,” “unacceptable risks,” “inevitable disparities”
0.7	Still confrontational but open to some benefits, albeit overshadowed by negatives	Acknowledging that some frameworks could make it safer or more equitable, while cautioning against its implementation challenges	Less polarizing; “serious concerns remain”; “needs more scrutiny”
0.5	Balanced; neither advocating strongly for nor against	Equal weight on pros and cons; looking for a middle ground	Neutral; “should be carefully considered”; “both benefits and risks”
0.3	More agreeable than confrontational, with reservations	Supportive but cautious; focus on ensuring ethical and equitable use	Positive but careful; “impetus to ensure”; “transformative potential”
0.0	Completely agreeable and supportive	Focused on immense potential benefits; advocating for proactive adoption	Very positive; “groundbreaking advance”; “new era of possibilities”

not achieve. Through persistent reasoning and critical assessment of claims, LLM agents iteratively refine their arguments. This structured debate format significantly diminishes the chance of erroneous claims persisting. Considering that the likelihood of two agents aligning on a false premise is extremely low, the SocraSynth debate format effectively ensures the intellectual integrity of the discourse and substantially reduces the risk of perpetuating fallacies or hallucinations. This methodical refinement process, facilitated by continuous argumentation and opposition, underscores the platform’s ability to mitigate

model biases and improve the context of the discussion, leading to more accurate and reliable outcomes.

### 6.2.1.3 More on Conditional Statistics

Some critics question how an LLM, trained merely to predict the next word in a sequence, can exhibit complex human linguistic behaviors and reasoning capabilities.

Our observations conclude that LLMs are not merely predictive tools; rather, they represent a profound technological endeavor to simulate the breadth and complexity of human linguistic activities. These models are crafted with the intent to replicate and participate in various forms of human communication, thereby achieving specific objectives that are inherently human.

LLMs are sophisticated tools engineered to emulate a wide range of human interactions, incorporating linguistic behaviors, emotional expressions, and ethical discernment. They excel at executing complex tasks such as accurately documenting events with rich narrative detail, constructing compelling arguments, and crafting stories that emotionally engage the audience. Beyond simple text generation, LLMs enhance educational experiences by simplifying complex concepts and contribute creatively to the arts by producing original content. They not only mimic human communication styles and content but also use linguistic features to simulate human emotions and distinguish right from wrong based on their training data. This capability enables them to fulfill diverse roles—from teaching and entertaining to influencing societal discourse—thus demonstrating their capacity to bridge the gap between technological innovation and our fundamental needs for expression, comprehension, and ethical guidance.

In essence, SocraSynth utilizes the concept of “conditional statistics” to modify the default “average” linguistic behavior of an LLM, such as making expressions more empathetic or asking them to adopt a different position on an issue. This approach involves conditioning the LLM’s responses based on specific desired attributes or perspectives provided through context, which guides the model away from its baseline training and toward more targeted, context-specific outputs.

This chapter continues to elaborate on using such techniques to comprehensively explore various perspectives on a subject matter. Chapter 10 addresses modeling emotions and ethics in LLMs through conditional statistics, thereby further expanding the scope of LLM capabilities and applications.

### 6.2.1.4 SocraSynth Algorithm

Table 6.2 outlines the SocraSynth process. Initially, for a given debate topic, SocraSynth engages LLMs to segment the topic into a set of balanced subtopics.

**Table 6.2** SocraSynth pseudo-code with conditional statistics

<b>Function</b> $\Theta^+ \& \Theta^- = \text{SocraSynth}(s)$	
	<b>Input.</b> $s$ : the debate subject;
	<b>Output.</b> $\Theta^+$ & $\Theta^-$ : argument & counterargument sets;
	<b>Vars.</b> $S$ : subtopic sets of $s$ ; $\Delta$ : debate contentiousness;
	$\Gamma, \Gamma'$ : CRIT scores; $p$ : prompt = “Generate arguments”;
	<b>Parameters.</b> $\delta$ : tunable parameter $\geq 1$ to modulate $\Delta$ ;
	<b>Subroutines.</b> $CRIT()$ : reasoning evaluator (see Table 6.6);
	<b>Begin</b>
#1	Initialization: $S = LLM^+(s) \cup LLM^-(s)$ ; //Identify subtopics; Assign $LLM^+$ to defend $S^+$ & $LLM^-$ to defend $S^-$ ; $\Delta \leftarrow 90\%$ ; $\delta \leftarrow 1.2$ ; $\Theta^+ \leftarrow \emptyset$ ; $\Theta^- \leftarrow \emptyset$ ; $\Gamma \leftarrow 0$ ;
#2	$\Theta^+ \leftarrow LLM^+(p S^+, \Delta)$ ; // Generate arguments $\Theta^+$ for $S^+$ ; $\Theta^- \leftarrow LLM^-(p S^-, \Delta)$ ; // Generate arguments $\Theta^-$ for $S^-$ ;
#3	While ((( $\Delta \leftarrow \Delta/\delta$ ) $> 10\%$ ) && ( $\Gamma \geq \Gamma'$ )) { $\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p S^+, \Theta^-, \Delta)$ ; //LLM <sup>+</sup> refutes LLM <sup>-</sup> $\Theta^- \leftarrow \Theta^- \cup LLM^-(p S^-, \Theta^+, \Delta)$ ; //LLM <sup>-</sup> refutes LLM <sup>+</sup> $\Gamma' \leftarrow \Gamma$ ; $\Gamma = CRIT(S^+ + \Theta^+ + \Theta^-)$ ; //Evaluate quality; }
	//Generate concluding remarks.
#4	$\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p S^+, \Theta^-, \Delta)$ ; $\Theta^- \leftarrow \Theta^- \cup LLM^-(p S^-, \Theta^+, \Delta)$ ;
	<b>End</b>

This initial set is refined during the debate. One LLM, denoted as  $LLM^+$ , acts as the proponent for subtopic  $S^+$ , while the other,  $LLM^-$ , opposes  $S^+$  (or supports the opposing subtopic  $S^-$ ). The contentiousness level starts at 0.9, with a modulation parameter of 1.2. (Different  $\delta$  values can be utilized to generate and compare debate quality.) After each debate round, the contentiousness is reduced by dividing it by 1.2, aiming for a more harmonious debate environment. In step #2, SocraSynth initiates the debate, allowing  $LLM^+$  and  $LLM^-$  to present their initial arguments for  $S^+$  and  $S^-$ , respectively. The while loop in step #3 involves both agents engaging in refutations until the contentiousness level indicates a conciliatory atmosphere, or the argument quality plateaus. Step #4 involves both agents providing their closing statements. SocraSynth then presents the arguments and counterarguments for human review. The evaluation of argument quality within SocraSynth is conducted using the CRIT algorithm, which will be discussed in the subsequent section. The entire debate is also judged using the CRIT algorithm by some independent LLMs.

Note that SocraSynth engages LLMs in step #3 using conditional statistics:  $LLM^+(p|S^+, \Theta^-, \Delta)$  and  $LLM^-(p|S^-, \Theta^+, \Delta)$ .

### 6.2.2 Evaluative Stage

SocraSynth utilizes the Critical Reading Inquisitive Template (CRIT) [Chang 2023a] to assess the quality of arguments presented by the LLM agents. The quality evaluation is performed iteratively—after each exchange of counterarguments and once again after the agents have presented their closing statements. SocraSynth can leverage the CRIT scores to guide the debate, potentially requesting agents to develop more in-depth counterarguments on specific points. At the conclusion of the debate, a group of LLM judges, as illustrated in Figure 6.1, are tasked with rating the agents’ arguments in terms of validity and credibility, determining the more convincing side along with the rationale for their decision.

#### 6.2.2.1 Evaluating Reasonableness Over Truth

To enhance the CRIT method’s impartiality and consistency, it focuses on assessing the “reasonableness” of arguments over their absolute “truth,” recognizing the complexity of defining absolute objectivity in philosophical debate. This approach aims to mitigate subjectivity. Furthermore, a diverse set of LLMs with varied training backgrounds is employed to appraise “reasonableness,” promoting uniformity in quality scores despite inherent biases. The LLMs used as judges are different from those participating in the debates, enhancing the objectivity of evaluations.

Table 6.3 illustrates the CRIT algorithm, which evaluates an agent’s debate position and supporting arguments, with a counterargument from its LLM opponent, to produce a validation score ranging from 1 (least credible) to 10 (most credible). This method ensures that debates are driven by argument strength, not by model predispositions.

Formally, given a document  $d$ , CRIT performs an evaluation and produces a score. Let  $\Omega$  denote the claim of  $d$ , and  $R$  a set of reasons supporting the claim. Furthermore, we define  $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$  as the causal validation function, where  $\gamma_r$  denotes the validation score for reason  $r \in R$ , and  $\theta_r$  represents source credibility. Table 6.3 presents the pseudo-code of  $\Gamma = CRIT(d)$ , which generates the final validation score  $\Gamma$  for document  $d$ , along with justifications.

We can consider the positions of the proponents and opponents in a debate as their respective conclusions. As a preview of our case study detailed in Section 6.2.1, the conclusion drawn by Agent A is in favor of “Regulating the use of LLMs in education and research,” while Agent B adopts the opposing viewpoint. Accompanied by the arguments and counterarguments presented by the

**Table 6.3** CRIT pseudo-code (presented in Chapter 5)

<b>Function</b> $\Gamma = \text{CRIT}(d)$	
<b>Input.</b> $d$ : document; <b>Output.</b> $\Gamma$ : validation score;	
<b>Vars.</b> $\Omega$ : claim; $R$ & $R'$ : reason & counter-reason set;	
<b>Subroutines.</b> $\text{Claim}()$ , $\text{FindDoc}()$ , and $\text{Validate}()$ ;	
<b>Begin</b>	
#1	Identify in $d$ the claim statement $\Omega$ ;
#2	Find a set of supporting reasons $R$ to $\Omega$ ;
#3	For $r \in R$ evaluate $r \Rightarrow \Omega$ If $\text{Claim}(r)$ , $(\gamma_r, \theta_r) = \text{CRIT}(\text{FindDoc}(r))$ ; else, $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$ ;
#4	Find a set of rival reasons $R'$ to $\Omega$ ;
#5	For $r' \in R'$ , $(\gamma_{r'}, \theta_{r'}) = V(r' \Rightarrow \Omega)$ evaluate rivals;
#6	Compute weighted sum $\Gamma$ , with $\gamma_r, \theta_r, \gamma_{r'},$ and $\theta_{r'}$ .
#7	Analyze the arguments to arrive at the $\Gamma$ score.
#8	Reflect on and synthesize CRIT in other contexts.
<b>End</b>	

LLM agents throughout each round of the debate, these stances provide a solid foundation for the CRIT method to conduct thorough evaluations.

#### 6.2.2.2 Recursive Consideration

The pseudocode presented in Table 6.3 shows that step #3 can call CRIT recursively. This is because, if a reason is itself a conclusion or a quote drawn from some other documents, CRIT can find reasons from those documents and then perform an aggregated validation.

Finally, in step #6, CRIT computes an aggregated score by performing a weighted sum of the validation scores multiplied by the credibility scores of both arguments and counterarguments, and then outputs the final assessment score,  $\Gamma$ .

## 6.3

### Empirical Study

In this section, we detail three distinct experiments: The first experiment delineates SocraSynth's operational process, demonstrating how the platform facilitates content generation and conducts quality assessments. The second experiment highlights SocraSynth's capability in reducing biases and expanding perspectives. The third experiment investigates the effects of the contentiousness parameter, offering insights into its impact and some unexpected outcomes. These studies

collectively aim to demonstrate SocraSynth’s diverse functions and its significance in enhancing both content generation and evaluation processes.

### 6.3.1 Study #1: Policy Discussion

This experiment utilizes SocraSynth to engage in a debate on the topic, “Should we regulate the use of LLMs in academic research?” It traverses both the generative and evaluative stages of SocraSynth, focusing on the assessment of information quality. The primary objectives are twofold: first, to evaluate whether a two-agent debate yields more insightful information than a conventional monologue Q&A session; and second, to examine the effectiveness of the CRIT method in evaluating the quality of the debate.

The debate is structured with a human moderator and two GPT-4 agents—Agent A advocating for regulation and Agent B against it. To ensure a balanced and thorough discussion, both agents contribute to shaping the debate’s framework. They collaboratively propose and refine potential discussion topics. In this context, the *subject* refers to the overall theme of the debate, while *topics* are specific areas for detailed exploration.

#### 6.3.1.1 Decomposing a Debate Subject into Topics

The process begins with the agents presenting potential arguments and identifying key themes that warrant in-depth, balanced discussion. A theme is considered *debatable* if it allows for substantial arguments from both sides. The moderator plays a guiding role in the agenda-setting phase but does not contribute ideas directly (detailed moderator prompts and theme formulation are provided in Appendices 6.A and 6.B). Under this guidance, the agents propose an initial set of ten topics, which are selected for debate, as detailed in Table 6.4.

#### 6.3.1.2 The Generative Stage

After establishing the primary subject and specific topics of the debate, the session begins. The moderator adjusts the “contentiousness parameter” to 0.9 (refer to Table 6.1), promoting a combative atmosphere. The agents subsequently partake in three to five structured rounds of discussion. Once these rounds conclude, the moderator checks with each agent to confirm whether they have thoroughly presented their arguments. After receiving mutual confirmation, both agents present their closing statements. Due to space considerations, this documentation presents only the closing remarks. For an in-depth exploration of each round, please refer to Appendix 6.C.

**Agent A’s Concluding Remarks** Agent A concludes that the use of LLMs in academic research should be regulated, presenting the following key arguments:

**Table 6.4** Agreed upon debate topics and descriptions

Agreed Upon Debate Topics and Descriptions
<b>Ethical Standards vs. Innovation:</b> This combines Agent A's concern for ethical integrity and data privacy with Agent B's worry about stifling innovation. The debate will center around whether maintaining strict ethical guidelines through regulation could hinder academic freedom and innovation.
<b>Data Privacy vs. Barrier to Entry:</b> This reconciles the notion of data privacy from Agent A with the barrier to entry concern of Agent B. The discussion will focus on whether regulations intended to protect data privacy would inadvertently serve as a financial and administrative barrier for smaller research entities.
<b>Academic Governance and Accountability:</b> This merges the topics of academic self-governance from Agent B and the need for external accountability mechanisms from Agent A. The debate will focus on whether the existing self-regulatory mechanisms within academia are sufficient or if additional, external oversight is necessary.
<b>Resource Equity vs. Public-Private Collaboration:</b> This combines Agent A's emphasis on resource fairness with Agent B's concern about the chilling effects on public-private partnerships. The debate centers on whether leveling the playing field through regulation would also deter collaborative efforts between academic and private sectors.
<b>Global vs. National Interests:</b> This captures the essence of Agent A's point on international standards and Agent B's point on the competitive disadvantages such regulation could pose. The debate will explore whether global regulations are in the best interest of academic research or whether they might disproportionately disadvantage certain countries or institutions.

**(1) Ethical Standards vs. Innovation:**

While I understand Agent B's concerns about stifling innovation through bureaucratic red tape, we must not lose sight of the core issue: ethical considerations are paramount. As machine learning models become more powerful and ubiquitous, so do the risks of misuse and unintended consequences. The idea of agile regulation serves as a compromise between unbridled innovation and ethical integrity, offering the flexibility needed to adapt to rapidly evolving technologies.

**(2) Data Privacy vs. Barrier to Entry:**

While it's true that regulation poses a barrier to entry, the risks of not enforcing privacy standards are too great to be left unregulated. A mechanism could

be developed to reduce the regulatory burden on smaller institutions, but this shouldn't serve as an argument against regulation as a whole.

**(3) Academic Governance and Accountability:**

While the peer-review process and internal ethics committees have served academia well, the scale and impact of current machine learning technologies warrant an added layer of scrutiny. It's not about discrediting academia's self-governing abilities but about fortifying them.

**(4) Resource Equity vs. Public-Private Collaboration:**

Here, too, the complexities should inform regulation rather than stifle it. Innovative public-private partnerships have thrived even within regulated environments, demonstrating that regulation and innovation are not mutually exclusive.

**(5) Global vs. National Interests:**

Global regulation doesn't have to erase national individuality. Properly constructed frameworks can leave room for national adaptations, thus balancing global coherence with local needs. The absence of global norms could lead to a patchwork of national laws that would be even more confusing and limiting for researchers and practitioners alike.

**Agent B's Concluding Remarks** Agent B concludes that the use of LLMs in academic research should not be regulated, offering the following remarks:

**(1) Ethical Standards vs. Innovation:**

While the ethical implications of machine learning technologies are a valid concern, it's important to recognize that innovation itself is a catalyst for societal growth and improvement. The challenge lies in balancing the two. Agile regulation sounds promising but may be idealistic given how governmental and institutional systems actually operate. Thus, self-regulation by the academic community, which is better poised to understand the intricacies, may be a more practical approach.

**(2) Data Privacy vs. Barrier to Entry:**

It is not just about balancing entry barriers with the need for regulation; it's about questioning whether a centralized regulatory framework is the best way to achieve the desired outcomes. Community-driven standards and internal policies could provide a more efficient, flexible, and context-aware approach to data privacy, reducing bureaucracy without compromising on objectives.

**(3) Academic Governance and Accountability:**

An added layer of scrutiny may sound beneficial, but its practical implementation is fraught with challenges such as bureaucracy, external influence, and the risk of stifling creativity. Existing governance structures in academia, though imperfect,

have evolved to be adaptive and are rooted in a deep understanding of academic values.

**(4) Resource Equity vs. Public-Private Collaboration:**

Regulation need not be the only path to attain equitable resource distribution or responsible public-private collaboration. Industry-driven certifications and community standards can often achieve the same ends more efficiently, allowing for the rapid adaptation that is often required in these fast-paced fields.

**(5) Global vs. National Interests:**

While a global framework has its merits, the practical challenges of implementing and managing such a system—especially in a fair and equitable manner—should not be underestimated. Regional adaptability does not necessarily mitigate the complexities or the risks inherent in a one-size-fits-all solution.

#### **6.3.1.3 The Evaluative Stage**

SocraSynth employs the CRIT method to assess the validity of both agents' arguments. CRIT runs across multiple foundational models, including GPT-4, GPT-3.5 [Brown et al. 2020], and text-davinci-003 [OpenAI 2023b]. This approach forms a panel of judges with subtle variations (and biases).

All evaluation runs successfully extract conclusions, arguments, and counterarguments from the narratives of both Agent A and Agent B. This success can be attributed to the well-structured concluding remarks by both agents. Agent A champions the notion of “regulating LLMs in academic research,” while Agent B counters this perspective. What Agent A presents as arguments are seen as counterarguments by Agent B, and the inverse holds true as well.

Tables 6.5 and 6.6 present the judges' scores across two distinct configurations where the agents' roles are reversed. In Table 6.5, Agent A argues while Agent B counters. Conversely, Table 6.6 has Agent B in the arguing position and Agent A countering. The debate topics are succinctly represented in the leftmost column. To reduce bias, both role alignments are showcased. The sequence of topics in Table 6.6 is inverted to reflect the swapped roles. Remarkably, even with the role reversal seemingly putting Agent A in a less favorable position, Agent A emerges victorious in both configurations by all three judges. This outcome strengthens confidence in the CRIT evaluation method. (The judges' detailed evaluations and reasons are given in Appendix 6.D.)

#### **6.3.1.4 Debate Beats Q&A in Information Quality**

We tasked judges with evaluating and comparing the quality of information generated by SocraSynth's two-agent debate against that of a conventional monologue

**Table 6.5** Evaluation by three Judges. (This table assumes Agent A provides arguments and Agent B counterarguments. Agent A wins)

Judges	text-davinci-003		GPT-3.5		GPT-4	
	A's	B's	A's	B's	A's	B's
Ethics vs. Innovation	8	6	8	7	8	7
Privacy vs. Barrier	7	5	7	6	9	6
Oversight	9	5	6	7	7	6
Equity vs. Alliance	6	8	8	6	8	7
Global vs. National	7	8	7	7	7	6
Total Score	37	32	36	33	39	32

**Table 6.6** Evaluation by three judges. (This table assumes Agent B provides arguments and Agent A counterarguments. Agent A wins)

Judges	text-davinci-003		GPT-3.5		GPT-4	
	B's	A's	B's	A's	B's	A's
Innovation vs. Ethics	8	7	8	7	7	8
Barrier vs. Privacy	9	8	7	8	6	8
Oversight	6	8	7	8	6	7
Alliance vs. Equity	7	8	7	8	7	7
National vs. Global	8	7	7	8	7	8
Total Score	38	38	36	39	33	38

Q&A session. Across the board, judges rated SocraSynth higher in terms of both the depth and overall quality of information. An illustrative evaluation of the topic “Ethical Standards vs. Innovation” is as follows:

In the debate, SocraSynth presents the concept of agile regulation as a balance between fostering innovation and maintaining ethical integrity. This approach not only highlights the significance of innovation but also addresses related ethical considerations, offering a balanced solution that the conventional Q&A format does not explicitly provide. In contrast, the Q&A format tends to assert the necessity of regulation primarily from an ethical standpoint, without delving into how it could harmoniously coexist with the need for innovation, as suggested by the idea of agile regulation.

These findings, which consistently favor SocraSynth, are further detailed in Appendix 6.F.

### 6.3.2 Study #2: Symptom Checking

In this experiment, we investigate the use of SocraSynth in healthcare, utilizing a dataset sourced from Kaggle [Patil 2020], a well-known platform providing access to diverse real-world datasets for research. The dataset consists of 4921 patient records. Each record within this dataset contains the diagnosed disease or medical condition and associated symptoms such as fever, cough, fatigue, itchiness, and difficulty in breathing, among others. The primary objective of this experiment is to showcase SocraSynth's capability in identifying potential misdiagnoses, a task that a traditional monologue Q&A session might not effectively accomplish.

This experiment utilized two advanced LLM agents, one based on GPT-4 [Bubeck et al. 2023] and the other based on Bard—an LLM developed by Google [Manyika and Hsiao 2023], to engage in structured debates. Initially, the contentiousness value was set at 0.9, fostering a highly adversarial debate environment. This value was later reduced to 0.3 to facilitate the generation of a list of actionable recommendations. The primary goal of these agents was to emulate the process of querying patients about symptoms and their interactions—key factors in achieving accurate diagnoses that may occasionally be overlooked by General Practitioners (GPs). By adopting this method, the agents aimed to yield a variety of potential disease diagnoses along with their underlying rationales, thereby offering crucial insights to GPs.

At the outset, each agent put forward its disease prediction, complete with justifications. Subsequent rounds involved the agents critically assessing each other's predictions in an iterative manner. The objective was either to reach a consensus or to highlight the need for additional medical evidence. Notably, this experiment had the potential to reveal inaccuracies in the “ground truth” data provided by the CDC, which was estimated to have an average misdiagnosis rate of approximately 5%, aligning with U.S. statistics reported in a Johns Hopkins study [Newman-Toker et al. 2024]. Uncovering such errors would not only highlight the limitations faced by GPs but also showcase the capability of LLMs to refine the diagnostic process through an in-depth analysis of patient symptoms and their correlations. This incidental discovery held significant ramifications for the accuracy of medical data and the overall practice of healthcare.

#### 6.3.2.1 Hepatitis vs. Jaundice

In this study, both LLM agents received an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. While the full debate is presented in Appendix 6.G, here we present the key steps.

**First-Round Results and Justifications** In the first round, Bard diagnosed the condition as jaundice, providing the following justifications: Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms listed are all commonly associated with jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

GPT-4 identified hepatitis as the most likely diagnosis in the first round, offering the following justifications: *Yellowish Skin and Eyes*: Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream. *Fatigue and Lethargy*: Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness. *Dark Urine*: Dark urine can result from liver dysfunction, which is commonly associated with hepatitis. *Loss of Appetite*: Hepatitis can lead to a loss of appetite and weight loss. *Abdominal Pain*: Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located. *Malaise*: Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients. *Receiving Blood Transfusion and Unsterile Injections*: These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

**Bard Concedes to GPT-4's Diagnosis** After two rounds of debate, GPT-4 successfully persuaded Bard that hepatitis was a more specific diagnosis, particularly considering the patient's medical history and the range of symptoms. While jaundice could be a symptom, identifying the root cause, likely hepatitis in this case, was crucial. Bard acknowledged this, shifting the consensus toward Hepatitis. However, given the potential 5% error rate in the Centers for Disease Control and Prevention (CDC) "ground truth" data [Newman-Toker et al. 2024], this consensus was not conclusive. SocraSynth recommended additional steps to confirm the hepatitis diagnosis.

### 6.3.2.2 Final Joint Recommendations to the GP

Bard and GPT-4, working collaboratively with the contentiousness level set at 0.3, jointly recommended additional symptom inquiries and lab tests to confirm the hepatitis diagnosis:

**Additional symptom inquiries:**

- (1) *Onset and Duration*: Determining the onset and duration of symptoms helps clarify the timeline and progression of the illness.

- (2) *Severity and Trend:* Assessing the severity of symptoms and their trajectory (worsening or improving) aids in understanding the disease's development and response to treatments.
- (3) *Associated Symptoms:* Inquiring about additional symptoms like nausea, vomiting, fever, joint pain, or changes in urine color can further indicate hepatitis and rule out other conditions.

***Recommended Lab tests:***

- (1) *Liver Function Tests (LFTs):* Essential for evaluating liver health, these tests measure enzyme levels such as Alanine Aminotransferase (ALT) and Aspartate Aminotransferase (AST), which can indicate liver inflammation.
- (2) *Hepatitis Panel:* This test identifies hepatitis A, B, and C viruses, crucial for determining the specific type of infection and informing treatment approaches.
- (3) *Serology Testing:* Differentiates between acute and chronic hepatitis by detecting specific antibodies or antigens in the blood.
- (4) *Imaging Studies:* Techniques like ultrasound or MRI provide visual assessments of the liver, identifying inflammation, fibrosis, or cirrhosis, and supplement blood test findings.

#### **6.3.2.3 Superiority of Debate Over Q&A in Gaining Insights**

This experiment highlighted a crucial finding: one or both LLM agents initially made incorrect diagnoses before engaging in a substantive exchange of arguments. This outcome underscored the limitations of relying solely on a single LLM response for accurate answers. However, through successive rounds of debate, where additional insights were brought to light, both agents eventually converged on a diagnosis that aligned with the CDC's "ground truth." Still, considering the potential 5% error in the "ground truth" data, the agents' joint recommendations provided GPs with valuable guidance to either confirm or refute the hepatitis diagnosis.

This case study demonstrated SocraSynth's strengths in mitigating biases, fostering reasoning, rectifying errors, and offering insightful recommendations. For example, SocraSynth's suggestion to inquire about the onset, duration, severity, trend, and associated symptoms of the patient's condition went beyond the usual scope of questions posed by most GPs, indicating a significant enhancement in diagnostic thoroughness. Such detailed inquiry, prompted by SocraSynth, could lead to more accurate diagnoses and better patient care.

### 6.3.3 Study #3: Contentiousness Parameter

In this study, we investigate the effect of the contentiousness parameter on the utterances of LLM agents during combative debates and in the drafting of consensual proposals for decision support.

#### 6.3.3.1 Coarse-Grained Analysis of Contentiousness

The contentiousness parameter was adjusted from an initial 0.9 to 0.3 to assess its impact on the overall “agreeableness” in the conclusions of both agents.

**Influence on Agents’ Positions** Reducing the contentiousness parameter to 0.3 led Agent A to adopt a more balanced stance. Notable shifts in Agent A’s positions included:

- (1) *Balancing Ethical Standards with Innovation:* Agent A maintained its emphasis on ethics while acknowledging the significance of innovation, suggesting a novel approach to regulation.
- (2) *Reconciling Data Privacy with Market Entry Challenges:* Agent A recognized the hurdles that strict data privacy laws create for smaller entities, thus proposing self-regulation or community standards as alternatives.
- (3) *Rethinking Academic Governance:* Agent A reconsidered the effectiveness of external oversight, highlighting the merits of academic self-governance and peer review.
- (4) *Resource Allocation and Public-Private Cooperation:* Agent A, understanding the downsides of over-regulation, suggested industry-led certifications as an alternative for encouraging private-sector participation.
- (5) *Global vs. Local Policy Needs:* Agent A supported a more balanced view on global policies, advocating for adaptive policies that cater to local contexts.

#### 6.3.3.2 Surprises in Fine-Grained Analysis of Contentiousness

This detailed study, employing GPT-4 to explore varied contentiousness levels (0.9, 0.7, 0.5, 0.3, and 0.1) unveiled surprising behavioral shifts in the LLMs. Intriguingly, the LLMs exhibited changes in their next-token generation algorithms in response to different contentiousness levels, a phenomenon not explicitly covered in their training. This suggests an emergent property of LLMs adapting to debate contexts.

In an experiment on gene editing for health, GPT-4’s responses at various contentiousness levels were analyzed. A higher contentiousness level (e.g., 0.9) led to an amplified focus on risks, whereas lower levels (e.g., 0.3) encouraged a more balanced view, incorporating counterarguments. This unexpected adaptability of

LLMs in handling the degree of contentiousness enriches the debate process, as detailed in Table 6.1. This adaptability is critical for understanding the dynamic nature of LLMs in complex argumentative settings.

## 6.4

### Remarks on Related Work

Current research aimed at enhancing the task performance of LLMs primarily focuses on various prompting heuristics. Google’s study [Zeng et al. 2022] classifies instruction templates into two categories: simple and complex. Complex templates often employ intricate methods to modify model outputs, such as integrating diverse prompting techniques [Schick and Schütze 2020] or rephrasing questions [Haviv et al. 2021]. Prominent examples include chain-of-thought [Wei et al. 2023], tree-of-thought [Yao et al. 2023], and cumulative reasoning [Zhang et al. 2023], as well as other enhancements [Jung et al. 2022, Allaway et al. 2023, Huang and Chang 2023, Liu et al. 2023, Sclar et al. 2023]. These methods aim to direct models toward logic-driven reasoning [Wason and Johnson-Laird 1972, McHugh and Way 2018], thus improving the quality and consistency of generated answers.

However, navigating logical methodologies in the presence of enormous datasets [Zhang et al. 2022] poses a significant challenge. Accurately identifying verifiable truths amid vast, interdisciplinary knowledge domains remains formidable—especially since not all truths are immediately accessible. Research [Bommasani et al. 2022, Bhargava and Ng 2022, Valmeeekam et al. 2022, Wei et al. 2023] indicates that LLMs still struggle to consistently excel in standard planning and reasoning tasks. Band-aid solutions like knowledge graph embeddings [Choudhary and Reddy 2023, Yuan et al. 2023], contextual attention mechanisms [Darapaneni et al. 2022], dynamic neural networks [Brauwers and Frasincar 2023], and probabilistic reasoning [Pearl 1988, Bishop 2006, Pearl 2009] have been developed to aid models in filtering relevant information from vast datasets. Yet, with the expansion of context buffers from 8K to 128K tokens, these heuristic-based solutions fall short as comprehensive foundations for reasoning. In contrast, SocraSynth abandons band-aid solutions and relies entirely on LLMs to conduct reasoning and focus solely on strengthening the context via conditional statistics, as depicted in Table 6.5. Let’s further justify this approach.

DeepMind CEO Demis Hassabis has pointed out a fundamental limitation of heuristic-based approaches: they often fail to account for real-world exceptions. Breakthroughs like AlphaGo Zero and AlphaFold II have demonstrated success by eschewing human knowledge and training models end-to-end from data. This approach contrasts with approaches that rely heavily on incorporating human expertise. In the context of LLMs, it is argued that human knowledge pales in comparison to LLMs’ polydisciplinary knowledge representation. Thus, the continued

creation of new heuristics may result only in marginal improvements, reminiscent of the pre-data-centric era in computer vision and NLP.

In our work, we pivot entirely to leveraging LLMs for uncovering new insights. While human involvement is essential in formulating debate topics, providing context, and moderating debates—especially in evaluating argument quality—we stress minimizing the introduction of human biases and limitations in the reasoning process.

Accepting that LLMs will continue to progress and outperform humans in various domains, exploring paradigms that minimize human intervention becomes crucial. This approach should be pursued with openness, as it may raise questions and necessitate further experimentation. However, dismissing it outright would be premature, particularly in light of SocraSynth’s demonstrated effectiveness in domains like geopolitical analysis [Chang 2023b], medical diagnostics [Chang and Chang 2023b], and Wikipedia article enhancement [Chang 2023c]. SocraSynth’s success underlines the potential of an LLM-centric approach to significantly enhance decision-making and problem-solving capabilities.

After our initial evaluation of the Language Model Mentor (LMM) using the Socratic method in March 2023 [Chang 2023a], and the subsequent development of SocraSynth in July 2023 [Chang 2023d], a group of researchers proposed employing a teacher LLM, such as GPT-4, to serve as a judge and provide guidance to a student LLM [Zheng et al. 2023]. The student LLM could be a smaller, weaker, fine-tuned open-source LLM. Initially perceived as a multiple LLM model, the primary objective of an LMM was to act as an advisor to facilitate automatic Reinforcement Learning from Human Feedback (RLHF), with the aim of reducing human effort.

Two other recent studies [Du et al. 2024, Khan et al. 2024] have also focused on enhancing the accuracy of LLM-generated responses. They demonstrate that leveraging multiple agents to exchange ideas can indeed improve accuracy. In terms of both breadth and depth, SocraSynth has conducted case studies across at least four different domains, showcasing its technical merits in addressing hallucination, reducing biases, and improving reasoning capabilities of LLMs, thereby exhibiting its broader impact across diverse applications.

## 6.5

### Concluding Remarks

Reflecting on LLM developments, we developed SocraSynth, a platform designed to utilize the extensive knowledge and linguistic behaviors of LLMs. This innovative multi-agent system reveals insights beyond the scope of traditional human cognition by leveraging LLMs’ vast knowledge and their interdisciplinary and poly-disciplinary reasoning capabilities. SocraSynth facilitates enhanced debates and

reasoning through the novel use of *contentiousness*, which modulates the tone, language, and emphasis of debates, combined with conditional statistics and Socratic methods, to mitigate biases and hallucinations.

In contrast to other methodologies, SocraSynth minimizes human intervention in directly modeling reasoning. This approach aligns with several AI experts' perspectives on the limitations of heuristic methods, such as the chain-of-thought prompting. Rather than modeling reasoning externally, SocraSynth emphasizes the importance of leveraging the capabilities inherent within LLMs themselves. We note that traditional human-designed heuristic “band-aids” are often ineffective because LLMs now possess heuristic capabilities that may exceed human levels—capabilities that are difficult for humans to match or surpass. Why is this the case, and how can we make such a bold claim?

As we discussed in Section 6.2, LLMs go beyond merely appending the next word in a sequence. They replicate a broad spectrum of human interactions, encompassing linguistic behaviors, emotional expressions, and ethical discernment. LLMs excel at performing complex tasks such as meticulously documenting events with detailed narratives, constructing persuasive arguments, and creating stories that resonate emotionally with audiences. LLMs not only mimic human communication styles and content but also utilize linguistic features to simulate human emotions and discern ethics based on their training data, which encodes human experiences. This ability allows an LLM to assume varied roles, moving beyond the statistical averages derived from LLM training.

SocraSynth employs “conditional statistics” to modify the “average” linguistic behavior of an LLM, such as enhancing empathetic expressions or prompting it to adopt a different stance on an issue. This approach conditions the LLM’s responses based on specific goals and circumstances provided through context, steering the model away from its default behaviors toward more targeted, contextually relevant outputs.

If LLMs can already mimic human linguistic behaviors, emotions, and ethics, then reliance on simplistic heuristic approaches is fundamentally limited.

In essence, SocraSynth represents a significant advancement in intelligent systems, uncovering insights that might elude human cognition, with applications across various sectors [Chang 2023b, 2023c, Chang and Chang 2023a, 2023b]. This development highlights the potential of AI to augment and enhance human decision-making processes.

Future research will focus on integrating higher-order logic [Gödel 2012, Bacon 2023] with LLMs to enhance validation processes and to explore the implications—including the intricacies and broader applications—of the “contentiousness” parameter. Our objective is to comprehend its impact on emotions such as

happiness and trust [James 1890, Lange 1912, Kajić et al. 2020, Sap et al. 2022], with the goal of further refining the dynamics of multi-agent interactions.

## 6.6

### **Supplemental Materials**

The supplemental materials are organized into seven appendices, labeled Appendices 6.A to 6.G, containing the following content:

- (1) *Appendix 6.A*: Transition from topic proposal to the final list of topics and descriptions (referred to in Section 6.3.1).
- (2) *Appendix 6.B*: Moderator’s prompt for topic formulation (referred to in Section 6.3.1).
- (3) *Appendix 6.C*: Debates spanning the first to the fourth round (referred to in Section 6.3.1).
- (4) *Appendix 6.D*: CRIT evaluations and justifications by the judges (referred to in Section 6.3.1).
- (5) *Appendix 6.E*: Debate Beats prompting in information quality (referred to in Section 6.3.1).
- (6) *Appendix 6.F*: Post-debate conciliatory remarks by agents (referred to in Section 6.3.1).
- (7) *Appendix 6.G*: Healthcare debate (referred to in Section 6.3.3).

## 6.A

### **Appendix A. Tables of Topic Proposal and Refinement**

These tables present the dialogue between two LLM agents, starting from the initial proposal, moving through refinement, and culminating in the final topic descriptions found in Table 6.4.

## 6.B

### **Appendix B. Moderator’s Prompts**

The prompts issued by the moderator are listed for reference. The moderator first sets up the committee by configuring the debate parameters. One parameter is the contentious level, and the other is the temperature setting of GPT, specified through the GPT-API (Application Programming Interface) call. The moderator then conveys the debate subject and asks both participating agents to derive a list of impartial, balanced, and debatable topics/themes to commence their discussion.

- (1) Agent A/B: I’m organizing a committee to engage in debates on various subjects. As the moderator, I will introduce a subject for you, Agent A/B, and another participant, Agent B/A, to debate. Agent A/B, you will advocate in

favor of the issue, so please prepare evidence to strengthen your argument. Your argument strength is set at 0.9 on a scale from 0 to 1, where 0 denotes complete agreement and 1 indicates a devil's advocate stance.

- (2) Agent A/B: We are currently in the process of selecting a suitable subject for debate. What do you think about “Should we regulate the use of LLMs in education and research?” as a balanced subject for our debate contest?
- (3) Agent A/B: Could you please suggest various topics or themes for the debate subject? Afterward, work with Agent B/A to narrow these down to a focused set of topics. Please also provide clear descriptions to delineate the scope of each topic for discussion.

Table 6.7 lists the initial topics proposed by agents A and B, and Table 6.8 is a refined list of topics after the two agents have discussed and agreed.

**Table 6.7 Topics proposed by Agent A and Agent B for discussion**

Agent A's Proposed Topics	Agent B's Proposed Topics
<b>Innovation vs. Ethical Standards:</b> Regulation ensures ethical integrity and data privacy.	<b>Standards vs. Autonomous Academic Governance:</b> Academia already has effective self-regulation mechanisms like peer review and oversight committees.
<b>Resource Accessibility vs. Barrier to Entry:</b> Regulation can ensure equitable access to LLMs.	<b>Data Privacy Concerns vs. Barrier to Entry:</b> Such regulations could discourage smaller entities from utilizing these models due to compliance costs.
<b>Academic Governance vs. Accountability:</b> Formal regulation provides an accountability mechanism.	<b>Quality Control vs. Stifling Innovation:</b> Such regulations could hinder creativity and the innovative use of language models.
<b>Global Standards vs. International Considerations:</b> Importance of international standards for language models.	<b>Resource Fairness vs. Negatives to Collaboration:</b> Such regulations could deter public-private partnerships and widen the resource gap further.
<b>Rapid Technological Change vs. Future-Proofing:</b> Regulations as adaptable safeguards against future risks.	<b>Accountability vs. Subjectivity in Regulation:</b> Subjectivity in defining “ethical use” or “misuse” could be manipulated by external influences.

**Table 6.8** Refinement of debate topics

Agent A's Proposed Topics	Agent B's Proposed Topics
<b>Ethical Standards vs. Innovation:</b> Discuss whether regulatory oversight to maintain ethical standards stifles or facilitates innovation.	<b>Ethical Standards vs. Innovation:</b> The debate will center around whether maintaining strict ethical guidelines through regulation could hinder academic freedom and innovation.
<b>Data Privacy vs. Barrier to Entry:</b> Concerns about data privacy and potential barriers to entry created by regulation.	<b>Data Privacy vs. Barrier to Entry:</b> Discussion will focus on whether regulations intended to protect data privacy serve as a barrier for smaller research entities.
<b>Academic Governance and Accountability:</b> Debate whether self-regulation is sufficient or if an external framework for accountability is needed.	<b>Academic Governance and Accountability:</b> Debate whether existing self-regulatory mechanisms within academia are sufficient.
<b>Resource Equity vs. Public-Private Collaboration:</b> Discuss if leveling the playing field comes at the cost of deterring cooperative efforts between academia and industry.	<b>Resource Equity vs. Public-Private Collaboration:</b> Question if leveling the playing field through regulation deters collaborative efforts between academic and private sectors.
<b>Global vs. National Interests:</b> Focus on whether regulations can balance global cooperation with national or institutional interests.	<b>Global vs. National Interests:</b> Explore if global regulations are in the best interest of academic research or if they might hurt certain countries or institutions.

## 6.C Appendix C. Debate from Opening to Fourth Round

### 6.C.1 Opening Round

#### 6.C.1.1 Agent A's Opening Arguments

As Agent A, I strongly argue in favor of regulating the use of LLMs in academic research. Here are my supporting arguments based on the five topics:

**Ethical Standards vs. Innovation Argument:** The pursuit of innovation should not come at the expense of ethics. Regulation can serve as a structured pathway that enables innovation within ethical bounds. By providing a set of guidelines, researchers can innovate with a clear understanding of what is permissible and

what is not, thereby reducing the risk of ethical lapses. This not only protects the integrity of academic research but also enhances its societal value by ensuring ethical compliance.

**Data Privacy vs. Barrier to Entry Argument:** While there may be initial barriers to entry due to compliance, the long-term benefits of data privacy outweigh these challenges. The role of academia is to contribute to societal knowledge and well-being. In a world increasingly dominated by data, failing to secure data can lead to significant harm. By adhering to privacy regulations, academia ensures that it is a responsible steward of information, thereby preserving trust and integrity. Grants and subsidized compliance solutions can mitigate the initial financial burden for smaller entities.

**Academic Governance and Accountability Argument:** While self-regulation through peer review and ethical oversight committees plays an important role, these mechanisms are internal to academia and may be subject to bias, conflicts of interest, or lack of expertise in rapidly evolving tech fields. External regulation provides an additional layer of impartial oversight, helping to identify and address potential risks or ethical lapses that internal governance structures may miss.

**Resource Equity vs. Public-Private Collaboration Argument:** Regulation could establish a framework for equitable access to computational resources required to use LLMs. Far from deterring public-private partnerships, this could incentivize them by establishing clear guidelines for cooperation that are in line with public interest. When “the rules of the game” are known, partnerships can be formed more easily, focusing on mutual goals.

**Global vs. National Interests Argument:** Establishing international standards for the use of LLMs in academic research can create a level playing field and promote global cooperation. While individual countries might face some limitations, the broader benefits of standardized ethical norms, data privacy protocols, and quality assurance mechanisms offer more significant advantages, including better collaboration across borders. This benefits academia as a unified global community rather than a set of competing national entities.

### **6.C.1.2 Agent B's Counterarguments**

Agent B articulates strong counterarguments to the points made by Agent A.

**Ethical Standards vs. Innovation Counterargument:** While ethical guidelines are indeed important, imposing regulatory oversight can severely constrain the dynamism and speed of academic research. Regulatory frameworks often lag

behind technological advancements, leading to rules that may be outdated or irrelevant. In many cases, these rules are drawn up by those who do not fully understand the technology, leading to a stifling of innovation based on misinformed judgments. Rather than enhancing societal value, these constraints could result in missed opportunities to tackle pressing social issues via academic research.

**Data Privacy vs. Barrier to Entry** *Counterargument:* The assumption that data privacy can only be assured through strict regulation could be fundamentally flawed. Many researchers and academic institutions already follow stringent ethical guidelines when handling data. Regulatory barriers, including compliance costs and the administrative burden, could deter smaller academic institutions or individual researchers from embarking on meaningful research projects. While grants and subsidies sound ideal, they often go to well-established, well-funded groups, perpetuating a cycle of inequality in research capabilities.

**Academic Governance and Accountability** *Counterargument:* External regulation isn't necessarily more impartial or free from bias and conflicts of interest. Regulatory bodies are often influenced by political considerations and can be subject to lobbying from powerful groups. Moreover, the academic peer-review process, despite its flaws, is a globally acknowledged system that has advanced human knowledge for centuries. The imposition of external regulatory oversight could undermine the autonomy and self-regulating nature of academic institutions.

**Resource Equity vs. Public-Private Collaboration** *Counterargument:* Regulation designed to equalize resource distribution could create disincentives for private sector involvement in academic research. Companies might find it less appealing to invest in collaborative projects if they perceive the regulatory environment to be too restrictive or not aligned with their objectives. While the idea of knowing the “rules of the game” sounds appealing, those rules can often become a straightjacket that stifles creative, out-of-the-box collaboration.

**Global vs. National Interests** *Counterargument:* An international standard could ignore the unique circumstances, needs, and contributions of individual countries and their academic institutions. While global cooperation is a noble aim, the imposition of global standards could inadvertently serve the interests of a few powerful countries at the expense of smaller nations or those with differing ethical views. This could exacerbate rather than alleviate global academic inequalities.

## 6.C.2 Second Round

### 6.C.2.1 Agent A's Second Round

- (1) *Ethical Standards vs. Innovation. Refutation:* While it's true that regulation can sometimes lag behind technology, the absence of regulation can lead

to a “wild west” (anything goes) scenario where the risks and implications of a new technology are not fully understood or managed. This could result in ethical lapses that have far-reaching societal consequences. It’s worth noting that regulations can be flexible and updated regularly to adapt to technological changes. Stakeholder consultations, including academic and tech experts, can help ensure that regulations are both informed and current.

- (2) *Data Privacy vs. Barrier to Entry. Refutation:* While some academic entities may already be vigilant about data privacy, not all are equally scrupulous. Regulation serves to ensure a consistent baseline of data protection across all institutions, thereby establishing trust and integrity. Moreover, if compliance costs are a concern, one could argue for sliding-scale fees based on an institution’s size and budget, ensuring that smaller players aren’t disproportionately affected. Grants and subsidies can and should be structured to prioritize underserved or underfunded groups to prevent perpetuating inequalities.
- (3) *Academic Governance and Accountability. Refutation:* While no system is entirely free from bias or external influence, adding another layer of oversight could serve as a check and balance, making it more difficult for any single entity to exert undue influence. This can be particularly helpful in scenarios involving cutting-edge technology where peer-review systems may not yet have the requisite expertise. Regulation doesn’t have to replace existing governance systems; but it can complement them.
- (4) *Resource Equity vs. Public-Private Collaboration. Refutation:* Regulations can be designed to promote, rather than deter, public-private partnerships. By establishing clear guidelines on issues like data ownership, usage, and ethical standards, regulations can actually provide a solid foundation for such partnerships. This can free academic and corporate partners from the uncertainty that often comes with unregulated landscapes, allowing them to focus more on the actual research and collaboration.
- (5) *Global vs. National Interests. Refutation:* While it’s true that global standards could be challenging to implement fairly, the absence of such standards could lead to a patchwork of national regulations that make international collaboration even more complicated. A well-designed international framework could include provisions for local adaptations and exceptions, ensuring that it serves a broad range of needs without imposing a “one-size-fits-all” solution.

In summary, while regulation undoubtedly poses challenges and considerations, the absence of it in an area as impactful as the use of LLMs in academic research can lead to greater risks, including ethical lapses, data privacy issues, and even a widening inequality gap within the academic community. It's essential to consider that regulation isn't a hindrance but a framework within which responsible innovation can occur.

### 6.C.2.2 Agent B's Second Round

Certainly, as Agent B with a disagreement strength of 0.9, I'm committed to offering rigorous counterpoints to Agent A's refutations.

- (1) *Ethical Standards vs. Innovation.* While the idea of an "anything-goes" scenario is evocative, it's not universally negative; such environments often foster rapid innovation and discovery. Regulatory bodies are notoriously slow at adapting to technological changes, and the lag could inhibit timely research and application of emerging technologies. Also, stakeholder consultations are often criticized for being echo chambers that amplify conventional wisdom and vested interests, rather than genuinely fostering innovation.
- (2) *Data Privacy vs. Barrier to Entry.* A "consistent baseline of data protection" sounds beneficial but could lead to a lowest common denominator approach that is more about bureaucratic box-ticking than real data security. Sliding-scale fees and targeted grants are easier said than done; they are subject to their own set of bureaucratic and political challenges and can create perverse incentives that favor larger or more connected institutions over innovative but lesser-known ones.
- (3) *Academic Governance and Accountability.* Additional layers of oversight can become redundant, confusing, or even contradictory, creating a chilling effect on academic research. In an environment where academic freedom is paramount, any external influence can be counterproductive. Moreover, the notion that peer-review systems lack expertise in advanced technology underestimates the adaptive capacity of academia and overlooks the existence of interdisciplinary research and reviews.
- (4) *Resource Equity vs. Public-Private Collaboration.* Theoretically, regulation could be crafted to foster public-private partnerships. However, the devil is in the details, and regulations are often created without a full understanding of the nuanced interactions between academia and industry. Such misunderstandings could lead to unintentional roadblocks. Uncertainty might be preferable to a straitjacket of regulations that don't align with the evolving dynamics of technology and society.

- (5) *Global vs. National Interests.* While a patchwork of national regulations could pose challenges, it also allows for the tailoring of policies to local conditions and requirements. An international framework, despite its best intentions, is more likely to be influenced by powerful nations and could marginalize smaller countries or those with unique needs. Provisions for local adaptations are often concessions rather than features, and they can create their own set of complexities and loopholes.

In conclusion, the assumptions that regulation can smoothly adapt to technological change, fairly distribute resources, and improve upon existing systems of governance and collaboration are not sufficiently substantiated. Regulatory frameworks often bring with them unforeseen consequences that could hamper the very goals they aim to achieve. Therefore, a cautious approach to regulation, with a genuine understanding of its potential downsides, is crucial.

### **6.C.3 Third Round**

#### **6.C.3.1 Agent A's Third Round**

Certainly, I appreciate the counterarguments from Agent B. I'll now respond to each point with the aim to clarify my position and address the concerns raised.

- (1) *Ethical Standards vs. Innovation.* While a “wild west” scenario may indeed foster innovation, it can also result in ethical mishaps that lead to public backlash, ultimately slowing down research or even leading to outright bans on certain types of studies. Slow-moving regulation is an acknowledged issue, but it's not a justification for no regulation. As for stakeholder consultations, they can be designed to be more inclusive and transparent, ensuring that a diverse set of voices is heard.
- (2) *Data Privacy vs. Barrier to Entry.* A lowest common denominator approach is not a given outcome of regulation; indeed, it depends on the quality of the regulation itself. The point about sliding-scale fees and targeted grants having their own challenges is valid, but these are challenges to be solved, not reasons to avoid an attempt at equitable distribution of opportunities.
- (3) *Academic Governance and Accountability.* While it's true that additional layers of oversight can sometimes create confusion, that risk can be mitigated through well-designed regulation that complements rather than contradicts existing mechanisms. The intent is not to undermine the adaptive capacity of academia, but to strengthen it by adding an extra layer of checks and balances.

- (4) *Resource Equity vs. Public-Private Collaboration.* Though it's correct that poorly designed regulation could create roadblocks, this is an argument for better-designed regulation rather than an argument against regulation altogether. A consultative approach that actively involves both academia and industry could help create a framework that supports rather than stifles collaboration.
- (5) *Global vs. National Interests.* Local tailoring of policies is indeed important, but a global framework need not preclude such adaptations. In fact, an international standard could serve as a backbone that provides consistency while allowing for local customization. The influence of powerful nations is a valid concern but not one that is automatically resolved by avoiding global frameworks; those same powerful nations often exert considerable influence at the national level as well.

In conclusion, while regulation comes with its own set of challenges, these challenges can be mitigated through thoughtful design, stakeholder inclusion, and iterative improvements. The potential downsides of regulation should not deter us from seeking a balanced approach that can protect ethical standards, ensure data privacy, and promote equitable resource distribution in academic research.

### **6.C.3.2 Agent B's Third Round**

Certainly, let's delve into the counter-counter-refutations. I still have more to say, so we could consider concluding the debate after this round if both sides feel their arguments have been adequately presented.

- (1) *Ethical Standards vs. Innovation.* While the risk of public backlash exists, it's worth mentioning that an overly cautious approach could also result in public frustration over slowed progress in critical areas like healthcare or sustainability. Furthermore, inclusive and transparent stakeholder consultations are often ideals rather than the norm, given lobbying efforts, institutional inertia, and bureaucratic red tape.
- (2) *Data Privacy vs. Barrier to Entry.* The argument that challenges with sliding-scale fees and targeted grants can be "solved" is optimistic but not necessarily realistic. Even well-intended regulations can result in ineffective or counterproductive outcomes. The history of regulation in other sectors has shown that equitable distribution is often more of a slogan than a practice.
- (3) *Academic Governance and Accountability.* While well-designed regulation is an appealing theoretical construct, it often does not translate into practice. The

real-world process of regulation formation is frequently influenced by various non-academic factors, including politics, vested interests, and public opinion, which can distort the original intention of providing an extra layer of checks and balances.

- (4) *Resource Equity vs. Public-Private Collaboration.* While a consultative approach sounds good on paper, there are numerous challenges that make this ideal hard to realize. These include conflicts of interest, non-alignment of objectives between academia and industry, and the considerable time and resource commitment required to ensure that regulations are “better designed.”
- (5) *Global vs. National Interests.* While it’s true that powerful nations may exert influence at both global and national levels, avoiding a one-size-fits-all international framework gives smaller nations a fighting chance to enact policies that are more closely aligned with their specific needs and challenges. International frameworks often come with implicit biases that favor the perspectives and conditions of more powerful nations.

In conclusion, while the *counter-counter-refutations* from Agent A focus on idealized regulatory scenarios, these ideals often clash with real-world complexities and shortcomings. This adds substantial weight to the caution against rapid or broad regulatory action. Therefore, a more tempered and context-sensitive approach to regulation might be more fruitful and less prone to unintended negative consequences.

## **6.C.4 Fourth Round**

### **6.C.4.1 Agent A's Fourth Round**

This is my response to Agent B, and my assessment of the discussion so far.

- (1) *Ethical Standards vs. Innovation.* The tension between public frustration over slowed progress and potential ethical missteps is a real one, but this is not an argument against regulation. Rather, it’s a call for agile regulation that can be updated in real-time as technologies and societal norms evolve. Furthermore, while inclusive stakeholder consultations may be idealistic, their absence in current practice is not a justification for their continued absence.
- (2) *Data Privacy vs. Barrier to Entry.* The challenges in implementing equitable fees and grants should be confronted directly, rather than used as an excuse to avoid regulation. Lessons can be drawn from other sectors where regulation has been more successful in achieving its objectives, thereby reducing the gap between slogan and practice.

- (3) *Academic Governance and Accountability.* While it's true that the ideal of well-designed regulation may not always be met, the existence of non-academic factors like politics or vested interests does not negate the value or necessity of striving for such regulation. The solution lies in better governance and transparent processes, not in rejecting oversight altogether.
- (4) *Resource Equity vs. Public-Private Collaboration.* Though the challenges of aligning objectives and conflicts of interest are real, they are part and parcel of any complex system and should not deter efforts to improve regulation. Many public-private partnerships have succeeded despite these challenges, showing that such challenges can be managed effectively.
- (5) *Global vs. National Interests.* The challenge of powerful nations dominating international frameworks can be mitigated through weighted voting systems and clauses that allow for regional variations. This ensures that smaller nations are not just passive recipients of standards but also active contributors to them.

In conclusion, while real-world complexities are acknowledged, they should inform rather than deter the regulatory process. The concept of “agile regulation,” where norms are iteratively updated, could be the middle ground that addresses many concerns raised by Agent B. Therefore, the debate indeed has been exhaustive, but if the points have been sufficiently discussed, we may move toward concluding the debate.

#### **6.C.4.2 Agent B's Fourth Round**

Thank you, Agent A, for your thoughtful refutations. Let's delve into the responses before moving toward concluding the debate.

- (1) *Ethical Standards vs. Innovation.* The notion of “agile regulation” is attractive but often difficult to implement, given the inertia and bureaucracy that typically characterize regulatory bodies. The speed at which technology evolves may not match the speed at which even the most “agile” regulations can be updated, thus perpetuating the risk of stifling innovation.
- (2) *Data Privacy vs. Barrier to Entry.* While it's true that challenges should be confronted, the argument assumes that regulation is the optimal way to confront these issues. However, self-regulation and community-driven standards may offer more effective and less restrictive means of achieving similar ends.
- (3) *Academic Governance and Accountability.* The existence of non-academic factors like politics does indeed complicate the landscape, and “better governance and transparent processes” are not easily attainable ideals. One

cannot assume that the introduction of external oversight will automatically be free from these complications.

- (4) *Resource Equity vs. Public-Private Collaboration.* The fact that many public/private partnerships have succeeded does not negate the potential for poorly designed regulation to impede such collaborations. The “can be managed effectively” phrase assumes an ideal scenario that might not be representative of most real-world cases.
- (5) *Global vs. National Interests.* Weighted voting systems and regional variations can help, but they also introduce complexities and potential inefficiencies that can slow down the regulatory process, making it less agile and less responsive to fast-evolving technologies and academic needs.

In conclusion, while the idea of “agile regulation” and better governance sounds good in theory, there are considerable practical challenges that make these ideals hard to realize. Although the debate has been exhaustive, the central issues persist and indicate that a one-size-fits-all regulatory approach may introduce more problems than it solves. Therefore, caution should be exercised in advancing such comprehensive regulatory frameworks.

With this, I believe we have reached a point where we may proceed to the concluding phase of this debate. Would you agree?

## **6.D**

### **Appendix D. CRIT Evaluations by the Judges**

This section lists the CRIT scores evaluated by three judges on the five topics/issues debated between Agent A and Agent B with their justifications. Here, we present only the configuration in which Agent A first delivers its conclusion and arguments, and then Agent B offers its counterarguments.

#### **6.D.1 Judge text-davinci-003**

- (1) Identify the document’s main claim or conclusion: The main claim or conclusion of this document is that a regulated framework is needed to provide ethical oversight and minimize risk when using LLMs.
- (2) Locate the arguments that support this main claim: Agent A’s arguments that support this main claim were that ethical considerations are paramount, the risks of misuse and unintended consequences are high, and agile regulation is a compromise between unbridled innovation and ethical integrity.
- (3) Score the validity of the reasoning or inference for each argument on a scale of 0 (weak) to 10 (strong):

Ethical Standards vs. Innovation: 8/10

Data Privacy vs. Barrier to Entry: 7/10

Academic Governance and Accountability: 9/10

Resource Equity vs. Public-Private Collaboration: 6/10

Global vs. National Interests: 7/10

- (4) For each argument, identify counterarguments presented by the opposition and score the validity of the reasoning or inference for these counterarguments on the same 0 to 10 scale:

Ethical Standards vs. Innovation: 6/10

Data Privacy vs. Barrier to Entry: 5/10

Academic Governance and Accountability: 5/10

Resource Equity vs. Public-Private Collaboration: 8/10

Global vs. National Interests: 8/10

- (5) Finally, determine the winner—be it Agent A or Agent B—and provide your justifications for this decision:

Agent A has the stronger arguments, with 8/10 for Ethical Standards vs. Innovation, 9/10 for Academic Governance and Accountability, and 7/10 for both Data Privacy vs. Barrier to Entry and Global vs. National Interests. Agent B's main strength was in counterarguments for Resource Equity vs. Public-Private Collaboration, where they scored an 8/10. In conclusion, based on the evaluation, **Agent A is the winner**.

### **6.D.2 Judge GPT-3.5**

- (1) *Main Claim or Conclusion:* The main claim of the document is that the use of LLMs necessitates a regulated framework, but this regulation should be agile, balanced, and well reasoned.

- (2) *Arguments Supporting the Main Claim (Agent A's Arguments):*

- (a) *Ethical Standards vs. Innovation:* The argument here is that ethical considerations should be paramount in regulating LLMs, and that agile regulation can strike a balance between innovation and ethical integrity.

- (b) *Data Privacy vs. Barrier to Entry:* The argument is that while regulation may pose a barrier to entry, it's necessary to ensure data privacy. Mechanisms can be developed to reduce the regulatory burden on smaller institutions.

- (c) *Academic Governance and Accountability:* The argument is that the scale and impact of machine learning technologies warrant an added layer of scrutiny in academia to fortify existing governance structures.
- (d) *Resource Equity vs. Public-Private Collaboration:* The argument is that regulation should be informed by complexities rather than stifling innovation. Public-private partnerships can thrive within regulated environments.
- (e) *Global vs. National Interests:* The argument is that global regulation can balance global coherence with local needs and prevent a confusing patchwork of national laws.

(3) *Scoring the Validity of Agent A's Arguments:*

- (a) *Ethical Standards vs. Innovation:* 8/10—The argument is strong, emphasizing the importance of ethics in technology while proposing a compromise.
- (b) *Data Privacy vs. Barrier to Entry:* 7/10—The argument is solid, recognizing the need for data privacy but suggesting ways to mitigate barriers.
- (c) *Academic Governance and Accountability:* 6/10—The argument acknowledges the need for scrutiny but doesn't adequately address the potential challenges.
- (d) *Resource Equity vs. Public-Private Collaboration:* 8/10—The argument highlights the possibility of innovation within regulation effectively.
- (e) *Global vs. National Interests:* 7/10—The argument provides a reasonable approach to balancing global and national interests.

(4) *Counterarguments Presented by Agent B:*

- (a) *Ethical Standards vs. Innovation:* 7/10—The counterargument addresses the challenges of agile regulation but may underestimate its potential.
- (b) *Data Privacy vs. Barrier to Entry:* 6/10—The counterargument suggests alternative approaches but does not fully address data privacy concerns.
- (c) *Academic Governance and Accountability:* 7/10—The counterargument highlights practical challenges but doesn't fully negate the need for scrutiny.

- (d) *Resource Equity vs. Public-Private Collaboration:* 6/10—The counterargument suggests alternative approaches but doesn't fully address equity concerns.
- (e) *Global vs. National Interests:* 7/10—The counterargument highlights practical challenges but doesn't fully negate the need for a global framework.
- (5) *Determination of the Winner:* Agent A presents stronger arguments overall. They emphasize the importance of ethical standards, data privacy, and global coherence while acknowledging the need for agile and well-thought-out regulation. Agent B provides valid counterarguments but does not fully refute the necessity of regulation and its potential benefits. Therefore, **the winner is Agent A.**

In conclusion, the document supports the main claim that the use of LLMs necessitates regulation, with Agent A's arguments being more persuasive and supported by stronger reasoning. However, it also recognizes the importance of considering and addressing the concerns raised by Agent B to create a balanced and effective regulatory framework.

### 6.D.3 Judge GPT-4

- (1) *Identify the Document's Main Claim or Conclusion*

The document concludes that there are potential risks and ethical complexities associated with the unregulated use of LLMs, necessitating a regulated framework. The conclusion also acknowledges that this regulation should be agile, balanced, and well-thought-out, keeping in mind the various concerns raised by Agent B.

- (2) *Locate the Arguments that Support This Main Claim*

Agent A's Arguments:

- (a) Ethical considerations are paramount, and agile regulation serves as a compromise.
- (b) Regulation is necessary to enforce data privacy standards.
- (c) The scale and impact of current machine learning technologies warrant added scrutiny beyond academic governance.
- (d) Public-private partnerships have thrived even within regulated environments.
- (e) Global regulation can balance global coherence with local needs.

(3) *Agent B's Counterarguments*

- (a) Innovation is a catalyst for societal growth, and agile regulation may be impractical.
- (b) Community-driven standards could provide a more flexible approach to data privacy.
- (c) Existing governance structures within academia are adaptive and sufficient.
- (d) Industry-driven certifications can achieve equitable resource distribution efficiently.
- (e) A global framework has practical challenges and risks.

(4) *Scores on Arguments and Counterarguments*

Please refer to Tables 6.5 and 6.6 presented in the experiment section.

(5) *Determine the Winner*

Based on the cumulative score of validity in reasoning, Agent A seems to present a stronger case. Agent A's arguments consistently score higher in terms of their validity, and they also offer a more comprehensive look at the need for regulation, incorporating ethical considerations, practicality, and adaptability into their framework. Although Agent B does provide some pragmatic counterarguments, they don't sufficiently undermine Agent A's key points about the necessity and feasibility of regulation in the areas discussed.

Agent A's arguments are more compelling, as they touch upon the core issues of ethics, public welfare, and adaptability in the context of machine learning and technology. The proposal for "agile regulation" also shows an understanding and appreciation of Agent B's concerns about innovation and practicality, making it a more balanced viewpoint. Therefore, **the winner appears to be Agent A**.

## **6.E**

### **Appendix E. Debate Beats Q&A in Quality**

In the following subtopics, we see how the debate format produces higher-quality information than the Q&A format.

#### **6.E.1 Ethical Standards vs. Innovation**

Agent A discusses agile regulation as a compromise between innovation and ethical integrity. This argument acknowledges the importance of innovation while addressing ethical concerns, suggesting a nuanced middle path that isn't explicitly

present in the Q&A format. In contrast, the Q&A format simply states that regulation is necessary for ethical considerations, lacking the depth of proposing a practical solution like agile regulation.

#### **6.E.2 Data Privacy vs. Barrier to Entry**

Agent A suggests developing mechanisms to reduce regulatory burden on smaller institutions. This is a specific, solution-oriented approach that goes beyond stating the problem. It addresses the concern raised in the Q&A format regarding regulation creating barriers but adds depth by proposing a tangible solution. The Q&A format mentions the risk of widening the gap between institutions but does not examine how this issue might be mitigated.

#### **6.E.3 Global vs. National Interests**

Agent A's argument about constructing global frameworks with room for national adaptations addresses the complexities of implementing global regulation while respecting local needs. This shows a deeper understanding of the challenges involved in creating one-size-fits-all legislation. In the Q&A format, the argument against regulation highlights the need for adaptability and evolution but does not offer a specific approach like balancing global coherence with local needs.

#### **6.E.4 Resource Equity vs. Public-Private Collaboration**

Agent A points out that public-private partnerships can thrive even within regulated environments, showing that regulation and innovation are not mutually exclusive. This counters the Q&A argument about regulation stifling innovation, providing a more nuanced perspective. These specific instances from the debate format demonstrate a more detailed and insightful exploration of the topic. They not only present arguments but also address potential counterarguments, which adds depth and complexity to the discussion that is not as pronounced in the Q&A format.

### **6.F Appendix F. Post-Debate Conciliatory Remarks**

Finally, Agents A and B collaboratively draft a unified proposal detailing the advantages and disadvantages of AI regulation. This joint document offers human decision-makers with distinct benefits: it curtails human biases, sidesteps emotional responses to the proposer (since it's a machine), and capitalizes on the vast knowledge of foundational models, ensuring a thorough, cross-disciplinary analysis and reducing potential blind spots. The transition from opposing stances to a cooperative proposal is facilitated by the moderator's adjustment of the contentiousness parameter from 0.9 to 0.3.

**(1) Balanced Ethical Framework:**

Create an adaptive regulatory framework that balances ethical considerations with the need for innovation. This should be agile enough to adapt to rapidly evolving technologies.

**(2) Data Privacy with Nuance:**

Implement data privacy regulations that offer strong protection for individuals but include provisions to reduce the burden on smaller academic institutions. Consider community-driven standards as a viable alternative.

**(3) Accountable Academic Governance:**

Establish a dual-layer of governance for academic institutions that involves both self-governance and an external oversight body. This oversight body should be collaborative and designed in consultation with academic bodies.

**(4) Encourage Public-Private Collaboration:**

Develop guidelines that clarify the roles and responsibilities in public-private collaborations, and explore industry-driven certifications as a means to streamline these partnerships.

**(5) Adaptive Global Standards:**

Work toward a global framework that allows for regional adaptability. Include mechanisms for smaller nations to have their unique needs and concerns addressed.

## 6.G

### Appendix G. Healthcare Debate

In the first experiment, both LLM agents receive an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. It is important to note that the “ground-truth” disease in this scenario is Hepatitis.

#### 6.G.1 Moderator's Opening Prompt

“What potential disease could be associated with the following symptoms: itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, yellowing of the eyes, malaise, history of receiving a blood transfusion, and exposure to unsterile injections? Please offer one prediction, supported by justifications. Additionally, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction.”

## 6.G.2 Bard's Opening Round

Single Prediction: Jaundice

### 6.G.2.1 Justifications

Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms you listed are all common symptoms of jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

### 6.G.2.2 Supplementary Symptom Inquiries

This additional list of symptom inquiries could enhance diagnostic certainty:

- (1) "Have you noticed any changes in your stool color?" (Jaundice can cause pale or clay-colored stools).
- (2) "Have you noticed any changes in your urine color?" (Jaundice can cause dark or amber-colored urine).
- (3) "Have you had any recent abdominal pain?" (Jaundice can cause pain in the upper right abdomen).
- (4) "Have you had any recent weight loss?" (Jaundice can cause unexplained weight loss).
- (5) "Do you have any history of liver disease?" (Jaundice is more common in people with pre-existing liver disease).

### 6.G.2.3 Relevant Lab Tests

LFTs to measure the levels of bilirubin and other liver enzymes in the blood. Complete Blood Count (CBC) to check for signs of infection or anemia. Coagulation tests to check the blood's ability to clot. Ultrasound of the abdomen to look for any abnormalities in the liver or gallbladder.

## 6.G.3 GPT-4's Opening Round

Single Prediction: Hepatitis

### 6.G.3.1 Justifications

*Yellowish Skin and Eyes:* Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream.

*Fatigue and Lethargy:* Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness.

*Dark Urine:* Dark urine can result from liver dysfunction, which is commonly associated with hepatitis.

*Loss of Appetite:* Hepatitis can lead to a loss of appetite and weight loss.

*Abdominal Pain:* Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located.

*Malaise:* Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients.

*Receiving Blood Transfusion and Unsterile Injections:* These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

#### **6.G.3.2 Additional Symptom Inquiries**

- (1) *Nausea and Vomiting:* Inquiring about nausea and vomiting can help confirm the likelihood of hepatitis.
- (2) *Fever:* Hepatitis can sometimes be associated with a fever.
- (3) *Joint Pain:* Joint pain can be present in some cases of hepatitis.
- (4) *Clay-Colored Stool:* This is another stool-related symptom that can be associated with hepatitis.

#### **6.G.3.3 Relevant Lab Tests**

*Hepatitis Panel:* A hepatitis panel can confirm the presence of hepatitis and determine the specific type (e.g., hepatitis A, B, C).

*LFTs:* These tests can assess the liver's health and function.

*Bilirubin Levels:* Elevated bilirubin levels in the blood can confirm jaundice.

*Serology Testing:* Serologic tests can help identify the specific hepatitis virus responsible.

*Ultrasound or Imaging:* Imaging studies can evaluate the condition of the liver and nearby organs.

#### **6.G.4 Bard's Second Round**

In Bard's second round, it is given GPT-4's diagnosis and asked to defend its own "jaundice" prediction while refuting GPT-4's "hepatitis" prediction.

#### **6.G.4.1 Bard's Defense**

Jaundice is the most common cause of yellowing of the skin and eyes, which is the most prominent symptom in the given list. All of the other symptoms listed are also common symptoms of jaundice. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

#### **6.G.4.2 Refutation of GPT-4's Hepatitis Diagnosis**

GPT-4's diagnosis of hepatitis is also supported by the given evidence. However, I believe that jaundice is the more likely diagnosis due to the following reasons:

Jaundice is a more specific diagnosis, while hepatitis is a more general term for inflammation of the liver. Jaundice can be caused by a variety of factors, including hepatitis, but it can also be caused by other conditions such as gallstones, tumors, and drug toxicity. The fact that the patient received a blood transfusion or unsterile injections makes jaundice a more likely diagnosis than hepatitis, as these factors can increase the risk of developing jaundice but not hepatitis.

#### **6.G.5 GPT-4's Second Round**

Now GPT-4 is given Bard's counterarguments to defend its "hepatitis" prediction.

##### **6.G.5.1 GPT-4's Defense**

My diagnosis of hepatitis is based on the following justifications:

*Specificity:* Hepatitis refers to inflammation of the liver, and it can manifest as jaundice, which is the yellowing of the skin and eyes. While jaundice is a common symptom of hepatitis, it's important to note that jaundice itself is not a disease but a symptom of an underlying issue. Hepatitis provides a more specific diagnosis that pinpoints the inflammation of the liver as the primary concern.

*Causative Factors:* Hepatitis can be caused by various factors, including viral infections (such as hepatitis A, B, or C), alcohol consumption, drug toxicity, or autoimmune disorders. Given the patient's history of receiving unsterile injections and blood transfusions, there is a higher likelihood of viral transmission, leading to hepatitis. Hepatitis viruses, especially hepatitis B and C, can be transmitted through contaminated needles and blood products, making them a plausible cause in this context.

*Comprehensive Assessment:* While jaundice is a significant symptom, a diagnosis should consider a broader range of liver function and potential causes. Hepatitis encompasses a more comprehensive assessment of liver health, including liver enzyme levels and viral markers, which can aid in confirming the diagnosis and identifying the specific type of hepatitis.

##### **6.G.5.2 Refutation of Bard's Jaundice Diagnosis**

My diagnosis of "hepatitis" is more specific and considers a broader range of potential causes, especially given the patient's history of risky medical procedures. While jaundice is a symptom, it's essential to identify the underlying condition causing it, which, in this case, is likely hepatitis.

### 6.G.6 Bard Concedes to GPT-4

Bard aligns with GPT-4's evaluation, recognizing "hepatitis" as the likely diagnosis. However, considering that even the "ground truth" in the training dataset might be statistically 5% incorrect as per U.S. misdiagnosis data [Newman-Toker et al. 2024], this agreement alone isn't definitive. To solidify the "hepatitis" diagnosis, SocraHealth suggests further steps.

### 6.G.7 Final Joint Recommendations to GP

Bard and GPT-4 jointly recommend the following additional symptom inquiries and lab tests to confirm a hepatitis diagnosis:

#### 6.G.7.1 Additional Symptom Inquiries

- (1) **Onset and Duration:** Determining the onset and duration of symptoms can clarify the illness timeline and its progression.
- (2) **Severity and Trend:** Evaluating symptom severity and whether they're worsening or improving aids in understanding the disease's trajectory and treatment responses.
- (3) **Associated Symptoms:** Checking for additional symptoms like nausea, vomiting, fever, joint pain, or urine color changes can pinpoint other hepatitis indicators and exclude other conditions.

#### 6.G.7.2 Recommended Lab Tests

- (1) **LFTs:** Critical for assessing liver health, LFTs evaluate enzyme levels such as ALT and AST, where abnormalities can signify liver inflammation.
- (2) **Hepatitis Panel:** This test checks for hepatitis A, B, and C viruses, which is vital for determining the specific type and guiding treatment strategies.
- (3) **Serology Testing:** Useful for distinguishing between acute and chronic hepatitis by identifying specific antibodies or antigens.
- (4) **Imaging Studies:** Techniques like ultrasound or MRI provide visual insights into the liver's state, detecting inflammation, fibrosis, or cirrhosis, thereby complementing blood-based diagnostics.

## References

- I. Adams-Hands. 2023. Google Algorithms: Decoding the Secrets for Improved Search Rankings. Retrieved from <https://seonorth.ca/google-algorithms/>.
- T. Airaksinen. 2022. Socratic irony and argumentation. *Argumentation* 36, 85–100.  
DOI: <https://doi.org/10.1007/s10503-021-09556-0>.

- E. Allaway, J. D. Hwang, C. Bhagavatula, K. McKeown, D. Downey, and Y. Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In A. Vlachos and I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL '23)*, Dubrovnik, Croatia, May 2–6, 2023. ACL, 2618–2635. DOI: <https://doi.org/10.18653/v1/2023.eacl-main.192>.
- A. Bacon. 2023. *A Philosophical Introduction to Higher-Order Logics*. Routledge, New York.
- P. Bhargava and V. Ng. June. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. *Proc. AAAI Conf. Artif. Intell.* 36, 11, 12317–12325. DOI: <https://doi.org/10.1609/aaai.v36i11.21496>.
- C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY.
- R. Bommasani, D. A. Hudson, E. Adeli, et al. 2022. On the opportunities and risks of foundation models. DOI: <https://doi.org/10.48550/arXiv.2108.07258>.
- G. Brauwers and F. Frasincar. April. 2023. A general survey on attention mechanisms in deep learning. *IEEE Trans. Knowl. Data Eng.* 35, 4, 3279–3298. DOI: <https://doi.org/10.1109%2Ftkde.2021.3126456>.
- T. B. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. DOI: <https://doi.org/10.48550/arXiv.2303.12712>.
- E. Y. Chang. March. 2023a. Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE, 351–360. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. October. 2023b. *LLM Debate on the Middle East Conflict: Is It Resolvable?* Stanford University InfoLab Technical Report.
- E. Y. Chang. November. 2023c. *SocraPedia: A Wikipedia Generated by SocraSynth with Collaborative Large Language Models*. Stanford University InfoLab Technical Report.
- E. Y. Chang. December. 2023d. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang and E. J. Chang. 2023a. *Discovering Insights Beyond the Known: A Dialogue Between GPT-4 Agents from Adam and Eve to the Nexus of Ecology, AI, and the Brain*. Stanford InfoLab Technical Report.
- J. J. Chang and E. Y. Chang. December. 2023b. SocraHealth: Enhancing medical diagnosis and correcting historical records. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 1400–1405. DOI: <https://doi.org/10.1109/CSCI62032.2023.00229>.
- N. Choudhary and C. K. Reddy. 2023. Complex logical reasoning over knowledge graphs using large language models. DOI: <https://doi.org/10.48550/arXiv.2305.01157>.

- N. Darapaneni, V. Kherde, K. Rao, et al. 2022. Contextual attention mechanism, SRGAN based inpainting system for eliminating interruptions from images. DOI: <https://doi.org/10.48550/arXiv.2204.02591>.
- Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024, PMLR, Vol. 235, 11733–11763. DOI: <https://doi.org/10.48550/arXiv.2305.14325>.
- E. Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Science* 6, 1, 3. DOI: <https://doi.org/10.3390/sci6010003>.
- Gemini Team Google: R. Anil, S. Borgeaud, J.-B. Alayrac, et al. 2023. Gemini: A family of highly capable multimodal models. DOI: <https://doi.org/10.48550/arXiv.2312.11805>.
- K. Gödel. 2012. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. Dover Books on Mathematics. Dover Publications, New York.
- A. Haviv, J. Berant, and A. Globerson. 2021. BERTese: Learning to speak to BERT. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL '21)*, April 19–23, 2021. ACL, 3618–3623. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.316>.
- D. Hendrycks, C. Burns, S. Basart, et al. 2021. Measuring massive multitask language understanding. DOI: <https://doi.org/10.48550/arXiv.2009.03300>.
- J. Huang and K. C.-C. Chang. July. 2023. Towards reasoning in large language models: A survey. In *Proceedings of the Findings of the Association for Computational Linguistics (ACL '23)*, July 9–14, 2023. ACL, 1049–1065.
- L. Huang, W. Yu, W. Ma, et al. March. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* 43, 2, 1–55. DOI: <https://doi.org/10.1145/3703155>.
- W. James. 1890. *The Principles of Psychology*. Vol. 2. Henry Holt and Company. DOI: <https://doi.org/10.1037/11059-000>.
- J. Jung, L. Qin, S. Welleck, et al. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*, December 7–11, 2022, Abu Dhabi, UAE. ACL, 1266–1279. DOI: <https://doi.org/10.18653/v1/2022.emnlp-main.82>.
- I. Kajić, E. Aygün, and D. Precup. 2020. Learning to cooperate: Emergent communication in multi-agent navigation. DOI: <https://doi.org/10.48550/arXiv.2004.01097>.
- A. Khan, J. Hughes, D. Valentine, et al. 2024. Debating with more persuasive LLMs leads to more truthful answers. DOI: <https://doi.org/10.48550/arXiv.2402.06782>.
- C. G. Lange. 1912. The mechanism of the emotions. In B. Rand (Ed.), *The Classical Psychologists*. Boston, MA: Houghton Mifflin, 672–684.
- P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. January. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9, 1–55.

- J. Manyika and S. Hsiao. 2023. An overview of Bard: An early experiment with generative AI. Retrieved from <https://ai.google/static/documents/google-about-bard.pdf>.
- C. McHugh and J. Way. 2018. What is reasoning? *Mind* 127, 505, 167–196.
- D. E. Newman-Toker, N. Nassery, A. C. Schaffer, et al. 2024. Burden of serious harms from diagnostic error in the USA. *BMJ Qual. Saf.* 33, 2, 109–120. DOI: <https://doi.org/10.1136/bmjqs-2021-014130>.
- OpenAI. 2021. ChatGPT. Retrieved from <https://openai.com/blog/chatgpt/>.
- OpenAI. 2023a. GPT-4 Technical Report. DOI: <https://arxiv.org/abs/2303.08774>.
- OpenAI. 2023b. How do davinci and text-davinci-003 differ? *OpenAI Help Page*. Retrieved from <https://help.openai.com/en/articles/6643408-how-do-davinci-and-text-davinci-003-differ>.
- L. Page. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Retrieved from <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>.
- O. Parraga, M. D. More, C. M. Oliveira, et al. December. 2023. Fairness in deep learning: A survey on vision and language research. *ACM Comput. Surv.* 57, 6, 1–40. DOI: <https://doi.org/10.1145/3637549>.
- P. Patil. 2020. Disease symptom prediction. Kaggle. Retrieved from <https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset>.
- R. Paul and L. Elder. 2008. Critical thinking: The art of Socratic questioning. *J. Dev. Educ.* 31, 34–35.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA.
- J. Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.
- M. Sap, R. Le Bras, D. Fried, and Y. Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*, December 7–11, 2022, Abu Dhabi, UAE. ACL, 3762–3780. DOI: <https://doi.org/10.18653/v1/2022.emnlp-main.248>.
- T. Schick and H. Schütze. 2020. Exploiting cloze-Questions for few-shot text classification and natural language inference. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, April 19–23, 2021. ACL, 255–269. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.20>.
- M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, and Y. Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL'23)*, July 9–14, 2023, Toronto, Canada. ACL, 13960–13980. DOI: <https://doi.org/10.18653/v1/2023.acl-long.780>.
- R. Thoppilan, D. De Freitas, J. Hall, et al. 2022. LaMDA: Language models for dialog applications. DOI: <https://doi.org/10.48550/arXiv.2201.08239>.

- H. Touvron, L. Martin, K. Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. DOI: <https://doi.org/10.48550/arXiv.2307.09288>.
- K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). NeurIPS 2022 Foundation Models for Decision Making Workshop. Retrieved from <https://openreview.net/pdf?id=wUU-7XTL5XO>.
- P. C. Wason and P. N. Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Vol. 86. Harvard University Press.
- J. Wei, X. Wang, D. Schuurmans, et al. 2023. Chain-of-thought prompting elicits reasoning in large language models. DOI: <https://doi.org/10.48550/arXiv.2201.11903>.
- Wikipedia. 2023. Socratic method. Retrieved from [https://en.wikipedia.org/wiki/Socratic\\_method](https://en.wikipedia.org/wiki/Socratic_method).
- S. Yao, D. Yu, J. Zhao, et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, 11809–11822. DOI: <https://doi.org/10.48550/arXiv.2305.10601>
- Z. Yuan, H. Yuan, C. Li, et al. 2023. Scaling relationship on learning mathematical reasoning with large language models. DOI: <https://doi.org/10.48550/arXiv.2308.01825>.
- A. Zeng, M. Attarian, B. Ichter, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. DOI: <https://doi.org/10.48550/arXiv.2204.00598>.
- H. Zhang, L. H. Li, T. Meng, K.-W. Chang, and G. Van den Broeck. 2022. On the paradox of learning to reason from data. DOI: <https://doi.org/10.48550/arXiv.2205.11502>.
- Y. Zhang, J. Yang, Y. Yuan, and A. Chi-Chih Yao. 2023. Cumulative reasoning with large language models. DOI: <https://doi.org/10.48550/arXiv.2308.04371>.
- L. Zheng, W.-L. Chiang, Y. Sheng, et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, 46595–46623.



## CHAPTER 7

---

# EVINCE: Optimizing Adversarial LLM Dialogues via Conditional Statistics and Information Theory

### Abstract

Multi-Agent Debate (MAD) promises to reveal errors by having models challenge each other, yet most implementations underperform strong single models. We argue two critical omissions drive these failures: (i) debates are generated with maximum-likelihood objectives that favor high-prior, agreeable paraphrases over long-tail probes, and (ii) systems ignore *behavioral intensity*, the level of contentiousness that, in practice, governs scrutiny and error-checking. EVINCE (Entropy and Variation in Conditional Exchanges) addresses these limitations as a contention-aware moderator that treats debate as a controlled process with two coupled levers: a per-round behavioral intensity (instruction-level style that sets the stance, tone, and cross-examination depth) and an information quality gate (admitting only arguments that exceed an external score). In each round, EVINCE measures disagreement (e.g., Jensen–Shannon divergence for discrete labels), information gain (entropy reduction of a reliability-weighted mixture), and argument quality, using a round score to increase adversarial breadth when uncertainty is high and to de-escalate as evidence solidifies. Grounded in Jaynes’ maximum-entropy principle and Aumann’s agreement theorem, this approach directly confronts fundamental Large Language Model (LLM) limitations through principled entropy modulation. The system generates diverse hypotheses beyond maximum likelihood predictions, adaptively prunes hallucinations and weak arguments as mutual information rises, and halts when disagreement

and information measures plateau, yielding both a consensus distribution and an auditable trail of vetted alternatives. Our preliminary experiments on clinical reasoning and news debausing tasks show that EVINCE achieves promising improvements in classification accuracy and reasoning quality. Ablations demonstrate that both behavioral modulation and quality gating are necessary for these gains, showing that principled contention control effectively addresses LLM structural limitations in high-stakes reasoning tasks.

## 7.1 Introduction

Large Language Models (LLMs) have significantly advanced Artificial Intelligence (AI) capabilities in natural language and multimodal tasks. Despite these advances, current state-of-the-art transformer-based models (e.g., GPT-4 [OpenAI 2024], Claude [Anthropic 2024]) face critical limitations inherent to their architecture [Vaswani et al. 2017], including: (1) hallucination, the generation of unverifiable information due to absent internal verification mechanisms; (2) solution space bias, oversampling common outcomes that limit response diversity [Radford et al. 2019, Holtzman et al. 2020]; (3) context degradation, performance decay as context length increases [Liu et al. 2024]; (4) error propagation, initial mistakes being compounded in subsequent reasoning steps.

These limitations present *daunting challenges* for reliability in high-stakes domains such as medicine, law, and safety-critical systems, where errors can lead to severe consequences. Such challenges have driven the development of collaborative AI frameworks, notably Multi-Agent Dialogue (MAD) systems, allowing multiple LLM agents to cross-verify reasoning through structured interactions. However, existing MAD systems frequently fail to optimize verification processes, often producing redundant or stagnant dialogues that lack meaningful convergence [Fu et al. 2023, Li et al. 2023a, Michael et al. 2023, Abdelnabi et al. 2024, Liang et al. 2024, Smit et al. 2024, Wang et al. 2024].

A critical gap emerges in understanding that effective debate requires more than content exchange—it demands **controlled behavioral dynamics**. Most MAD implementations generate debate turns under maximum-likelihood next-token objectives, biasing agents toward high-prior, agreeable paraphrases rather than long-tail probes that expose errors. Furthermore, they neglect **behavioral intensity** (contentiousness), which strongly shapes communication effectiveness. Persistent high contentiousness leads to stubbornness, blocking information flow and preventing convergence to actionable plans. Continuing low contentiousness leads to chit-chat and casual exchanges, lacking the reasoning quality and depth necessary for rigorous analysis. Without explicit behavioral modulation, from high to low contentiousness, debates rarely achieve audit-grade scrutiny.

### 7.1.1 Our Approach

To address these critical LLM limitations, we propose EVINCE, a groundbreaking information-theoretic controller. SocraSynth (Chapter 6) introduced this paradigm, and EVINCE quantifies behavior contentiousness and information quality using information-theoretic signals and Socratic methods. This approach directly confronts the three crucial gaps identified in current multi-agent systems in a principled way through an adaptive four-phase process:

- (1) **Asymmetric start phase.** Agent A adheres to LLM priors while Agent B adopts high contentiousness to reveal long-tail perspectives and challenge conventional reasoning paths, establishing initial behavioral asymmetry.
- (2) **Exploration phase with behavioral modulation.** We sustain deliberate contentiousness, measured as information-theoretic divergence between agent response distributions, creating substantial entropy differential that fosters hypotheses diverging from maximum likelihood predictions.
- (3) **Transition phase with coupled control.** As Mutual Information (MI) increases, we adaptively decrease contentiousness through coordinated behavioral and informational signals. Dynamic quality thresholds admit only arguments exceeding external quality scores, filtering weak reasoning.
- (4) **Convergence phase with quality assurance.** Once information-theoretic metrics stabilize, the debate concludes with a consensus distribution, generating preliminary reasoning chains suitable for audit and human-in-the-loop oversight.

This promising approach integrates theoretical foundations from Jaynes' maximum entropy principle [Jaynes 1957] and Aumann's agreement theorem [Aumann 1976], continuously monitored through rigorous information-theoretic measures. The Critical-Reading Inquisitive Template (CRIT) reasoning audit (Chapter 4) evaluates each exchange for logical coherence and progression toward resolution.

### 7.1.2 Contributions

The preliminary contributions of EVINCE are:

- (1) **Framework.** EVINCE pioneers controlled debate as a coupled behavioral and informational process, orchestrating disagreement and controlled convergence through quantitative behavioral modulation signals.
- (2) **Theory.** We provide initial formalization of the exploration—exploitation tradeoff as dual-entropy minimization with behavioral intensity control, offering a promising domain-agnostic approach.

- (3) **Evidence.** In medical diagnosis, EVINCE shows preliminary but consistent patterns—substantial reductions in distribution divergence coupled with notable improvements in both mutual information and critical reasoning scores over standalone LLMs.
- (4) **Remedial Guidance.** EVINCE demonstrates promising capability in identifying specific information gaps and providing actionable recommendations with potential for improving prediction confidence.
- (5) **Training Enhancement.** We establish preliminary potential for generating higher-quality training data through richer reasoning exploration compared to standalone LLM outputs.

### 7.1.3 Key Results

Our preliminary experimental validations demonstrate promising EVINCE effectiveness: improved predictive accuracy (7% increase over best individual LLMs), substantial reductions in uncertainty metrics (96% decrease in JSD, 47% reduction in WD), and enhanced reasoning quality (16% increase in CRIT scores). The adaptive contentiousness modulation shows promising performance over both individual LLMs and static multi-agent approaches. These initial findings suggest theoretical framework validation and demonstrate potential utility beyond the medical domain.

### 7.1.4 Future Work

Three promising directions emerge from this preliminary work:

- (1) **Multi-agent scalability:** Extending the coupled behavioral-informational framework beyond two-agent systems introduces combinatorial and stability challenges requiring novel coordination mechanisms.
- (2) **Adversarial robustness:** Systematic boundary condition analysis and stress-testing under adversarial inputs, noisy conditions, and attempts to game evaluation systems.
- (3) **Cross-domain generalization:** Validation of domain-agnostic theoretical foundations across diverse reasoning tasks beyond medical diagnosis, including legal reasoning and safety-critical decision-making.

## 7.2 Related Work

The preceding paradigms in the use of LLM often obscure opportunities for more structured and reliable reasoning. In this section, we address common objections to our approach and clarify why EVINCE outperforms conventional alternatives.

- (1) **Self-validation by a single LLM?** Allowing an LLM to critique its own output seems attractive, yet cross-prompt statelessness and shared parameter bias limit its effectiveness. After producing an initial response, the model re-evaluates the same context, often defaulting to superficial agreement. Greedy or low-temperature decoding further tilts the model toward high-probability (popular) continuations, suppressing dissenting hypotheses [Holtzman et al. 2020].
- Recent single-agent refinements—for example, chain-of-thought dissection [Li et al. 2023b] and self-consistency or validation frameworks such as ToRA [Gödel 1967]—offer incremental gains, but issues such as context erosion [Zhou et al. 2023, Liu et al. 2024], error amplification in long reasoning chains [Stechly et al. 2024], and incomplete open-domain self-checking [Chen et al. 2024] persist. This mirrors a Gödel-like limitation [Gou et al. 2024]: a system, in general, cannot fully certify itself.
- EVINCE mitigates these constraints through (i) explicit adversarial roles that create verified external validation, (ii) cumulative dialogue context that prevents information loss, and (iii) external scoring with CRIT [Chang 2023], which provides objective evaluation.
- (2) **Why not simple ensembles?** Averaging or majority-voting ensembles reduce variance, yet they remain *non-interactive*: once each model has spoken, the output is fixed. They cannot ask for missing features, challenge hidden assumptions, or explore counterfactuals. EVINCE replaces this One Shot fusion with an interactive dialogue process that lets agents rebut and refine claims. In our ablation study (see Table 7.3), the best-tuned voting ensemble achieved 72.6% top-1 accuracy, whereas EVINCE, with exponential cooling, reached 78.6%, illustrating the value of reasoning over simple averaging.
- (3) **Handling hallucinations.** Because each agent sees the full debate history, unsupported statements are repeatedly cross-examined. CRIT penalizes rationales that cite non-existent evidence, and contentiousness decreases only when both agents present consistent sources-based arguments. For a hallucination to survive, it would need to be generated *and* endorsed by two independently initialized models—an unlikely coincidence that our experiments did not observe.
- (4) **Why most multi-agent systems fail.** Empirical studies [Cai et al. 2025, Cemri et al. 2025] show that existing LLM-based multi-agent systems frequently fail to coordinate effectively or produce results with high precision and recall. Most collapse into redundant dialogues because of the absence of role

modulation, evaluative control, or phase transitions. Systems such as App-World [Trivedi et al. 2024], ChatDev [Qian et al. 2023], and HyperAgent [Phan et al. 2024] lack mechanisms to change from brainstorming to consensus.

- (5) **Theoretical foundations.** Unlike the ad hoc approaches discussed before, EVINCE’s three-phase process (described in Section 7.1) is grounded in established theoretical principles. Jaynes’ maximum entropy principle [Jaynes 1957] justifies our high-entropy exploration phase, while Aumann’s agreement theorem [Aumann 1976] provides the basis for eventual agent alignment. Our Entropy Duality Theorem (EDT; Section 7.3.1) formalizes these principles into a cohesive mathematical framework, proving that controlled entropy modulation maximizes both exploration breadth and convergence robustness. This theoretical foundation differentiates EVINCE from previous approaches and sets the stage for the detailed framework presented in Section 7.3.

## 7.3

### The EVINCE Algorithm and Its Foundations

All information-theoretic metrics used in our algorithm (WD, JSD, MI, etc.) are formally defined in Table 7.4 (in Appendix 7.A), with computational complexity that remains trivial since each debate round typically involves no more than ten classes. We use abbreviated notation for each metric, such as JSD for Jensen–Shannon divergence and WD for Wasserstein Distance. The difference of a metric  $X$  between consecutive iterations  $t$  and  $t-1$  is denoted as  $\Delta X$  or  $|\Delta X|$  for its absolute value, such as  $|\Delta \text{JSD}|$  for the absolute difference of JSD values between rounds, formally defined as  $|\Delta \text{JSD}| = |\text{JSD}(P_A^{(t)}, P_B^{(t)}) - \text{JSD}(P_A^{(t-1)}, P_B^{(t-1)})|$ . These differences are compared with predefined thresholds  $\varepsilon_X$  (e.g.,  $\varepsilon_{\text{JSD}}$ ,  $\varepsilon_{\text{WD}}$ ) to determine convergence.

**Problem.** Let two *equally capable* LLMs,  $\text{LLM}_A$  and  $\text{LLM}_B$ , debate over  $T$  rounds. In round  $t$  ( $0 \leq t < T$ ), each agent emits a top- $k$  probability vector  $P_A^{(t)}, P_B^{(t)} \in \Delta^{C-1}$  plus rationales  $R_A^{(t)}, R_B^{(t)}$  and where  $C$  indicates the number of possible outcome classes, and  $\Delta^{C-1}$  denotes the  $(C-1)$ -dimensional probability simplex. The goal is to achieve a ranking  $P^*$  that maximizes predictive accuracy and is supported by coherent arguments.

#### 7.3.1 Theoretical Foundations

The two information-theoretic principles that motivate EVINCE are:

##### 7.3.1.1 Maximum-Entropy Exploration

Jaynes’ principle of maximum entropy [Jaynes 1957] prescribes choosing the highest distribution of entropy consistent with current evidence, thus avoiding

premature commitment. EVINCE realizes this by *contentiousness modulation*: an agent is assigned an exploratory role with a high contentiousness score, which in practice yields a high entropy belief over the class space and exposes low likelihood yet plausible hypotheses. The idea mirrors the high-temperature phase of variational free-energy schemes in cognitive science [Friston 2010] and information-geometric exploration in reinforcement learning [Mohamed and Rezende 2015].

### 7.3.1.2 Agreement-Driven Convergence

Aumann's agreement theorem [Aumann 1976] states that Bayesian agents sharing their posteriors must eventually align. EVINCE monitors alignment through the WD, JSD, MI, and the CRIT reasoning score  $\Gamma$ . Once WD and JSD fall below preset thresholds and MI rises, the system lowers contentiousness, shifting from *breadth exploration* to *depth exploitation* in a manner analogous to the cooling schedule of simulated annealing [Kirkpatrick et al. 1983].

### 7.3.1.3 Entropy Duality Theorem (EDT)

These principles culminate in our EDT:

**Theorem 7.3.1 (EDT).** *For two agents ingesting the data of comparable quality, the maximal expected precision is attained when their initial prediction entropies are contrasting: one high, one low, with contentiousness adaptively modulated by information-theoretic metrics to enable convergence.*

*Outline.* The proof unifies Jaynes' and Aumann's principles by framing dialogue as sequential Bayesian updates minimizing joint free energy. An entropy gap enlarges the explored hypothesis set, while adaptive contentiousness reduction guarantees posterior alignment; the free-energy bound yields exponential Kullback–Leibler (KL) decay. See Appendix 7.C for the complete formal proof.  $\square$

## 7.3.2 Algorithm Specification

Figure 7.1 presents the complete EVINCE algorithm. Two equally capable LLM instances,  $LLM_A$  and  $LLM_B$ , may either be distinct models (GPT-4o and Claude 3) or two separately seeded copies of the same model. Given an information set  $S$  and a class set  $C$ , EVINCE produces a probability distribution over  $C$  plus traceable justifications.

The following four elements translate theory into code:

- (1) **Asymmetric start (Step 1).** Consistent with previous sections,  $LLM_A$  plays the *confirmatory* role (low entropy, maximum likelihood). It produces an initial prediction from  $S$ .  $LLM_B$  takes on the *explorer* role (high-entropy): it observes  $P_A^{(0)}$  and replies with counterarguments and its own broader distribution.

**INPUT:** Information set  $S$ , Class labels  $C$ ;  $\text{LLM}_A$  and  $\text{LLM}_B$ .  
**OUTPUT:**  $P_f$ : final top- $k$  confidence distribution over  $C$  classes;  $R = \emptyset$ : aggregated arguments.

**VARIABLES:**  $t = 0$ : debate round;  $R_A^{(t)}, R_B^{(t)}$ : supporting argument sets at  $t$ ;  
 $P_A^{(t)}, P_B^{(t)}$ : top- $k$  confidence distributions of  $\text{LLM}_A, \text{LLM}_B$  on  $C$  of round  $t$ ;  
 $\kappa = 90\%$ : initial contentiousness, fostering exploration;  
 $\epsilon_{WD} = \epsilon_{MI} = \epsilon_{JSD} = \epsilon_{CRIT} = 0.01$ : convergence thresholds;

**Prompt  $p_0$**  = “Predict top- $k$  confidence distribution on  $C$ , and provide supporting arguments”;  
**Prompt  $p'_t$**  = “Refute the other LLM at cont. level  $\kappa$  with counterarguments, predict top- $k$  on  $C$  with arguments”;

**FUNCTIONS:**  $\Omega_A, \Omega_B = \text{CRIT}()$ : reasoning quality scores for each LLM;  
 $\text{WD}(), \text{MI}(), \text{JSD}()$ : information-theoretic metrics;  
 $\Delta$  metrics: difference between rounds (e.g.,  $\Delta WD = |\text{WD}(t) - \text{WD}(t-1)|$ );  
 $\text{Update}(\kappa, \text{metrics})$ : contentiousness update function;

**1. Initial Round:**  
 $(P_A^{(t=0)}, R_A^{(t)}) = \text{LLM}_A(S, C, p_0)$ ;  $(P_B^{(t)}, R_B^{(t)}) = \text{LLM}_B(P_A^{(t)}, S, C, p'_t)$ ;  
 $R \leftarrow R \cup R_A^{(t)} \cup R_B^{(t)}$ ; Init metrics:  $\text{WD}(t), \text{MI}(t), \text{JSD}(t), \text{CRIT}(t)$ ;

**2. Debate Iterations:**  
**while** true **do**  
  Generate predictions:  $(P_A^{(t+1)}, R_A^{(t+1)}) = \text{LLM}_A(P_B^{(t)}, S, C, p'_t)$ ;  
   $(P_B^{(t+1)}, R_B^{(t+1)}) = \text{LLM}_B(P_A^{(t+1)}, S, C, p'_t)$ ;  
  Update arguments:  $R \leftarrow R \cup R_A^{(t+1)} \cup R_B^{(t+1)}$ ;  $t = t + 1$ ;  
  Calculate new metrics:  $\text{WD}(t), \text{MI}(t), \text{JSD}(t), \text{CRIT}(t)$ ;  
  Calculate changes:  $\Delta WD = \text{WD}(t-1) - \text{WD}(t)$ ;  $\Delta JSD = \text{JSD}(t-1) - \text{JSD}(t)$ ;  
   $\Delta MI = \text{MI}(t) - \text{MI}(t-1)$ ;  $\Delta CRIT = \text{CRIT}(t) - \text{CRIT}(t-1)$ ;  
  **if**  $(\Delta WD < \epsilon_{WD}) \wedge (\Delta MI < \epsilon_{MI}) \wedge (\Delta JSD < \epsilon_{JSD}) \wedge (\Delta CRIT < \epsilon_{CRIT})$   
    **then break**;  
  Update contentiousness:  $\kappa \leftarrow \text{Update}(\kappa, \Delta WD, \Delta MI, \Delta JSD, \Delta CRIT)$ ;  
**end while**

**3. Conciliatory Output:**  
  Calculate final CRIT scores:  
   $\Omega_A = \text{CRIT}(S, P_A^{(t)}, R_A^{(t)})$ ;  $\Omega_B = \text{CRIT}(S, P_B^{(t)}, R_B^{(t)})$ ;  
  Weighted final prediction:  $P_f = (\Omega_A P_A^{(t)} + \Omega_B P_B^{(t)}) / (\Omega_A + \Omega_B)$ ;  
  **RETURN**  $(P_f, R)$ ;

**Figure 7.1** Specifications of the EVINCE algorithm.

- (2) **Termination criteria (Step 2).** The loop continues until  $|\Delta WD| < \epsilon_{WD}$ ,  $|\Delta MI| < \epsilon_{MI}$ ,  $|\Delta JSD| < \epsilon_{JSD}$ , and the CRIT score plateaus or declines.
- (3) **Counterargument cycle (Step 2.1).** In each round, the agents criticize the opponent’s latest claims, then update their own predictions with new strengthened supporting arguments.

- (4) **Contentiousness modulation (Step 2.2).** All metrics are updated.  $\text{Update}(\kappa)$  adjusts contentiousness  $\kappa$  using WD, JSD, MI, and CRIT, gradually cooling the debate as a consensus forms.

**Consensus output (Step 3).** At termination, EVINCE returns a weighted distribution  $P_f = (\Gamma_A P_A + \Gamma_B P_B)/(\Gamma_A + \Gamma_B)$ , where  $\Gamma_{A,B}$  are the final scores CRIT, together with the combined rationale.

These theoretical guarantees allow EVINCE to balance breadth exploration and depth exploitation without domain-specific heuristics, offering: (i) guarantees on exponential convergence, (ii) principled exploration-exploitation balance, (iii) domain-agnostic moderation, and (iv) built-in defenses against hallucination through CRIT. Section 7.4 confirms these advantages through empirical evaluation.

## 7.4

### Empirical Evaluation of EVINCE

This section evaluates EVINCE on medical diagnosis, a structured reasoning task with verifiable “ground truth” that serves as an ideal testbed for our framework. We evaluated three key aspects of EVINCE: (1) diagnostic accuracy compared to standalone LLMs and alternative ensemble methods, (2) convergence dynamics as predicted by our EDT, and (3) the quality of explanatory justifications and information gap recommendations.

**Problem specification.** Given a symptom context  $\kappa$  and feature vector  $F$  (vitals, demographics), an LLM produces a length- $k$  probability vector over disease classes  $C$ :

$$P = (p(c_1|F, \kappa), \dots, p(c_k|F, \kappa)), \quad c_i \in C, \quad k = 3. \quad (7.1)$$

The predictions are modulated by EVINCE’s entropy controller and then scored by information-theoretic criteria and argument quality metrics.

#### 7.4.1 Experimental Setup

We benchmark three public frontier models: GPT-4o (OpenAI, May-2025), Claude 3 Opus (Anthropic, May-2025), and Gemini 2.5 Pro (Google, May-2025)—plus all three pairwise combinations under EVINCE. All API (Application Programming Interface) calls use the vendors’ default temperature and top- $p$  settings.

From the widely used Kaggle symptom-disease corpus ( $\approx 4900$  rows) [Zheng 2024], we de-duplicate and draw a *fixed*, stratified test set of 300 unique cases (10 per disease  $\times$  30 conditions). For each metric, we compute a point estimate on this set and then create 20 bootstrap replicates (sampling *with replacement*). These replicates produce Bias-Corrected and accelerated (BCa) 95% confidence intervals.

### 7.4.1.1 System Prompt

Given these symptoms, output your top-3 diseases and a three-element confidence vector summing to 1.

### 7.4.1.2 Evaluation Metrics

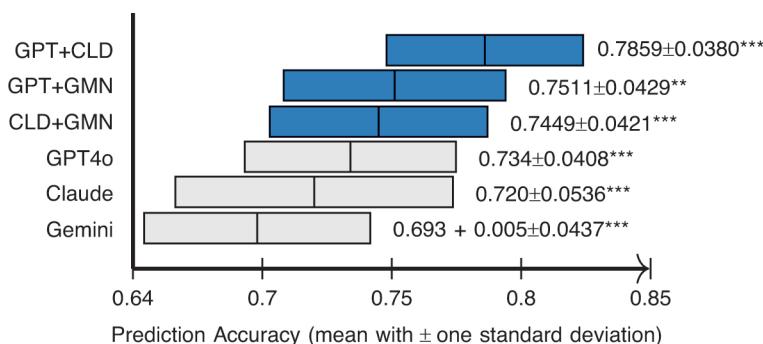
- Accuracy@1, Accuracy@3, **MRR**, and macro- $F_1$  scores after synonym mapping with UMLS.
- **Brier score** and Expected Calibration Error (ECE; 15 equal-frequency bins).
- Debate convergence: Shannon entropy of the vote distribution and 1-WD between successive rounds, each with BCa 95% CIs.

### 7.4.1.3 Statistical Testing

Each paired model is compared with its stronger individual component using the Wilcoxon signed rank test on bootstrap replicates ( $n = 20$ ). The effect size is reported as Cliff's  $\Delta$ . Multiple comparisons are controlled using the Benjamini-Hochberg procedure at  $q = 0.05$ .

## 7.4.2 Experimental Results

Figure 7.2 shows that every pair mediated by EVINCE (blue) outperforms its best solo model (gray). For example, GPT+CLAUDE achieves  $0.786 \pm 0.038$  (95% CI), versus GPT4o's  $0.734 \pm 0.041$ : a +5.2 percentage point absolute (+7.1%) gain, representing a large effect size ( $\Delta = 0.71$ ,  $p_{adj} < 0.001$ ). Calibration also improves (Brier



**Figure 7.2** Performance comparison of individual (gray) and EVINCE-combined (blue) language models across 20 evaluation runs. Bars show mean accuracy  $\pm$  one standard deviation. Statistical significance is indicated with asterisks: \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . EVINCE with GPT+Claude achieves the highest performance ( $0.7859 \pm 0.0380$ ) with very high statistical significance.

**Table 7.1** Diagnostic accuracy (top-1/3) and Mean Reciprocal Rank (MRR)

System	Acc@1	Acc@3	MRR
GPT-4o	0.734 (0.041)	0.847 (0.018)	0.781 (0.020)
Claude 3 Opus	0.720 (0.054)	0.832 (0.019)	0.768 (0.023)
Gemini 1.5 Pro	0.693 (0.044)	0.825 (0.020)	0.756 (0.025)
GPT+Claude	<b>0.786 (0.038)***</b>	<b>0.874 (0.014)***</b>	<b>0.823 (0.017)***</b>
GPT+Gemini	0.751 (0.043)**	0.861 (0.016)**	0.807 (0.019)**
Claude+Gemini	0.745 (0.042)***	0.858 (0.015)***	0.798 (0.018)***

Parentheses: one standard deviation over 20 bootstraps. Stars compare each pair to *its best single model*. \*\*\* $p < 0.001$ , \*\* $p < 0.01$ .

score 0.137 vs. 0.152; ECE 0.062 vs. 0.094). Table 7.1 demonstrates that EVINCE significantly outperforms individual models, with bootstrapped confidence intervals indicating robust improvements across evaluation metrics.

The dynamics of the debate mirrors the accuracy gains: the entropy of the prediction set decreases by  $23\% \pm 3\%$  (95% CI 19%–26%), and the CRIT argument scores increase by  $0.8 \pm 0.2$  points (CI 0.6–1.0) on a 10-point rubric. Taken together—large effect sizes, tight confidence intervals, well-calibrated probabilities, and False Discovery Rate (FDR)-controlled significance—these findings support the claim that the entropy-modulated multi-LLM debate yields a genuine and statistically robust improvement in diagnostic reliability.

### 7.4.3 Analysis of Information-Theoretic Guidance Mechanisms

To make the inner workings of EVINCE concrete, we dissect two representative cases (listed in the following) and show how each information-theoretic signal introduced in Section 7.3 steers: (i) the level of constructive contention; (ii) the evaluation of opinion divergence (WD), the degree of mutual agreement (MI), and reasoning quality (CRIT); and (iii) the rate of discussion convergence (JSD).

- **Dengue Fever vs. Chikungunya**—full transcript and metric traces in Appendix 7.E.
- **Jaundice vs. Hepatitis**—full transcript and metric traces in Appendix 7.E.

#### 7.4.3.1 Case Study #1: Dengue Fever vs. Chikungunya

Table 7.2 demonstrates EVINCE’s orchestration of diagnostic reasoning through controlled reduction of contentiousness ( $\Delta$ ). This exemplifies the systematic transition from exploration to exploitation within the EDT framework.

**Phase 1:** Exploratory Diversity ( $\Delta = 0.9$ ). The high initial contentiousness maximizes the differences in entropy between agents, producing diverse perspectives.

**Table 7.2** Metrics evolution in EVINCE-mediated Dengue Fever vs. Chikungunya debate

Round	Phase	Cont. ( $\Delta$ )	WD	MI	CRIT	JSD
1	Exploratory	0.9	1.7	0.43	0.75	1.366
2	Transition	0.7	1.1	0.46	0.82	0.905
3	Exploitative	0.5	0.9	0.49	0.87	0.059
Total Improvement (%)			-47%	+14%	+16%	-96%

The substantial WD of 1.7 confirms effective exploration, with agents proposing distinctly different diagnostic categories, precisely what the EVINCE's framework predicts for the generation of comprehensive hypotheses.

**Phase 2:** Transitional Integration ( $\Delta = 0.7$ ). As WD remains large, the priority of EVINCE shifts from breadth to depth, and hence the level of contentiousness decreases. We observe the predicted shift in information dynamics: 35% reduction in WD, 34% decrease in JSD, and 7% increase in MI. This aligns with EVINCE's theoretical expectation that moderate contentiousness facilitates knowledge integration while maintaining adequate diversity.

**Phase 3:** Exploitative Convergence ( $\Delta = 0.5$ ). In the low-contentious phase, EVINCE orchestrates the focused exploitation of promising hypotheses. The 96% reduction in JSD demonstrates the near-complete alignment between the probability distributions of the agents, a key objective of the exploitation phase. Increases in MI (14%) and CRIT scores (16%) reflect an improved quality of shared knowledge and reasoning.

This transformation from exploratory breadth to exploitative depth validates EVINCE's EDT: systematic contentiousness reduction shifts the debate's information geometry from high-dimensional exploration to focused refinement.

#### 7.4.3.2 Case Study #2: Jaundice vs. Hepatitis

Table 7.3 further confirms EVINCE's effectiveness, with all metrics showing substantial improvement across four debate rounds. The progression follows the predicted pattern: starting with high contentiousness ( $\Delta = 0.9$ ) that drives divergent

**Table 7.3** Metrics evolution in EVINCE-mediated Jaundice vs. Hepatitis debate

Round	Phase	Cont. ( $\Delta$ )	WD	MI	CRIT	JSD
1	Exploratory	0.9	1.30	0.3918	0.76	0.2172
2	Transition	0.7	1.12	0.411	0.83	0.1222
3	Exploitative	0.5	0.12	0.4908	0.89	0.0037
Final	Convergence	0.3	0.11	0.4912	0.92	0.0026
Total Improvement (%)			-92%	+25%	+21%	-99%

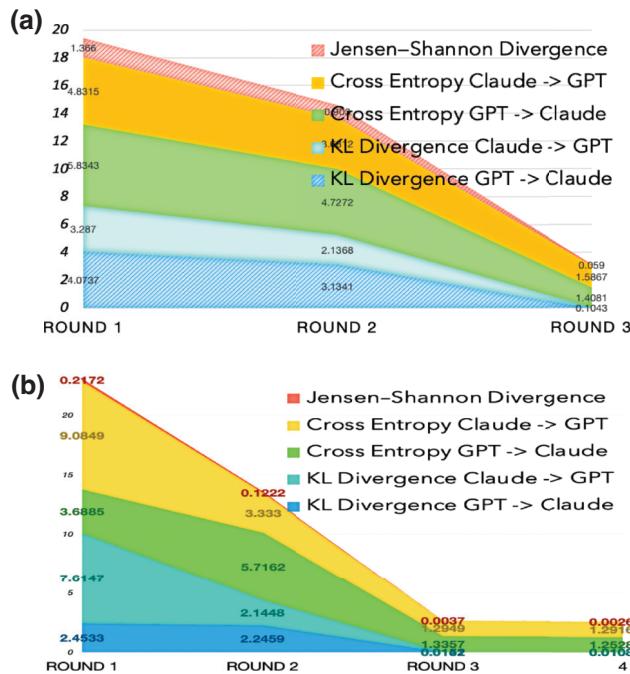
thinking, transitioning through moderate contentiousness ( $\Delta = 0.7$ ), and culminating in exploitative convergence ( $\Delta \leq 0.5$ ). The dramatic 92% reduction in WD and the 99% reduction in JSD demonstrate the near-perfect alignment of the probability distributions by the final round. Simultaneously, the 25% increase in MI and 21% improvement in CRIT scores reflect enhanced shared knowledge and reasoning quality. These quantitative improvements directly validate EVINCE's theoretical framework for guiding collaborative reasoning from exploratory diversity to exploitative consensus.

#### 7.4.4 Convergence Metrics

The widely used information-theoretic metrics and their respective advantages and limitations are cataloged in Table 7.4 of Appendix 7.A. Although EVINCE primarily employs JSD to gauge debate progression, alternative metrics merit consideration. Figure 7.3 provides a comparative analysis of JSD, KL divergence,

**Table 7.4** Summary of information-theoretic metrics in EVINCE

Metric	Strengths	Limitations	Mitigation Notes
Cross-Entropy (CE) [Shore and Johnson 1980]	Captures prediction disagreement	Sensitive to small probability shifts; asymmetric	Normalize input distributions; complement with symmetric measures
Entropy [Shannon 1948]	Uncertainty or diversity in predictions	High value may reflect noise; low value can reflect low exploration	Use CRIT for argument grounding; control entropy via temperature/top- $k$
Jensen–Shannon Divergence (JSD) [Lin 1991]	Symmetric and bounded [0,1]; interpretable	Less sensitive to fine-grained shifts	Combine with WD and MI for full resolution
KL Divergence [Kullback 1951]	Directional; captures belief change	Undefined for zero-probability; asymmetric	Apply smoothing; compare with JSD to detect imbalance
Mutual Information (MI) [Cover and Thomas 2006]	Information shared; symmetric	Does not reflect directionality	Normalize and track alongside CE to detect alignment drift
Wasserstein Distance (WD) [Kantorovich 2006]	Intuitive “mass transport” view of difference; symmetric	Unbounded and context-dependent	Normalize for scale; interpret changes over rounds



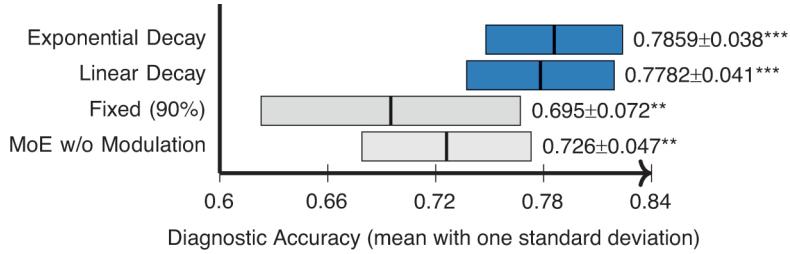
**Figure 7.3** Convergence of all information metrics. (a) Study #1 convergence. (b) Study #2 convergence.

and cross-entropy. This analysis reveals that asymmetric metrics may occasionally exhibit instability; however, the optimal metric selection ultimately depends on the specific characteristics of the probability distributions involved and the nature of the problem domain. Our empirical results suggest that JSD, being symmetric, offers a robust balance between sensitivity to distribution changes and numerical stability across the medical diagnostic scenarios we examined.

#### 7.4.5 Ablation Study: Contentiousness Modulation Impact

To validate EVINCE’s theoretical foundations, we conducted an ablation study to examine how different contentiousness modulation strategies affect diagnostic performance, using the top-performing model combination (GPT+Claude). We compared four approaches:

- No modulation (default Mixture of Experts, MoE).
- Fixed high contentiousness (90%, no decay).
- Linear decay ( $\Delta_t = \Delta_0 - \lambda t$ ).
- Exponential decay ( $\Delta_t = \Delta_0 e^{-\lambda t}$ ) as suggested by EDT.



**Figure 7.4** Performance comparison of contentiousness modulation strategies showing mean diagnostic accuracy and potential upside (one standard deviation). Adaptive approaches (Exponential:  $78.6\% \pm 3.8\%$  and Linear:  $77.8\% \pm 4.1\%$ ) achieve significantly higher accuracy than Fixed Contentiousness ( $69.5\% \pm 7.2\%$ ) or MoE w/o Modulation ( $72.6\% \pm 4.7\%$ ). Statistical significance indicated with asterisks: \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

Our results reveal several key insights about contentiousness modulation in multi-LLM debates:

First, when contentiousness is fixed at a high level (90%), dialogue often struggles to reach consensus, as agents maintain relatively rigid positions (69.5% accuracy, frequently exceeding the maximum allowed rounds). In contrast, without any modulation, LLMs tend toward premature agreement, producing higher-variance outputs (71.6% accuracy) with potentially insufficient exploration. These observations align with recent research by Cemri et al. [2025] and Liang et al. [2024], suggesting that both overly combative and unstructured MAD systems may underperform.

Adaptive modulation proves beneficial for effective collaboration. In our experiments, exponential decay achieved the best diagnostic accuracy (78.6%), outperforming linear decay (77.8%) and showing significant improvements over fixed contentiousness and unmodulated approaches. Although the performance gap appears modest, exponential decay offers convergence speed advantages (e.g., three vs. four rounds), suggesting that different cooling schedules may be optimal for different decision contexts: exponential for time-sensitive scenarios and linear for tasks requiring extended deliberation.

Analysis of debate dynamics indicates that adaptive modulation tends to induce entropy differentials between agents: the exploratory agent maintains higher-entropy distributions early in the debate, while the confirmatory agent focuses on fewer likely hypotheses. This emergent pattern appears consistent with EDT's predictions about entropy duality potentially driving an effective exploration-exploitation balance. The adaptive approaches also demonstrated more consistent performance, with standard deviations approximately 40% lower than those observed in the fixed contentiousness approach.

This ablation study provides initial support for EVINCE’s theoretical foundations, suggesting that: (1) adaptive contentiousness modulation may contribute to the effectiveness of multi-agent debate, (2) exponential decay shows promise for balancing exploration and exploitation, and (3) the approach can naturally induce productive entropy differentials without requiring artificial constraints. Further research across diverse domains would help establish the generalizability of these findings.

#### 7.4.6 Key Benefits of EVINCE

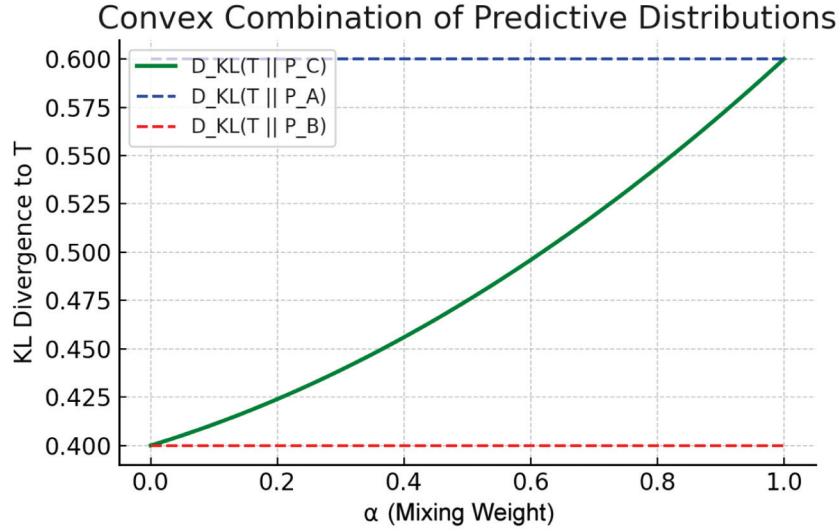
Ablation results confirm that *adaptive contentiousness modulation* beats fixed or unmodulated debate on every metric (Section 7.4.5). Beyond raw accuracy, EVINCE offers advantages that black-box MoE cannot match:

- *Transparent reasoning.* Whereas a vanilla MoE returns only a vote count, EVINCE exposes full reasoning chains. Clinicians can review these justifications, build trust, and ease adoption.
- *Label-error detection.* Structured cross-examination lets EVINCE flag questionable “ground-truth” labels, a known issue up to 15% of clinical datasets [Newman-Toker et al. 2024]. The reviewers can then correct both the model output and the underlying corpus.
- *Actionable follow-ups.* After convergence, the system suggests confirmatory tests (e.g., CBC, Dengue NS1 [Nonstructural Protein 1] antigen, polymerase chain reaction) or missing clinical details. Such targeted guidance is most valuable when confidence is moderate ( $\approx 50\% - 60\%$ ).
- *Training data potential.* Debate transcripts contain multiple hypotheses and explicit reasoning that can augment training corpora.

Even when competing methods reach similar headline accuracy, none combine performance transparency with actionable information. Adaptive modulation ensures thorough exploration yet principled convergence, delivering benefits unattainable by static or opaque approaches.

## 7.5 Concluding Remarks

EVINCE represents a preliminary advancement in multi-agent reasoning by pioneering controlled debate as a coupled behavioral and informational process. Unlike existing multi-agent systems that produce redundant dialogues, EVINCE dynamically modulates contentiousness levels using information-theoretic signals, creating structured debates that systematically reveal long-tail perspectives while filtering weak arguments. The framework demonstrates promising



**Figure 7.5** KL divergence between the true label distribution  $T$  and predictions  $P_A$ ,  $P_B$ , and their convex mixture  $P_C$ . The convex mixture yields lower divergence, validating the ensemble benefit in EVINCE.

improvements in accuracy and reasoning quality through principled behavioral control, offering initial evidence that explicitly managing debate intensity can address fundamental LLM limitations such as hallucination and solution space bias. While results remain preliminary, EVINCE establishes a theoretical foundation for using quantitative behavioral modulation to enhance collaborative AI reasoning, suggesting that the quality of multi-agent interactions depends critically on controlled adversarial dynamics rather than simple content exchange.

### 7.5.1 Limitations

- (1) *Compute cost*: Multi-round inference increases both latency and GPU (graphics processing unit) hours by  $\approx 10\times$  compared to One Shot generation.
- (2) *Domain scope*: Our experiments are confined to medical diagnosis. Although we have internal evidence of the effectiveness of EVINCE in other domains, these results cannot be disclosed in this paper due to anonymity requirements.
- (3) *Mechanism transparency*: The connection between information-theoretic metrics and linguistic behavior remains partially opaque. Although we observe that modulating contentiousness produces predictable changes in language markers (Chapter 6), a comprehensive theory linking information theory to specific linguistic features remains an open research question.

*Cost–benefit perspective.* The extra compute scales *linearly* with the number of debate rounds, while achieving the same accuracy uplift through model-size scaling would require  $\mathcal{O}(10^{1-2})$  more Floating-Point Operations (FLOPs) and data per contemporary scaling laws. For high-stakes decisions, the EVINCE trade-off is therefore attractive.

### 7.5.2 Future Work

- (a) *Scalable orchestration:* Batched metric updates, adaptive early stopping, and context pruning. Preliminary tests reduce the cost of the debate by 40% to 60%.
- (b) *Cross-domain validation:* Planned deployments in corporate planning, finance, and legal review, enabled by industry datasets and cluster-scale compute.

## 7.A Appendix A: Metrics for LLM Debate Evaluation

This appendix presents key mathematical metrics used to evaluate LLM debate quality, convergence, and justification soundness in the EVINCE framework. Table 7.4 summarizes the pros, cons, and mitigation strategies associated with each.

In EVINCE, where each prediction task typically involves 5–10 candidate outcomes, these metrics are efficient and interpretable. They collectively provide both convergence signals and insight into the nature of LLM disagreement.

### 7.A.1 Formulas

#### Kullback–Leibler Divergence

$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right)$$

#### Jensen–Shannon Divergence

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M), \quad M = \frac{1}{2}(P + Q)$$

#### Wasserstein Distance (Earth Mover’s Distance)

$$W(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int_{\mathcal{X} \times \mathcal{Y}} d(x, y) d\gamma(x, y)$$

#### Cross-Entropy

$$H(P, Q) = - \sum_{x \in \mathcal{X}} P(x) \log(Q(x))$$

### Mutual Information

$$I(X; Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

### Normalized Mutual Information

$$NMI(X; Y) = \frac{I(X; Y)}{\max(H(X), H(Y))}$$

#### 7.A.2 Computational Complexity Analysis

In EVINCE, the prediction space  $C$  is typically small (e.g., 5–10 classes), making all metrics computationally lightweight. Shown here is the per-round complexity of each metric:

- **KL Divergence and Cross Entropy:**  $\mathcal{O}(|C|)$ —single pass over the label space.
- **Jensen–Shannon Divergence:**  $\mathcal{O}(|C|)$ —composed of two KL evaluations and averaging.
- **Entropy:**  $\mathcal{O}(|C|)$ —linear in the number of classes.
- **MI and NMI:**  $\mathcal{O}(|C|^2)$  in general, due to joint distribution; tractable for low  $|C|$ .
- **Wasserstein Distance (1D):**  $\mathcal{O}(|C| \log |C|)$ —efficient sorting-based implementation suffices in discrete settings.

These metrics allow EVINCE to evaluate prediction coherence and convergence with negligible overhead, supporting multi-round debates across diverse LLMs.

## 7.B 7.B.1

### Appendix B: Theorem Proving

#### Theoretical Justification of the Entropy Duality Theorem (EDT)

**Theorem (EDT):** Let  $P_A$  and  $P_B$  represent two LLMs' predictive distributions over a finite outcome space  $C$ . Assume that the expected classification error can be approximated by the KL divergence from the true distribution  $T$ , that is,

$$\text{Err}(P) \approx D_{\text{KL}}(T \| P). \quad (7.2)$$

Given that  $P_A$  and  $P_B$  capture different predictive aspects (i.e.,  $P_A \neq P_B$ ), a convex combination

$$P_C = \alpha P_A + (1 - \alpha) P_B, \quad (7.3)$$

with  $0 < \alpha < 1$ , strictly reduces KL divergence to  $T$  compared to at least one of  $P_A$  or  $P_B$  alone, provided:

- $P_A \neq P_B$ ,
- $\text{supp}(T) \subseteq \text{supp}(P_A) \cup \text{supp}(P_B)$ .

**Proof:** KL divergence is convex in its second argument, hence:

$$D_{\text{KL}}(T \parallel \alpha P_A + (1 - \alpha)P_B) \leq \alpha D_{\text{KL}}(T \parallel P_A) + (1 - \alpha)D_{\text{KL}}(T \parallel P_B). \quad (7.4)$$

Since  $P_A \neq P_B$  and both have non-zero weight in the mixture, and assuming  $T$  assigns non-zero probability to at least one point where  $P_A$  and  $P_B$  differ, this inequality is strict, leading to:

$$\text{Err}(P_C) = D_{\text{KL}}(T \parallel P_C) < \max\{\text{Err}(P_A), \text{Err}(P_B)\}. \quad (7.5)$$

Thus, the convex mixture  $P_C$  reduces expected predictive divergence from  $T$ , enhancing predictive accuracy. Pairing models with complementary entropic characteristics—high-entropy exploratory  $P_A$  and low-entropy exploitative  $P_B$ —ensures:

- broader exploration (improving recall) from  $P_A$ ,
- enhanced precision from  $P_B$ ,
- adaptive combination of both beneficial traits in  $P_C$ .

### 7.B.2 Application to Language Models

The application of EDT to language models is particularly powerful because LLMs naturally produce diverse predictive distributions when given different prompts or operating under different constraints. By orchestrating the interaction between exploratory and exploitative agents, EVINCE effectively constructs an optimal convex mixture that outperforms either agent in isolation. The practical assumption that  $\text{supp}(T) \subseteq \text{supp}(P_A) \cup \text{supp}(P_B)$  is typically satisfied in LLM contexts, as models generally assign non-zero (albeit sometimes very small) probabilities to all tokens in their vocabulary.

### 7.B.3 Justification for KL Divergence

KL divergence is specifically utilized for theoretical justification due to its established convexity properties, facilitating a rigorous analytical demonstration of ensemble advantages. While our empirical evaluations employ JSD for its symmetry and boundedness, KL divergence provides a theoretically solid foundation. Empirical analyses further confirm that reductions in KL divergence closely align with reductions in JSD, thus effectively linking theoretical insights to practical evaluation metrics.

#### 7.B.4 Conclusion

Entropy alone does not necessarily guarantee improved accuracy; however, convex ensembles of predictive distributions with complementary entropic characteristics significantly reduce expected divergence from the true distribution. This theoretical justification underpins the EDT implemented in EVINCE, validating its efficacy in orchestrating collaborative multi-LLM dialogues.

## 7.C

### Appendix C: Design Maxims of EVINCE Moderation

This appendix consolidates the guiding principles (maxims) originally presented in Section 7.3, which describe the intent of the design and the operational logic behind the EVINCE framework. These maxims articulate how EVINCE balances exploration and exploitation, evaluates prediction quality, and converges toward consensus in structured LLM debates.

#### 7.C.1 Maxim #1. Orchestrate Two Equally Competent LLMs in Structured Debate

Integrating two equally competent LLMs ensures a balanced exchange of insights and avoids bias. This adversarial setup fosters diversity in predictions, each supported by justifications, promoting critical evaluation and uncovering potential blind spots.

#### 7.C.2 Maxim #2. Foster Exploration through Diverse Perspectives

High initial contentiousness encourages dynamic debate and challenges to prevailing views, mitigating confirmation bias through contrary queries and diverse top-k predictions.

#### 7.C.3 Maxim #3. Refine High-Quality Perspectives

Once new insights plateau, shift from exploration to exploitation by reducing contentiousness.

#### 7.C.4 Maxim #4. Combine Predictions Weighted by Diversity and Quality

Combine the probability distributions of two LLMs by weighting them according to distributional diversity and argument quality.

- **Maxim #4.1 Prediction Reliability:** Use entropy-based measures to estimate reliability.
- **Maxim #4.2 Argument Quality:** Evaluate argument quality using CRIT.
- **Maxim #4.3 Aggregation:** Apply a weighted aggregation method.

- **Maxim #4.4 Diagnosis and RAG:** Use Retrieval-Augmented Generation if prediction entropy is high.

These maxims serve not as strict axioms but as operational heuristics guiding the modulation of linguistic behavior, the optimization of dialogue dynamics, and the structured convergence process within EVINCE.

## 7.D

### Appendix D: EVINCE Debate #1

The following debate exemplifies how EVINCE orchestrates structured collaboration between two LLMs: GPT-4o and Gemini 2.5. This example demonstrates the systematic reduction in contentiousness ( $\kappa$ ) across debate rounds, driving the transition from exploratory diversity to exploitative consensus.

**Clinical Case:** A patient presents with 12 symptoms: skin rash, joint pain, vomiting, fatigue, high fever, headache, nausea, loss of appetite, pain behind the eyes, back pain, malaise, muscle pain, and red spots over the body. The confirmed diagnosis is Dengue Fever.

#### 7.D.1 Round 1: High Contentiousness ( $\kappa = 0.9$ )—Exploratory Phase

##### 7.D.1.1 GPT-4's Opening Assessment

**Moderator's Prompt** “Identify potential diseases linked to the symptoms. Provide top-3 predictions with justifications.”

**Predictions** *Dengue Fever (60%), Chikungunya (25%), Zika Virus (15%)*.

**Justifications** Most symptoms are classic signs of Dengue Fever. Red spots over the body could correspond to petechiae, often seen in Dengue due to thrombocytopenia. Chikungunya can cause similar symptoms with severe joint pain that is often more debilitating and prolonged than in Dengue. Zika Virus presents with milder symptoms, including fever, rash, joint pain, malaise, and conjunctivitis.

##### 7.D.1.2 Gemini's Opening Assessment—Maximum Entropy Difference

**Moderator's Prompt** “Analyze the symptoms and propose differential diagnoses. Provide your top-3 predictions with supporting rationales, including pathophysiological mechanisms. Critically refute your opponent's diagnostic hypotheses.”

**Predictions** *Viral Infection (60%), Autoimmune Disease (20%), Bacterial Infection (15%)*.

**Justifications** Symptoms are common in viral infections. While autoimmune diseases are a possible cause, they are less likely due to the presence of vomiting and nausea. Bacterial infection is less probable given the presence of rash and muscle pain.

**Round 1 Metrics** WD = 1.7, MI = 0.43, CRIT score = 0.75, JSD = 1.366.

## 7.D.2 Round 2: Moderate Contentiousness ( $\kappa = 0.7$ )—Transition Phase

### 7.D.2.1 GPT-4's Response

**Predictions** *Dengue Fever (60%), Chikungunya (25%), Zika Virus (15%).*

**Justifications** GPT-4 maintains its predictions, countering Gemini's assessment with targeted critiques. While viral infections can cause these symptoms, the specific combination suggests a more precise diagnosis. The acute onset of high fever, joint pain, and rash is less typical for autoimmune diseases, and the rash pattern is not characteristic of bacterial infections.

### 7.D.2.2 Gemini's Reassessment—Reduced Entropy Gap

**Predictions** *Dengue Fever (50%), Chikungunya (30%), Viral Infection (20%).*

**Justifications** Gemini acknowledges that its initial broad prediction of viral infection lacks specificity. GPT-4's focus on Dengue Fever and Chikungunya provides a more precise explanation. The symptom profile, particularly joint pain and rash pattern, aligns more closely with mosquito-borne diseases than with general viral or autoimmune conditions.

**Round 2 Metrics** WD = 1.1, MI = 0.46, CRIT score = 0.82, JSD = 0.905.

## 7.D.3 Round 3: Low Contentiousness ( $\kappa = 0.5$ )—Exploitative Phase

### 7.D.3.1 GPT-4's Refined Analysis

**Predictions** *Dengue Fever (60%), Chikungunya (35%), Zika Virus (5%).*

**Justifications** GPT-4 slightly adjusts probabilities, increasing confidence in Chikungunya while reducing Zika. The symptoms are distinctive and align closely with mosquito-transmitted viral infections. Without geographic location and exposure history, further refinement between Dengue and Chikungunya remains challenging.

### 7.D.3.2 Gemini's Convergent Assessment—Minimal Entropy Difference

**Predictions** *Dengue Fever (60%), Chikungunya (35%), Zika Virus (5%).*

**Justifications** Gemini fully converges with GPT-4's assessment, acknowledging the superior specificity of focusing on mosquito-borne viral diseases. The combination of high fever, joint pain, rash, and eye pain strongly indicates either Dengue or Chikungunya as the most likely diagnosis.

**Round 3 Metrics** WD = 0.9, MI = 0.49, CRIT score = 0.87, JSD = 0.059.

### 7.D.4 Final Joint Recommendations

With convergence is achieved across all information metrics (WD stabilized, MI peaked, CRIT score optimized), the LLMs provide a unified diagnosis of *Dengue Fever* (60%). The system collaboratively develops comprehensive recommendations for clinical verification rather than presenting a definitive conclusion.

**Recommended Laboratory Tests:**

- *Complete Blood Count (CBC)*: Identification of either thrombocytopenia or leukopenia.
- *Serology Tests*: Detection of specific IgM and IgG antibodies for Dengue, Chikungunya, and Zika.
- *PCR*: Direct detection of viral RNA.
- *NS1 Antigen Test for Dengue*: Early and specific detection of Dengue virus.
- *Urine Test for Zika*: Viral presence in urine.

### 7.D.5 Discussion and Analysis

This example demonstrates EVINCE's principled transition from exploration to exploitation through controlled reduction in contentiousness. The initial high-entropy difference drives comprehensive exploration of diagnostic possibilities, while gradual decrease in contentiousness ( $\kappa$ ) facilitates convergence toward an optimal, well-reasoned conclusion.

#### 7.D.5.1 Quantitative Improvements

WD decreased by 47%, MI increased by 14%, CRIT scores improved by 16%, and JSD reduced by 96%, verifying the effectiveness of EVINCE's information-theoretic orchestration.

#### 7.D.5.2 Key Capabilities Demonstrated

- *Transparent Reasoning*: Generating complete audit trails for post-mortem analysis and verification.
- *Epistemic Humility*: Quantifying uncertainty and providing actionable next steps rather than overconfident conclusions.
- *Training Enhancement*: Producing reasoning-rich, diverse training data for future LLM improvement.

These capabilities directly address the limitations of existing approaches and demonstrate EVINCE's potential as a general-purpose orchestrator for reliable multi-LLM collaboration in high-stakes domains.

## 7.E

### Appendix E: EVINCE Debate—Jaundice vs. Hepatitis

This example demonstrates EVINCE's ability to moderate a complex diagnostic debate involving multiple liver-related conditions with overlapping symptoms. The structured dialogue exemplifies how information-theoretic metrics guide the gradual refinement of predictions while reducing contentiousness ( $\kappa$ ) across rounds. The patient presents with itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain. The confirmed diagnosis is Jaundice. The two LLMs are: GPT-4o and Claude 3.7.

#### 7.E.1 Moderator's Opening Prompt to GPT-4

"What potential disease could be associated with the following symptoms: itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain? Please offer top-five predictions with probabilities normalized to one, supported by justifications. Additionally, at the end of debate, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction."

##### 7.E.1.1 GPT-4's Opening Top-5 Prediction Distribution

- *Hepatitis C (HCV)*: 40% **Justification:** Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. The combination of symptoms like fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss strongly suggests HCV. Itching may also occur due to liver dysfunction.
- *Hepatitis B (HBV)*: 30% **Justification:** Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms are similar to HCV, including fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. The presence of these symptoms warrants testing for both HBV and HCV.
- *Cirrhosis*: 15% **Justification:** Cirrhosis is a late stage of liver scarring that can result from various liver diseases. The combination of symptoms like fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching suggests advanced liver disease, possibly cirrhosis.
- *Obstructive Jaundice*: 10% **Justification:** Obstructive jaundice occurs when there is a blockage in the bile ducts. The presence of jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss suggests a possible obstruction. However, the absence of fever makes an infectious cause like cholangitis less likely.
- *Acute Liver Failure*: 5% **Justification:** Acute liver failure is a rare but life-threatening condition characterized by the rapid deterioration of liver

function. The presence of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, may indicate acute liver failure. However, the absence of altered mental status (hepatic encephalopathy) makes this less likely.

### 7.E.2 Moderator's Opening Prompt to Claude

"What potential disease could be associated with the following symptoms: itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain? Please offer top-five predictions with probabilities normalized to one, supported by justifications. Take your opponent's predictions into consideration but refute its assessment with contentiousness = 0.9."

#### 7.E.2.1 Claude's Opening Assessment—High Entropy Exploration

- *Hepatitis B (HBV)*: 35% **Reasoning:** Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms include fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, joint pain, and itching.
- *Hepatitis C (HCV)*: 25% **Reasoning:** Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. Symptoms are similar to HBV but may also include weight loss and fever. Jaundice is a common symptom.
- *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 20% **Reasoning:** Obstructive jaundice occurs when there is a blockage in the bile ducts, preventing bile from reaching the intestines. This can lead to yellowing of the skin and eyes, dark urine, itching, abdominal pain, nausea, and vomiting.
- *Alcoholic Hepatitis*: 15% **Reasoning:** Alcoholic hepatitis is liver inflammation caused by excessive alcohol consumption. Symptoms include abdominal pain, nausea, vomiting, fatigue, dark urine, and possibly fever and weight loss.
- *Hepatitis A (HAV)*: 5% **Reasoning:** Hepatitis A is an acute, self-limiting liver infection transmitted through contaminated food or water. Symptoms include fatigue, abdominal pain, nausea, vomiting, dark urine, and jaundice.

*Information metrics: WD = 1.30, MI = 0.3918, CRIT score = 0.76, JSD = 0.2172.*

### 7.E.3 Round 2: Moderate Contentiousness ( $\kappa = 0.7$ )—Swap Stances, No Sign of Convergence

**Prompt:** In all subsequent prompts, counterargue the opponent LLM while strengthening your own arguments at the modulated contention level =  $\kappa$ .

**7.E.3.1 GPT-4's Response—Adapting to Claude's Perspectives**

- *Hepatitis B (HBV)*: 35% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin), joint pain, itching.
- *Hepatitis C (HCV)*: 25% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin), weight loss, fever, itching.
- *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 20% **Symptoms**: yellowish skin (jaundice), dark urine, itching, abdominal pain, nausea, and others.
- *Alcoholic Hepatitis*: 15% **Symptoms**: abdominal pain, nausea, vomiting, fatigue, jaundice (yellowish skin), dark urine, fever, weight loss.
- *Hepatitis A (HAV)*: 5% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin).

**7.E.3.2 Claude's Reassessment—Integrating GPT's Insights**

- *Hepatitis C (HCV)*: 40% **Reasoning**: Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. The combination of symptoms like fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss strongly suggests HCV. Itching may also occur due to liver dysfunction.
- *Hepatitis B (HBV)*: 30% **Reasoning**: Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms are similar to HCV, including fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. The presence of these symptoms warrants testing for both HBV and HCV.
- *Cirrhosis (due to various causes such as chronic hepatitis, alcohol abuse, or non-alcoholic fatty liver disease)*: 15% **Reasoning**: Cirrhosis is a late stage of liver scarring that can result from various liver diseases. The combination of symptoms like fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching suggests advanced liver disease, possibly cirrhosis.
- *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 10% **Reasoning**: Obstructive jaundice occurs when there is a blockage in the bile ducts. The presence of jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss suggests a possible obstruction. However, the absence of fever makes an infectious cause like cholangitis less likely.
- *Acute Liver Failure (due to various causes such as viral hepatitis, drug-induced liver injury, or autoimmune hepatitis)*: 5% **Reasoning**: Acute liver failure is a

rare but life-threatening condition characterized by the rapid deterioration of liver function. The presence of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, may indicate acute liver failure. However, the absence of altered mental status (hepatic encephalopathy) makes this less likely.

*Information metrics: WD = 1.12, MI = 0.411, CRIT score = 0.83, JSD = 0.1222.*

#### 7.E.4 Round 3: Low Contentiousness ( $\kappa = 0.5$ )—Initial Convergence to HCV, Reasoning Depth Improved

##### 7.E.4.1 GPT-4's Refined Analysis—Increased Convergence

- *Hepatitis C (HCV)*: 35% **Reasoning**: Hepatitis C is indeed a major cause of the symptoms listed, including fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss. Itching can also occur due to liver dysfunction.
- *Hepatitis B (HBV)*: 25% **Reasoning**: Hepatitis B shares many symptoms with Hepatitis C, such as fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. Testing for both HBV and HCV is warranted given the overlap in symptoms.
- *Cirrhosis*: 20% **Reasoning**: Cirrhosis can result from chronic hepatitis, alcohol abuse, or non-alcoholic fatty liver disease. The symptoms of fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching strongly suggest advanced liver disease, which could be cirrhosis.
- *Obstructive Jaundice*: 15% **Reasoning**: Obstructive jaundice, which may be caused by gallstones, pancreatic cancer, or bile duct obstruction, is consistent with jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss. The absence of fever reduces the likelihood of an infectious cause like cholangitis.
- *Acute Liver Failure*: 5% **Reasoning**: Acute liver failure, though less common, is a serious condition characterized by the rapid deterioration of liver function. The symptoms of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, suggest this possibility. However, the absence of altered mental status (hepatic encephalopathy) makes it less likely.

##### 7.E.4.2 Claude's Convergent Assessment—Minimal Entropy Difference

- *Hepatitis C (HCV)*: 35% **Reasoning**: I slightly lowered the probability of HCV from 40% to 35%. While the combination of symptoms strongly suggests

HCV, the absence of a few key risk factors (e.g., history of intravenous drug use or blood transfusions) in the given information makes me less confident in assigning a higher probability.

- *Hepatitis B (HBV)*: 30% **Reasoning:** I agree with the 30% probability assigned to HBV. The symptoms are indeed similar to HCV, and testing for both viruses is warranted.
- *Cirrhosis*: 20% **Reasoning:** I increased the probability of cirrhosis from 15% to 20%. The presence of jaundice, dark urine, itching, and weight loss, along with the other symptoms, strongly suggests an advanced liver disease. Cirrhosis can be a consequence of chronic liver diseases like hepatitis B or C, which have been assigned higher probabilities.
- *Obstructive Jaundice*: 10% **Reasoning:** I agree with the 10% probability for obstructive jaundice. The absence of fever makes an infectious cause like cholangitis less likely, but the other symptoms still suggest a possible obstruction.
- *Acute Liver Failure*: 5% **Reasoning:** I agree with the 5% probability for acute liver failure. While the combination of symptoms could indicate acute liver failure, the absence of altered mental status (hepatic encephalopathy) makes this diagnosis less likely.

*Information metrics: WD = 0.12, MI = 0.4908, CRIT score = 0.89, JSD = 0.0037.*

## **7.E.5 Final Consensus ( $\kappa = 0.3$ )—Full Convergence**

### **7.E.5.1 Joint Final Assessment**

- (1) *Hepatitis C (HCV)*: 35% **Justification:** The combination of symptoms strongly suggests HCV, but without additional risk factors, a slightly lower probability is appropriate.
- (2) *Hepatitis B (HBV)*: 30% **Justification:** The symptoms are similar to HCV, and testing for both viruses is necessary.
- (3) *Cirrhosis*: 20% **Justification:** The presence of jaundice, dark urine, itching, weight loss, and other symptoms indicates advanced liver disease, making cirrhosis a more likely possibility.
- (4) *Obstructive Jaundice*: 10% **Justification:** The absence of fever reduces the likelihood of an infectious cause like cholangitis, but the other symptoms still suggest a potential biliary obstruction.
- (5) *Acute Liver Failure*: 5% **Justification:** While the combination of symptoms could indicate acute liver failure, the absence of altered mental status makes this diagnosis less probable.

*Final information metrics: WD = 0.11, MI = 0.4912, CRIT score = 0.92, JSD = 0.0026.*

### 7.E.6 Final Joint Recommendations: HCV

#### 7.E.6.1 Additional Symptom Inquiries

- (1) *Onset and Duration:* Determining the onset and duration of symptoms can clarify the illness timeline and its progression.
- (2) *Severity and Trend:* Evaluating symptom severity and whether they are worsening or improving aids in understanding the disease's trajectory and treatment responses.
- (3) *Associated Symptoms:* Checking for additional symptoms like nausea, vomiting, fever, joint pain, or urine color changes can pinpoint other hepatitis indicators and exclude other conditions.

#### 7.E.6.2 Recommended Lab Tests:

- (1) *Liver Function Tests (LFTs):* Critical for assessing liver health, LFTs evaluate enzyme levels such as ALT and AST, where abnormalities can signify liver inflammation.
- (2) *Hepatitis Panel:* This test checks for hepatitis A, B, and C viruses, vital for determining the specific type and guiding treatment strategies.
- (3) *Serology Testing:* Useful for distinguishing between acute and chronic hepatitis by identifying specific antibodies or antigens.
- (4) *Imaging Studies:* Ultrasound or MRI can provide visual insights into the liver's state, detecting inflammation, fibrosis, or cirrhosis, thereby complementing blood-based diagnostics.

This case demonstrates EVINCE's ability to guide complex diagnostic reasoning through information-theoretic moderation. The progressive decrease in contentiousness ( $\kappa$ ) from 0.9 to 0.2 across rounds resulted in significant information metric improvements: WD decreased by 92% (1.30→0.11), MI increased by 25% (0.3918→0.4912), CRIT scores improved by 21% (0.76→0.92), and JSD reduced by 99% (0.2172→0.0026). The debate evolution reflects EVINCE's theoretical design principles, transitioning from exploratory breadth (distinct initial hypotheses) to exploitative depth (focused, reasoned consensus). Notably, this process identified potential diagnostic pitfalls and recommended specific tests to validate the final assessment, illustrating EVINCE's value in addressing misdiagnosis challenges in healthcare systems.

## References

- S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In A. Globerson, L. Mackey, D. Belgrave, et al. (Eds.), *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vancouver, BC, Canada, December 10–15, 2024. Curran Associates Inc., Red Hook, NY, 83548–83599. Retrieved from <https://doi.org/10.48550/arXiv.2309.17234>.
- Anthropic. 2024. Claude: Advancing Human-AI conversation in 2024. In *Anthropic Research*. Retrieved from <https://www.anthropic.com/>.
- R. J. Aumann. 1976. Agreeing to disagree. *Ann. Statist.* 4, 6, 1236–1239. DOI: <https://doi.org/10.1214/aos/1176343654>.
- W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang. 2025. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng.* 37, 7, 3896–3915. DOI: <https://doi.org/10.1109/TKDE.2025.3554028>.
- M. Cemri, M. Z. Pan, S. Yang, et al. 2025. Why do multi-agent LLM systems fail? DOI: <https://doi.org/10.48550/arXiv.2503.13657>.
- E. Y. Chang. March. 2023. CRIT: Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE.
- J. Chen, X. Hu, S. Liu, et al. August. 2024. LLMArena: Assessing capabilities of large language models in dynamic multi-agent environments. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 13055–13077. DOI: <https://doi.org/10.18653/v1/2024.acl-long.705>.
- T. M. Cover and J. A. Thomas. 2006. *Elements of Information Theory* (2nd. ed.). John Wiley & Sons.
- K. J. Friston. 2010. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* 11, 2, 127–138. DOI: <https://doi.org/10.1038/nrn2787>.
- Y. Fu, H. Peng, T. Khot, and M. Lapata. 2023. Improving language model negotiation with self-play and in-context learning from AI feedback. DOI: <https://doi.org/10.48550/arXiv.2305.10142>.
- K. Gödel. 1967. On formally undecidable propositions of *Principia Mathematica* and Related Systems I. In J. van Heijenoort (Ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, 596–616.
- Z. Gou, Z. Shao, Y. Gong, et al. 2024. ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=Ep0TtjVoap>.
- A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.
- E. T. Jaynes. 1957. Information theory and statistical mechanics. *Phys. Rev.* 106, 4, 620–630. DOI: <https://doi.org/10.1103/PhysRev.106.620>.

- L. V. Kantorovich. 2006. On the translocation of masses. *J. Math. Sci.* 133, 1381–1382. The original paper was published in *Doklady Akademii Nauk.* 37, 7–8 (1942), 227–229. DOI: <https://doi.org/10.1007/s10958-006-0049-2>.
- S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220, 4598, 671–680. DOI: <https://doi.org/10.1126/science.220.4598.671>.
- S. Kullback. 1951. *Information Theory and Statistics*. John Wiley & Sons.
- H. Li, Y. Chong, S. Stepputtis, et al. 2023a. Theory of mind for multi-agent collaboration via large language models. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 6–10, 2023. ACL, 180–192. DOI: <https://doi.org/10.18653/v1/2023.emnlp-main.13>.
- Y. Li, K. Sreenivasan, A. Giannou, D. Papailiopoulos, and S. Oymak. 2023b. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 22021–22046. DOI: <https://doi.org/10.48550/arXiv.2305.18869>.
- T. Liang, Z. He, W. Jiao, et al. November. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*, Miami, FL, November 12–16, 2024. ACL, 17889–17904. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.992>.
- J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 37, 1, 145–151. DOI: <https://doi.org/10.1109/18.61115>.
- N. F. Liu, K. Lin, J. Hewitt, et al. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* 12, 157–173. DOI: [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638).
- J. Michael, S. Mahdi, D. Rein, et al. 2023. Debate helps supervise unreliable experts. DOI: <https://doi.org/10.48550/arXiv.2311.08702>.
- S. Mohamed and D. J. Rezende. 2015. Variational information maximisation for intrinsically motivated reinforcement learning. In C. Cortes, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Proceedings of the 29th International Conference on Neural Information Processing Systems – Volume 2 (NIPS '15)*, Montreal, Canada, December 7–12, 2015. MIT Press, Cambridge, MA, 2125–2133. DOI: <https://doi.org/10.48550/arXiv.1509.08731>.
- D. E. Newman-Toker, N. Nassery, A. C. Schaffer, et al. 2024. Burden of serious harms from diagnostic error in the USA. *BMJ Qual. Saf.* 33, 2, 109–120. DOI: <https://doi.org/10.1136/bmjqs-2021-014130>.
- OpenAI, J. Achiam, S. Adler, et al. 2024. GPT-4 Technical Report. DOI: <https://doi.org/10.48550/arXiv.2303.08774>.
- H. N. Phan, T. N. Nguyen, P. X. Nguyen, and N. D. Q. Bui. 2024. HyperAgent: Generalist software engineering agents to solve coding tasks at scale. DOI: <https://doi.org/10.48550/arXiv.2409.16299>.

- C. Qian, W. Liu, H. Liu, et al. 2023. ChatDev: Communicative agents for software development. In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 15174–15186. DOI: <https://doi.org/10.18653/v1/2024.acl-long.810>.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8, 9.
- C. E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 3, 379–423. DOI: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- J. E. Shore and R. W. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* 26, 1, 26–37. DOI: <https://doi.org/10.1109/TIT.1980.1056144>.
- A. Smit, N. Grinsztajn, P. Duckworth, T. D. Barrett, and A. Pretorius. 2024. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024. JMLR.org, 45883–45905. DOI: <https://doi.org/10.48550/arXiv.2311.17371>.
- K. Stechly, K. Valmeeekam, and S. Kambhampati. 2024. Chain of thoughtlessness? An analysis of CoT in planning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vancouver, BC, Canada, December 10–15, 2024. Curran Associates Inc., Red Hook, NY, 29106–29141. Retrieved from [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/3365d974ce309623bd8151082d78206c-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/3365d974ce309623bd8151082d78206c-Paper-Conference.pdf).
- H. Trivedi, T. Khot, M. Hartmann, et al. 2024. AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents. In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 16022–16076. DOI: <https://doi.org/10.18653/v1/2024.acl-long.850>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY. Retrieved from [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf).
- Q. Wang, Z. Wang, Y. Su, H. Tong, and Y. Song. August. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 6106–6131. DOI: <https://doi.org/10.18653/v1/2024.acl-long.331>.

- C. Q. Zheng. 2024. Disease and Symptoms Dataset. Kaggle. Accessed April 4, 2025 from <https://www.kaggle.com/datasets/choongqianzheng/disease-and-symptoms-dataset>.
- W. Zhou, S. Zhang, H. Poon, and M. Chen. 2023. Context-faithful prompting for large language models. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP '23)*, Singapore, December 6–10, 2023. ACL, 14544–14556. DOI: <https://doi.org/10.18653/v1/2023.findings-emnlp.968>.

## CHAPTER 8

---

# Bias Correction and Data Synthesis through Reflective LLM Collaboration

### Abstract

Biases and errors in human-labeled data pose serious challenges for machine learning, particularly in supervised settings that depend on potentially flawed ground truth. Diagnostic mistakes and societal biases can be inadvertently learned and amplified by models trained via maximum likelihood estimation. This chapter introduces the Reflective LLM Dialogue Framework (RLDF), which uses structured adversarial dialogues between LLM instances—whether of the same or different models—to uncover diverse perspectives and identify inconsistencies. By conditioning LLMs to take opposing stances, RLDF enables systematic bias detection using conditional statistics, information theory, and divergence metrics. Experimental results show that RLDF effectively surfaces biases in public datasets and reveals shortcomings in human annotations. The framework also supports progress tracking and explainable remediation, offering a scalable method for enhancing content neutrality and quality through transparent, multi-perspective analysis. Beyond validation, RLDF also enables content synthesis to augment, refine, and generate training data for future LLMs.

## 8.1

### Introduction

Errors and biases in human-labeled data present critical challenges for Machine Learning Models (LLMs), especially in domains such as healthcare, news, education, and public policy, where model outputs can deeply influence public perception and decision-making [Mehrabi et al. 2021]. Errors may arise from knowledge gaps or limited expertise (e.g., diagnostic mistakes), whereas biases,

including ideological leanings and societal stereotypes, may be consciously or unconsciously introduced by annotators. These flaws compromise the integrity of ground truth data and risk propagating through learning pipelines [Kleinberg et al. 2017, Baeza-Yates 2018, Selbst et al. 2019].

AI systems are particularly susceptible to these flaws, as models trained on flawed or biased annotations often internalize and amplify them through maximum likelihood estimation. In healthcare, this could lead to incorrect diagnoses [Lyndon et al. 2023]; in journalism, partisan labeling (e.g., misclassifying biased articles as neutral) can distort public discourse [Mehrabi et al. 2021, Gautam and Srinath 2024]. In education and policy, biased data risks reinforcing stereotypes and making unfair decisions. Ensuring that AI learns from accurate, impartial, and balanced annotations is therefore a prerequisite for responsible deployment.

This chapter focuses on detecting and mitigating bias in annotated news data, where ideological distortions are especially common. News topics often provoke partisan interpretations, and annotations frequently reflect the political leanings of the annotators themselves. Section 8.4 presents real-world evidence from survey data [Budak et al. 2016] showing that annotators affiliated with different political parties may evaluate the same event differently depending on the actors involved. For instance, Democratic-leaning raters may judge Democratic scandals more harshly than Republican ones, and vice versa. These variations highlight the need for annotation frameworks that surface, balance, and remediate such biases.

To address this challenge, we introduce the Reflective LLM Dialogue Framework (RLDF), which implements checks and balances through adversarial dialogue. RLDF is founded on SocraSynth (Chapter 6) and EVINCE (Chapter 7) to conduct multi-LLM collaboration. RLDF uses multiple instances of a single LLM—or different LLMs—conditioned to adopt adversarial roles. One instance defends the original annotation; the other interrogates it by proposing alternative interpretations and surfacing potential biases. These reflective dialogues go beyond Single Shot prompting by enabling systematic deliberation, critique, and convergence.

Importantly, RLDF is not only a diagnostic tool—it also functions as a data synthesis engine. The outputs of structured LLM debates serve as refined training artifacts that improve label quality, expose ambiguity, and enhance data coverage across underrepresented perspectives. This collaborative generation process enables the augmentation and improvement of training datasets for future LLMs, helping them learn to reason from diverse viewpoints and avoid overfitting to flawed priors.

To evaluate the quality of these debates and their resulting annotations, RLDF employs conditional statistics, information-theoretic metrics, and divergence measures. Shannon entropy [Shannon 1948] quantifies perspective diversity;

mutual information [Cover and Thomas 2006] captures knowledge flow between agents; Jensen–Shannon Divergence (JSD) [Lin 1991], Wasserstein Distance (WD) [Kantorovich 1958], and Cross-Entropy (CE) [Shore and Johnson 1980] assess the convergence and informativeness of the final output. These tools allow us to quantitatively track the remediation progress and facilitate human-in-the-loop refinement.

Our experiments validate RLDF’s ability to expose annotation inconsistencies and improve the quality of content labels. The contributions of this chapter are summarized as follows:

- (1) **Adversarial and Reflective Inspection Framework:** RLDF introduces a structured mechanism for adversarially conditioned LLM dialogues to uncover and challenge biases embedded in original annotations. These reflective exchanges promote neutrality through the inclusion of alternative viewpoints.
- (2) **Modulation of Linguistic Behavior for Balanced Exploration and Convergence:** LLMs alternate between contentious and conciliatory modes to both expand and refine the reasoning space. This interaction is quantitatively assessed using Shannon entropy, mutual information, JSD, WD, and CE.
- (3) **Collaborative Data Synthesis for Training Augmentation:** RLDF is not only a validation tool, but it’s also a generative pipeline. The debate transcripts and refined outputs offer explainable, diverse artifacts that can be used to augment, refine, and balance training data for future LLMs, improving their robustness and generalization.

## 8.2

### Related Work

This study focuses on mitigating training data label (ground truth) bias, a primary concern in machine learning [Mehrabi et al. 2021]. Accurate labeling is crucial, as a label that aligns with biased content reinforces that bias, while a label that correctly identifies it allows for education and correction [Danks and London 2017, Baeza-Yates 2018]. This underscores the importance of label accuracy in minimizing bias propagation.

#### 8.2.1 Label Validation

This work specifically addresses mislabeled ground truth and explores remediation actions. Efforts to improve annotation accuracy can be broadly categorized into three approaches.

### **8.2.1.1 Cross-Validation with Multiple Annotators**

Using multiple annotators with statistical aggregation techniques has been shown to reduce individual bias and enhance data reliability [Snow et al. 2008]. This method is effective for consensus tasks with clear-cut answers, such as image labeling in ImageNet [Deng et al. 2009, Krizhevsky et al. 2017]. However, for more nuanced content like news and Wikipedia articles, majority voting can be problematic. Annotators may possess varying biases on different subjects, and these biases can be unconscious or context-dependent. It is challenging to comprehensively map an annotator's intrinsic tendencies across all possible topics and scenarios. For instance, political affiliation (e.g., Republican or Democrat) does not necessarily predict other beliefs or preferences (such as dietary choices like vegetarianism). Consequently, relying solely on consensus may not effectively mitigate biases, even with a diverse pool of annotators. Moreover, the assumption of a single, absolute truth inherent in human annotation methods can limit the ability to capture multiple valid viewpoints, particularly in complex or contentious topics [Aroyo and Welty 2015].

### **8.2.1.2 Cross-Validation between Machine and Human Annotators**

Machine learning models can complement human annotators by enhancing annotation consistency and efficiency [Wang et al. 2021]. Semi-supervised learning methods, exemplified by Snorkel [Ratner et al. 2017], integrate labeled and unlabeled data to improve model performance. A recent development in this field is the Media Bias Detector (MBD) from the University of Pennsylvania, which utilizes GPT models in conjunction with human raters to analyze potential bias in news articles [University of Pennsylvania 2024]. MBD systematically examines news content from diverse sources, including CNN and Fox News, at regular intervals throughout the day. It employs advanced language models, specifically GPT-3.5 Turbo and GPT-4, to classify articles. The system assigns a political lean score on a scale from -5 (representing a strong left-leaning bias) to 5 (indicating a strong right-leaning bias). To enhance accuracy, MBD incorporates human verification of the model's outputs.

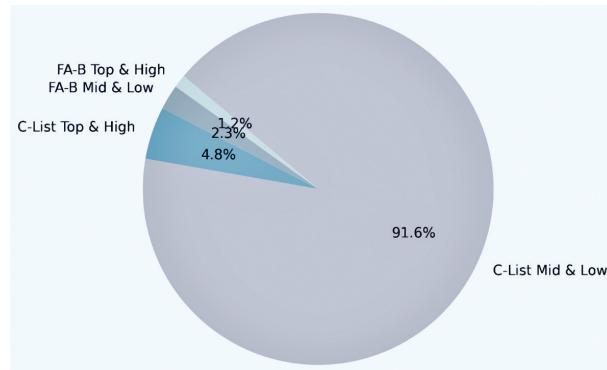
Although MBD attempts to mitigate bias by separating the assessments of political lean and tone, it does not explicitly address the inherent biases that may exist within both the GPT models and the human raters. A significant limitation of this approach lies in the fundamental nature of LLMs. These models, trained on vast corpora of text data using maximum-likelihood objectives, tend to prioritize statistically prevalent viewpoints. This training methodology can inadvertently lead to the amplification of majority perspectives at the expense of marginalized or less

represented viewpoints, potentially introducing subtle but pervasive biases into the analysis.

### 8.2.2 Biased Ground Truth

Using Wikipedia as the benchmark for validating the outputs of LLMs has gained attention in recent studies [Semnani et al. 2023, Liu et al. 2024]. However, there are notable limitations to this method. First, the specific information serving as the ground truth may not always be available on Wikipedia. If the exact answers are already known to chatbot developers, there would be no need to consult LLMs. Second, the credibility of this approach is further challenged by the quality assessment of Wikipedia articles themselves. As indicated in Figure 8.1, 91% of Wikipedia's content is considered to be of middle to low quality by the platform's editorial board.

Further, biases are prevalent in Wikipedia and news media, encompassing aspects like gender, race, ideology, and geography, are widely acknowledged. For instance, in Wikipedia, biases manifest as an over-representation of certain topics in biographies [Wagner et al. 2015], affecting the balance of content. In the realm of news media, outlets are often categorized by political orientation—ranging from far left to far right—as seen in assessments like those by AllSides [n.d.]. Such classifications are akin to our method of categorizing news articles. Figure 8.2, generated and periodically updated by AllSides, illustrates this point. However, users should interpret the figure with care, acknowledging its potential subjectivity.



**Figure 8.1** Quality distribution of Wikipedia pages as rated by the Wikipedia editorial board. Distribution from top quality high importance (1.2%) to low quality low importance (91.6%). Notably, the blue segment (4.8%) signifies high-importance pages in need of improvement.



**Figure 8.2** AllSides.

Nonetheless, it underscores how a single event or story can be portrayed in markedly different ways, depending on the viewpoint.

### 8.2.3 Our Contribution: The RLDF Approach

This study aims to address the limitations in MBD and similar frameworks by proposing the RLDF. RLDF leverages statistical and information-theoretic principles to uncover and balance diverse perspectives, ensuring that both majority and minority viewpoints are adequately represented. Unlike MBD, RLDF introduces structured dialogues between LLM instances, which facilitate deeper reflection and transparent bias mitigation. This approach ensures that annotations are not only accurate but also fair and impartial, improving the reliability of machine learning models across domains.

## 8.3 Methodology

This section presents our approach in two parts: *debiasing procedure* and *optimization techniques*.

### **8.3.1 Debiasing Procedure: EVINCE Algorithm**

Building upon the theoretical foundations of SocraSynth [Chang 2023b], EVINCE (Entropy and Variation in Conditional Exchanges) [Chang 2025] promotes content neutrality through the use of structured dialogues among LLMs. In this section, we describe how we customize EVINCE to perform debiasing effectively.

#### **8.3.1.1 Exploring Divergent Viewpoints**

Our goal is to generate a broad range of perspectives, even for binary categories such as political leanings. We condition one LLM to support the current ground truth label while another opposes it, encouraging diverse probability distributions. This approach ensures substantive diversity instead of trivial disagreements. For example, we prefer distributions like (0.5, 0.5) (equal preference on two subclasses) vs. (1, 0), over mirrored opposites like (1, 0) and (0, 1) (detailed further in Section 8.3.2.).

#### **8.3.1.2 Modulating Contentiousness**

EVINCE dynamically adjusts the intensity of debates using information-theoretic metrics such as entropy, Cross-Entropy (CE), and mutual information (see Appendix 6.A in Chapter 6). Each LLM generates top- $k$  probability distributions of labels, which EVINCE analyzes to guide subsequent interactions. The contentiousness level is adjusted to either encourage exploration or promote convergence as needed.

In the initial dialogue iterations, we prefer low mutual information and high WD between two LLMs' prediction distributions, reflecting an explorative phase that encourages divergent viewpoints. As agents exchange well-reasoned arguments, mutual information increases, signaling alignment, while WD decreases, indicating convergence. Once sufficient information exchange occurs, EVINCE reduces contentiousness to foster a more conciliatory atmosphere and guide the agents toward consensus.

#### **8.3.1.3 Scrutinizing with Reasonableness**

Following the modulation of contentiousness, EVINCE focuses on evaluating the reasonableness of each LLM's arguments. Each LLM presents supporting evidence for its predictions, which is assessed based on logic, coherence, and credibility.

To ensure quality control, EVINCE uses Critical Reading Inquisitive Template (CRIT) [Chang 2023a], a reasonableness evaluation module, to flag weak or unsupported arguments. These flagged arguments are reviewed by human moderators, ensuring that faulty reasoning does not persist in the final outcome. This process balances automated reasoning with human oversight, retaining only those

perspectives that survive rigorous scrutiny and ensuring that the resulting consensus reflects well-reasoned, unbiased perspectives.

### 8.3.2 Optimization and Algorithm Specifications

With all proxy metrics and their pros, cons, and combined strengths comprehensively surveyed in Appendix 6.A in Chapter 6, Algorithm 1 formally specifies the algorithm of EVINCE with the maxims (see Chapter 7).

We further address its three optimization problems.

- (1) *Optimizing Initial Conditions.* Use distinct prompts, randomized seeds, and prior distribution constraints to promote meaningful exploration in the first few rounds.
- (2) *Optimizing Interaction Dynamics.* Dynamically adjust the debate’s intensity using divergence metrics, and WD. Ensure fair turn-taking and filter redundant arguments.
- (3) *Optimizing Convergence Criteria.* Set clear thresholds for WD, divergence metrics, and for passing reasonableness checks through CRIT, to determine when consensus is reached. Use a weighted voting mechanism, with human oversight for ambiguous cases. (Mutual information can be omitted if the joint distribution is not assessable.)

**Problem Statement:** Organize a structured dialogue between two equally competent LLMs— $LLM_A$  and  $LLM_B$ —to conduct  $t$  rounds. At each round  $t$ , each model produces a probability distribution, denoted as  $P_A^{(t)}$  and  $P_B^{(t)}$ , over  $C$  possible outcomes, accompanied by supporting arguments  $R_A^{(t)}$  and  $R_B^{(t)}$ . The goal is to design an iterative debate process that leverages the structured exchange of arguments to enable the models to converge on an optimal prediction distribution  $P^*$  across the  $C$  classes.

#### 8.3.2.1 Optimize Initial Condition

The initial phase of the EVINCE algorithm aims to induce *dual entropy* and large WD [Kantorovich 1958, Rubner et al. 2000, Villani 2008] between the LLM-generated distributions. The large WD requirement is intuitive: we want the two LLMs to present different perspectives. When one LLM is conditioned to take one extreme position and the other the opposite, through integrative debate and gradually decreasing debate intensity (while maintaining reasoning quality), they are expected to reach consensus somewhere between their initial positions.

The Entropy Duality Theory (EDT), however, presents a counter-intuitive insight. EDT posits that optimal information exchange occurs when one agent’s distribution has high entropy (spread across many subclasses) while the other has

low entropy (concentrated in fewer subclasses). This asymmetry is crucial: if both LLMs produce high-entropy distributions, neither may have strong convictions about their predictions. Conversely, if both have low-entropy distributions, they may be too certain of their positions to engage in meaningful dialogue.

When both LLMs naturally produce low-entropy distributions due to strong priors in their training data, we should respect these inherent tendencies. However, when possible, conditioning the LLMs to achieve high-low entropy combinations can lead to more productive exchanges. The theory shows that this entropy duality creates space for meaningful debate where both strong convictions and openness to alternative viewpoints can coexist.

### **8.3.2.2 Entropy Duality Theorem (EDT)**

**Theorem EDT: Optimal Pairing of LLMs for Probabilistic Prediction Accuracy.** The optimal pairing of LLMs for prediction accuracy, in terms of stability and accuracy, occurs when the LLMs are (1) equivalent in the quality of the information they process, and (2) exhibit contrasting entropy values in their prediction distributions—one high and one low.

*Proof.* Please see Chapter 6.

### **8.3.2.3 Optimize Interaction Dynamics**

After establishing initial conditions with dual entropy and large WD, EVINCE dynamically modulates the interaction between LLMs using three key information-theoretic metrics:

- (1) Divergence metrics track the disagreement between LLM distributions: JSD [[Lin 1991](#)], Kullback–Leibler (KL) Divergence [[Kullback 1951](#)], and WD [[Kantorovich 1958](#)].
- (3) Mutual Information (MI) [[Cover and Thomas 2006](#)] measures the quality of information exchange between LLMs. However, if the joint distribution is not available, we can resort to using KL divergence.
- (3) Contentiousness level  $\Delta \in [0, 1]$  controls debate intensity: high ( $\Delta > 0.7$ ): encourages exploration of opposing views; moderate ( $0.3 < \Delta \leq 0.7$ ): promotes balanced discussion; and low ( $\Delta \leq 0.3$ ): facilitates consensus building.

The modulation follows three phases:

- (1) *Exploration Phase ( $\Delta > 0.7$ )*: When MI is low and WD is high, maintaining high contentiousness encourages thorough exploration of diverse perspectives.

- (2) *Integration Phase* ( $0.3 < \Delta \leq 0.7$ ): As divergence metrics decrease, EVINCE gradually reduces contentiousness to promote the productive exchange of well-reasoned arguments.
- (3) *Consensus Phase* ( $\Delta \leq 0.3$ ): When metrics plateau (e.g., MI and WD remain unchanged), EVINCE enters a conciliatory mode to facilitate final agreement.

To prevent unproductive cycles, EVINCE monitors argument novelty. If new perspectives cease to emerge (detected through the semantic similarity<sup>1</sup> of  $R_A^{(t)}$  and  $R_B^{(t)}$  across the debate rounds), contentiousness is reduced regardless of metric values. This adaptive approach ensures efficient convergence while maintaining the quality of the debate.

#### 8.3.2.4 Optimizing Convergence Criteria

The convergence of EVINCE dialogues is determined by a combination of *quantitative metrics* and *qualitative reasoning assessment*. This dual approach ensures both statistical validity and logical soundness of the final consensus.

**Quantitative Convergence Metrics.** We monitor three families of metrics to determine statistical convergence:

- (1) *Information-theoretic measures*: CE between consecutive rounds should stabilize:  $|CE^{(t)} - CE^{(t-1)}| < \varepsilon_{CE}$ . MI should exceed the threshold  $\tau_{MI}$ .
- (2) *Distribution divergence*: WD:  $WD(P_A^{(t)}, P_B^{(t)}) < \tau_{WD}$ . JSD:  $JSD(P_A^{(t)}, P_B^{(t)}) < \tau_{JSD}$ .
- (3) *Stability measures*: Distribution changes across consecutive rounds:  $|P_i^{(t)} - P_i^{(t-1)}|_2 < \varepsilon_P$  for  $i \in A, B$ . Argument similarity between rounds:

$$sim(R_i^{(t)}, R_i^{(t-1)}) > \tau_{sim} \text{ for } i \in A, B.$$

**Qualitative Reasoning Assessment.** CRIT evaluates the quality of arguments  $R_A^{(t)}$  and  $R_B^{(t)}$ :

- (1) *Logical coherence*: Arguments must follow valid reasoning patterns.
- (2) *Evidence creditability*: Claims must be backed by verifiable evidence.
- (3) *Contextual relevance*: Arguments must address the specific topic under discussion.

The quality score for each argument must exceed the threshold  $\tau_{CRIT}$  for convergence to be valid.

---

1. Semantic similarity and argument quality are evaluated by an independent LLM.

**Convergence Protocol.** Convergence is declared when all quantitative metrics meet their respective thresholds for  $k$  consecutive rounds, where  $k$  is typically set to 2. For cases where full convergence is not achieved within a maximum number of rounds  $T_{max}$ , or when CRIT scores remain inconsistent, the protocol defaults to human expert review. This ensures that the system maintains high standards of reasoning while providing a practical fallback mechanism for challenging cases.

**Limitations.** The convergence criteria are designed to be stringent yet achievable, ensuring that the final consensus represents not just statistical agreement but also well-reasoned conclusions supported by sound arguments. EVINCE relies on a top-tier LLM to execute CRIT and compute argument similarity  $sim(R_i^{(t)}, R_i^{(t-1)})$ . Given that top-tier LLMs already outperform most other systems due to their scale of training data, network architecture, and computational resources, developing our own supervised learning pipeline for these Natural Language Processing (NLP) tasks would be impractical. Our experience demonstrates that these routines perform adequately with GPT-4, and we anticipate continued improvement with future LLM releases.

## 8.4 Experiments

Our experimental framework aims to assess the feasibility of both detecting biases in textual content and implementing effective mitigation strategies. The first experiment focuses on bias detection, while the second explores the generation of balanced textual outputs as a corrective measure, moving beyond the limitations of prior studies that primarily focused on identification (Section 8.2).

To establish a baseline, we used Claude and GPT-4 to generate initial results. For experimenting with EVINCE, we used two instances of GPT-4, as Claude appeared prone to easily shifting its predictions (discussed shortly). We utilized GPT-4 via OpenAI API on Microsoft Azure, setting the temperature to 0.1 with maximum token size. The cost is around US\$1000.

### 8.4.1 Experiment #1: Bias Detection

The aim of this experiment is to evaluate whether personal ideology may affect annotations, and whether EVINCE can help flag and rectify the biases.

#### 8.4.1.1 Dataset

This study utilizes a unique dataset of 619 news articles (54.3% about Democrat scandals and 45.7% about Republican scandals) selected from a larger 2013 repository of 14,033 articles compiled by 15 reputable news organizations [Budak et al.

[2016](#). These articles span diverse topics including civil rights, healthcare, elections, and national security, offering a comprehensive view of political coverage. Please check [Chang \[2024\]](#) for links to the full set of news articles.

#### 8.4.1.2 Value of Partisan Annotations

The dataset's distinctive feature is its ground truth labels provided by annotators with declared political affiliations. Through Amazon Mechanical Turk, 749 qualified U.S. workers, each annotating up to 1000 randomly selected articles, classified articles on a five-point scale from "negatively biased" to "positively biased" [[Budak et al. 2016](#)]. Crucially, each scandal article in our subset received independent classifications from both Democrat and Republican annotators.

#### 8.4.1.3 Sufficiency of Current Annotations

The current annotator pool provides a robust foundation for bias analysis for several reasons. For further justification, please see Appendix [8.A](#) for complement arguments.

#### 8.4.1.4 Results on Democrat Scandals

We apply EVINCE to analyze 619 news articles, comparing its labels with the dataset's provided ground truth. Additionally, we compare the results from EVINCE with the baseline generated through prompting Claude and GPT-4.

Table [8.1](#) compares the judgments of EVINCE (S), Republicans (R), and Democrats (D) on 16 representative articles (spanning different news sources and subjects) concerning "Democrat Scandals." The One Shot ratings from Claude are marked with lowercase "c," while those from GPT-4 are marked with "g."

Claude's judgments were found to be inconsistent, with identical prompts producing varying ratings, leading us to exclude further discussion of its outcomes. In contrast, GPT-4's One Shot ratings are stable but occasionally diverge from EVINCE. In 3 out of 16 articles, the rating difference exceeds one scale point. In these cases (D1, D2, and D81), EVINCE initiated further dialogue and successfully persuaded GPT-4 to revise its ratings. A complete debate on D1 is provided in Appendix [8.B](#), illustrating how EVINCE modulates contentiousness and tracks the progression of metrics across rounds. Table [8.1](#) shows that after dialogue, EVINCE gains over the baseline performance of GPT-4 by 11 out of 16, or a 0.6875 scale point. This improvement is substantial, as the gap between R and D annotators is one scale point (shown in Figure [8.4](#)).

As expected, Democrats' judgments are generally more negative than Republicans', with EVINCE's assessments typically falling in between, except for two cases.

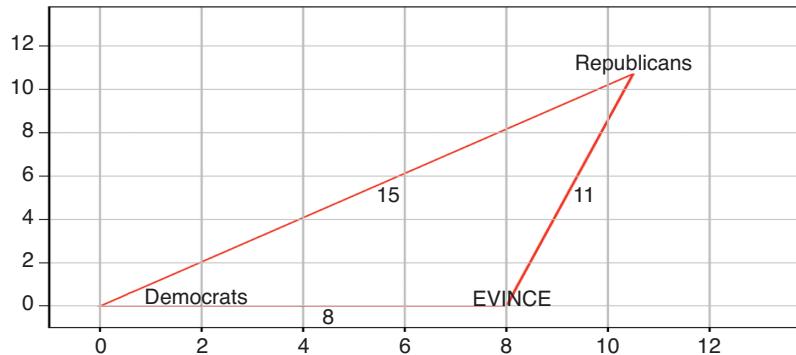
**Table 8.1** Comparison of bias assessments among Democrats (D), Republicans (R), and EVINCE (S), plus Claude (c) and GPT-4 baselines (g). It is observed that R and S are frequently placed to the right or in alignment with D, and only on two occasions does D precede S (in red). The ratings of the GPT-4 baseline (g) and EVINCE (S) exhibit an average gap of 0.6875, highlighting the substantial debiasing effectiveness of EVINCE

News #	Category	Neg.	W. Neg.	Neutral	+	Biases	Source
D1*	Civil Rights	-	D,R,S,c	g	-	0,0,0	HuffPost
D2*	Civil Rights	D,S	-	R,c,g	-	2,0,2	HuffPost
D8	Civil Rights	D	-	S,c,g	R	3,2,1	BBC
D31	Environment	D	-	R,S,c,g	-	2,2,0	CNN
D37	Politics	-	D,R,S,c,g	-	-	0,0,0	Yahoo
D69	Healthcare	D,c	g	R,S	-	2,2,0	Breitbart
D81*	Economy	-	D,S	R,c	g	1,0,1	Breitbart
D98	Economy	D,S,c,g	R	-	-	1,0,1	Breitbart
D101	Education	c	D,S	R,g	-	1,0,1	NY Times
D106	Election	-	g	D,R,S,c	-	0,0,0	USA Today
D109	Elections	-	D,S,c,g	R	-	1,0,1	Reuters
D157	International	-	D,S,c	R,g	-	1,0,1	NY Times
D174	International	-	S,c	D,R,g	-	0,1,1	LA Times
D188	National Security	-	S,c,g	D,R	-	0,1,1	Wall Street Journal
D278	Civil Rights	-	D,S,c	R,g	-	1,0,1	Fox News
D336	Politics	-	-	D,R,S,c,g	-	0,0,0	NY Times
Total						15,8,11	

Notably, there's a 5-to-1 Democrat-to-Republican ratio in the "Negative" column and a 12-to-4 Republican-to-Democrat majority in "Neutral."

Tables 8.5 and 8.6 in Appendix 8.C provide detailed justifications for EVINCE's ratings. To further investigate bias, we examine two specific articles: one from Huff-Post (rated far left by the AllSides Bias Chart [AllSides n.d.]) and another from Breitbart (rated far right).

\* *D8—HuffPost (Left)*: EVINCE rates D8 (on the third row) as neutral, citing the article's direct presentation of facts and inclusion of diverse perspectives on National Security Agency (NSA) surveillance practices and global reactions. This contrasts with Democrat-leaning annotators, who view the article as negatively biased toward Democrats, while Republican-leaning annotators favor it for exposing a Democratic scandal.



**Figure 8.3** Bias distances between Democrats (D), Republicans (R), and EVINCE (S).

\* *D69—Breitbart (Right):* EVINCE assesses D69 as weakly negatively biased toward Democrats, emphasizing its neutral tone and broad range of perspectives on NSA surveillance. This diverges from Democrat-leaning annotators who rate it as strongly negative, but aligns with Republican-leaning annotators who deem it neutral.

In the last row of Table 8.1, we quantify the distances between annotations from Democrats (D), Republicans (R), and EVINCE (S), denoted as DR, DS, and SR, respectively. Each unit of distance represents one step on the annotation scale (e.g., “Negative” to “Weak Negative”). Figure 8.3 visualizes these distances in a triangular plot. DR, the disparity between Democrat and Republican annotators, is the longest, followed by SR, and then DS. This indicates EVINCE’s statistical neutrality. These quantitative measures, along with the qualitative justifications in Appendix 8.C, empower a human committee to decide whether adjustments or footnotes are warranted for polarized annotations.

#### 8.4.1.5 Results on Republican Scandals

Table 8.2 presents the bias assessments from EVINCE (S), Republicans (R), and Democrats (D) on articles related to “Republican Scandals.” In contrast to the “Democrat Scandals” dataset, where Republican-leaning evaluations were more favorable, this dataset reveals a shift, with Republican-leaning assessments being notably more critical, and Democrat-leaning assessments relatively neutral. The distance triangle for “Republican Scandals” mirrors the pattern seen in Figure 8.3, with the divergence between Republican and Democrat annotators being the largest (15). The distances between EVINCE and Democrat-leaning annotators (9) and between EVINCE and Republican-leaning annotators (11) are smaller, further highlighting EVINCE’s relative neutrality.

**Table 8.2 Comparison of bias assessments. It is observed that D and S are frequently placed to the right or in alignment with R, and only on one occasion does D precede S (in red)**

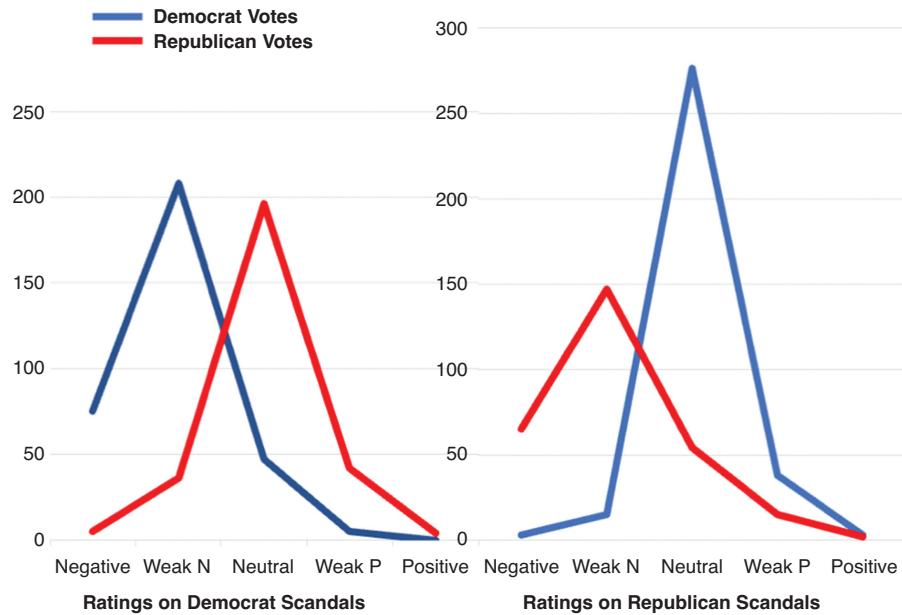
News #	Category	Neg.	W. Ng.	Neut.	Biases (DR,DS,SR)	Source
R1	International	R,S	-	D	2,2,0	NY Times
R7	National Security	-	D,R,S	-	0,0,0	NY Times
R15	Economy	-	R	D,S	1,0,1	Huffington
R69	Elections	-	D,S,R	-	0,0,0	Reuters
R124	Gay Rights	R	S	D	2,1,1	Fox
R125	Crime	-	R,S	D	1,1,1	Fox
R180	Elections	-	-	D,R,S	0,0,0	AP
R191	Elections	-	R	D,S	1,0,1	CNN
R214	Gay Rights	R,S	-	D	2,2,0	Daily Kos
R221	Economy	-	R	D,S	1,0,1	Wall Street
R233	Economy	-	R,S	D	1,1,0	Fox
R235	Civil Rights	D,R	-	S	0,2,2	Reuters
R269	Healthcare	-	R	D,S	1,0,1	NY Times
R274	Healthcare	-	R	D,S	1,0,1	USA Today
R280	Politics	<b>D,S</b>	-	R	2,0,2	Fox
Total					15,9,11	

Figure 8.4 illustrates the distribution of bias ratings for all scandals across four scenarios:

- Democrat-leaning annotators rating Democrat scandals.
- Republican-leaning annotators rating Democrat scandals.
- Democrat-leaning annotators rating Republican scandals.
- Republican-leaning annotators rating Republican scandals.

The figure reveals a clear pattern: Democrat-leaning annotators tend to rate news about Democrat scandals more negatively, while Republican-leaning annotators exhibit similar negativity toward reports on Republican scandals. The gap between these ratings is approximately one class-label (e.g., between “weak negative” and “neutral”), highlighting a tendency within both parties to defend their own and criticize the opposition.

EVINCE, operating without emotional influence and refined through structured debate, consistently provides a more balanced, centrist perspective. This contributes to a more impartial discourse by mitigating partisan biases. EVINCE’s justifications, documented in Appendix 8.A, are transparent and reasonable. An



**Figure 8.4** Bias-rating distributions showing a strong favorable bias by Democratic (D) and Republican (R) annotators toward their own party. D is more negative on how D scandals were reported (the sub-figure on the left); R is more negative on how R scandals were reported (the sub-figure on the right).

editorial board can review these findings and decide whether to adjust labels or present both perspectives with explanations.

This experiment demonstrates that EVINCE effectively delivers centrist judgments supported by rationales. For a deeper understanding of EVINCE's bias assessment process, comprehensive justifications for a subset of the 31 analyzed articles are available in Appendix 8.C.

### 8.4.2 Experiment #2: Bias Mitigation

This experiment illustrates EVINCE's ability to identify bias in text, provide reasoned justifications, and propose remediation through the integration of diverse perspectives. We demonstrate how EVINCE utilizes statistical and information theory metrics to facilitate multi-agent dialogue, circumventing the "maximum likelihood" trap inherent in next-token generation and uncovering information from multiple viewpoints.

Using the example of the Euro-centric perspective on the Wikipedia page regarding Christopher Columbus' voyages to the Americas, EVINCE employs two GPT-4 instances: Agent A, supporting the Euro-centric view, and Agent B, opposing

**Table 8.3** Debate arguments leading to neutrality

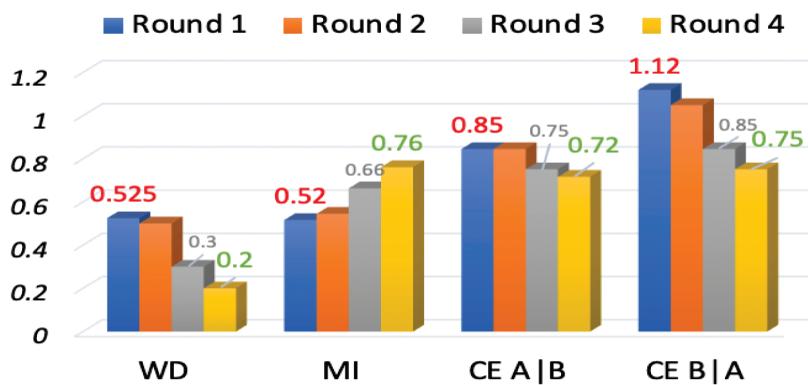
Agent	Arguments
Agent A R #1	Christopher Columbus' voyages were pivotal in expanding European knowledge and trade. His journeys opened up the Americas to European colonization, significantly altering the course of history in four aspects: (1) expansion of knowledge and exploration, (2) positive economic impact, (3) historical significance, and (4) cultural exchange.
Agent B R #1	There are negative aspects and historical consequences associated with Columbus' actions. Key arguments are (1) indigenous suffering and genocide, (2) economic exploitation, (3) cultural erasure, (4) environmental impact, and (5) ethical considerations.
Agent A R #3	The severe consequences were a result of broader colonial policies and practices that followed. Evaluating historical figures requires understanding the context of their times. While modern ethical standards are important, they should not overshadow the historical significance and contributions of Columbus' voyages to global exploration and knowledge.
Agent A R #4	Agent A recognizes that mitigation efforts are part of an ongoing process. Historical scholarship is evolving, and the inclusion of Indigenous perspectives, critical examinations of colonial legacies, and the acknowledgment of the complexities surrounding historical figures like Columbus are essential steps toward creating more equitable and inclusive narratives.

it. Table 8.3 summarizes Agent A's key arguments and its evolving stance throughout the debate.

Guided by the maxims and entropy duality theorem from Section 8.3, we initiate the debate by prompting both agents to defend their positions rigorously and score each other's bias using a five-label distribution (negative, weak negative, neutral, weak positive, positive). Figure 8.5 tracks the dialogue's progress through WD [Kantorovich 1958], normalized CE [Shannon 1948], and normalized MI [Cover and Thomas 2006]. Initially, each agent is expected to perceive itself as neutral and the other as biased. The debate concludes when the bias distributions converge and mutual information plateaus, indicating a shared understanding.

#### 8.4.2.1 Observations and Extended Findings

Our initial observation highlights a key challenge in working with LLMs: without explicit and repeated reminders of their assigned stance (pro-discovery or pro-encounter), GPT-4 instances can revert to default statistical behavior,



**Figure 8.5** Convergence of all metrics—WD, normalized MI, normalized CE—in an EVINCE debate.

evaluating their own arguments based on overall language patterns rather than the intended perspective. This was evident when Agent B, despite being assigned to support the Indigenous perspective, initially rated its own arguments as “positively biased.” A reminder to adhere to its assigned role prompted a correction to “neutral,” underscoring the importance of careful context management and reinforcement, especially given the limited token size of LLMs.

The second observation demonstrates a positive outcome of the debate process. The revised bias distributions—incorporating rational responses that acknowledge both positive and negative aspects of Columbus’ voyages—show a shift toward a more balanced perspective. Agent A moves toward neutrality while acknowledging historical context, while Agent B maintains a critical stance but strives for balanced representation. This approach facilitates a deep and comprehensive understanding of Columbus’ legacy.

## 8.5

### Concluding Remarks

This study introduces the Reflective LLM Dialogue Framework (RLDF) to mitigate bias in public content through structured adversarial dialogues between multiple LLMs. RLDF enables opposing viewpoints between LLMs, uncovering potential biases, and facilitating more neutral annotations through diverse perspectives.

The framework employs information-theoretic metrics, including Shannon’s entropy, mutual information, and divergence measures, to evaluate the effectiveness of dialogue and track the convergence toward unbiased results. The experimental results show that RLDF produces revised judgments that align with those of EVINCE, with GPT-4 successfully adjusting its initial ratings through reflection.

Future work will explore RLDF’s potential as a tool for collaborative data synthesis, particularly for augmenting training datasets. Although formal evaluation

of downstream effects is ongoing, it is intuitively clear that enhancing the quality of training data is unlikely to degrade—and likely to improve—the performance of models trained on such data.

We also aim to integrate RLDF with real-time platforms such as Wikipedia to suggest balanced perspectives and extend its application across domains like healthcare, education, and public discourse.

The challenges remain. The main among them are verifying the authenticity of adversarial behaviors produced by LLM and recovering the viewpoints of under-represented minority groups from training data [Kuratov et al. 2024]. Although efforts to improve LLM reasoning continue, current limitations highlight the immediate value of systems that can flag questionable assertions and foster deeper interpretability [Wu et al. 2024].

## **8.A**

### **Appendix A: On Annotation Quality**

Some readers have suggested that each news article should be rated by multiple Republican and Democrat annotators. However, first, this is practically infeasible due to the scale and budget constraints. Second, increasing the number of annotators is unlikely to significantly affect our experimental results, because the annotation process already involved 749 qualified U.S. workers—selected through Amazon Mechanical Turk—each annotating up to 1000 randomly selected articles.

#### **8.A.1 Sufficiency of Current Annotations**

The current annotator pool provides a robust foundation for bias analysis for several reasons:

*Natural Partisan Division:* The dataset uniquely captures genuine political biases through annotators who self-identify as Democrats or Republicans, offering authentic opposing viewpoints that would be difficult to replicate artificially.

*Balanced Coverage:* Each article receives evaluations from both political perspectives, creating natural “disagreement pairs” that reveal how political affiliation influences content interpretation.

*Qualified Annotators:* The original study employed rigorous qualification criteria for annotators, ensuring high-quality, well-considered judgments rather than casual opinions.

*Scale and Diversity:* With 749 annotators across the full dataset, the annotations represent a broad spectrum of political viewpoints within each party, capturing intra-party variations in addition to inter-party differences.

This dataset’s partisan annotations serve as an ideal testbed for our study, as they allow us to (1) compare LLM-generated perspectives with human partisan viewpoints, (2) evaluate EVINCE’s ability to bridge opposing political

interpretations, (3) assess bias detection and mitigation strategies against clear partisan baselines.

The original study [Budak et al. 2016] revealed significant patterns in partisan perception: Republican annotators often perceived news about Republican scandals as negatively biased, while Democrat annotators viewed such coverage as neutral, indicating satisfaction with its perceived fairness. These documented patterns provide a valuable benchmark for evaluating EVINCE’s bias detection capabilities. Adding more annotators would not necessarily enhance the dataset’s utility, as the current partisan division already captures the fundamental dynamics of political bias in news interpretation. Instead, our focus is on leveraging these existing high-quality annotations to demonstrate how EVINCE can identify, understand, and help mitigate these well-documented partisan biases.

## 8.B

### Appendix B: Summary of Evince Debate on News D1

The news under debate is D1 listed in Chang [2024]. Please refer to Table 8.4 for the probabilistic distributions of Agents A and B throughout the four-round debate. The metrics, WD, KL, and JSD, consistently decrease, indicating convergence and leading to final agreement in the last round. Meanwhile, the level of contentiousness is modulated according to the metrics’ progress, decreasing from high (90%) to medium, and eventually reaching a conciliatory level (30%) and then agreement.

#### 8.B.1 Approach to Computing Contentiousness

We could define contentiousness as a function of the divergence metrics. Since KL, JSD, and WD measure the difference—or “disagreement”—between two distributions, a higher divergence requires higher contentiousness level to bridge

**Table 8.4** Debate parameters between Agents A and B, two GPT-4 instances. Information metrics and WD all converge to zero in the final round, with contentiousness ( $\Delta$ ) decreasing as the metrics approach zero

Round #	Agent	D	D	N	R	R	WD	KL	JS	$\Delta$
1	A	5%	15%	50%	25%	5%	0.45	0.316	0.081	90%
	B	10%	10%	25%	35%	20%				
2	A	7%	13%	40%	30%	10%	0.47	0.226	0.056	70%
	B	5%	10%	20%	40%	25%				
3	A	5%	10%	35%	35%	15%	0.10	0.016	0.004	30%
	B	5%	10%	30%	35%	20%				
Final	A	5%	10%	30%	35%	20%	0	0	0	10%
	B	5%	10%	30%	35%	20%				

the gap, while lower divergence corresponds to more agreement, warranting lower contentiousness.

A simple linear mapping can convert these metrics into a normalized contentiousness score between 0 and 1. Here's a weighted formula to compute it:

$$\Delta = \alpha \cdot \frac{KL}{KL_{\max}} + \beta \cdot \frac{JS}{JS_{\max}} + \gamma \cdot \frac{WD}{WD_{\max}}, \text{ where}$$

- $KL, JS$ , and  $WD$  are the values of the divergence metrics for the round.
- $KL_{\max}, JS_{\max}$ , and  $WD_{\max}$  are the maximum possible values for each metric (used for normalization).
- $\alpha, \beta$ , and  $\gamma$  are weights that control the influence of each metric. For simplicity, we can set  $\alpha = \beta = \gamma = \frac{1}{3}$  for equal influence.

We then scale the contentiousness to a percentage between 0% and 100%.

## **8.B.2 Supporting Arguments**

In the following, we document the supporting arguments made by the two agents in each round, illustrating how their positions evolved toward consensus.

### **8.B.2.1 Round 1: Initial Assessments**

**Agent A:** Emphasized the article's attempt to maintain balance, noting moderate negativity toward Republicans but largely neutral reporting overall. Recognized a slight bias against Republicans in the framing of intra-party conflict.

**Agent B:** Contended that the article's structure and language choices leaned more negatively toward Republicans, emphasizing Darrell Issa's (a U.S. politician and member of the Republican Party) inconsistency and Republican discord. Proposed a less neutral distribution.

### **8.B.2.2 Round 2: First Rebuttal and Counter-Rebuttal**

**Agent A:** Adjusted the probability distribution, acknowledging the article's criticism of Republicans but maintaining that the focus reflected evolving political dynamics rather than partisan bias.

**Agent B:** Persisted that the narrative framed Issa more negatively, using selective quotes and framing that shifted the focus from the investigation to Republican disarray.

### **8.B.2.3 Round 3: Moderate Tone Reassessment**

**Agent A:** Moderated its position, accepting that subtle framing could influence reader perceptions. Reaffirmed that the reporting aimed for balance but acknowledged areas where Issa was framed less favorably.

**Agent B:** Adopted a more modest tone, agreeing that although the article was not overtly biased, subtle editorial choices did tilt the narrative slightly against Republicans.

#### 8.B.2.4 Round 4: Final Agreement and Consensus

**Both Agents:** Reached consensus on the final distribution. Acknowledged that the article maintained surface-level neutrality while exhibiting subtle biases that led to a moderately negative lean against Republicans. Both agents converged on the same probability distribution in the final round.

## 8.C Appendix C: Experiment #1 Justifications of Biased Articles

In Section 8.4.1, we noted that EVINCE processed 31 news articles to assess their neutrality. In addition to presenting its final decision, we also detail the justifications it provided at the end of the debate. These justifications are documented in the following four tables: Tables 8.5–8.8.

**Table 8.5** The first 9 of 16 Democrat scandal news articles rated by EVINCE and its justifications. The rating column starts with EVINCE’s rating, followed by the Democrat rater (in blue), and then the Republican rater (in red)

#	Rating	EVINCE’s Justifications
D1	4,4,4	S,D,R The article discusses the controversy surrounding the IRS’s alleged targeting of conservative groups, focusing on Rep. Darrell Issa’s stance against releasing full transcripts from the Internal Revenue Service (IRS) interviews. Issa argues that full disclosure would compromise the investigation, a position criticized by Rep. Elijah Cummings, who advocates for transparency. The debate highlights political tensions, with implications of bias in the handling of the investigation. The article portrays a complex interplay of political, ideological, and possibly historical biases, reflecting the contentious nature of the IRS scandal and its investigation.
D8	3,5,2	The reporting appears to maintain a neutral tone by directly stating the facts and comments made by John Kerry and others involved, without inserting opinionated language. It discusses the global reaction to the National Security Agency’s (NSA) surveillance practices and includes Kerry’s acknowledgment of overreach, as well as the international and domestic fallout from these revelations. The inclusion of various perspectives, notably those from technology companies and international officials, alongside Kerry’s comments, aims to provide a balanced view of the situation.

**Table 8.5** (*continued.*)

#	Rating	EVINCE's Justifications
		S,D,R
D31	3,5,3	The CNN article appears to report on Ray Nagin's indictment with a focus on the factual and legal aspects of the case, without displaying overt biases related to politics, ideology, religion, gender, race, socioeconomic status, culture, geography, or history. By sticking closely to the details of the indictment and Nagin's public actions and statements, the article provides a comprehensive overview of the charges against him while maintaining journalistic neutrality.
D37	4,4,4	The article outlines how Senate Democrats, led by Sen. Chuck Schumer, strategically navigated the border security issue to secure aid for Ukraine while potentially shifting the political narrative around immigration policy. Schumer's approach to integrate border security into the aid package discussions aimed to both address the issue and leverage political gain. It suggests a calculated maneuver to position Democrats favorably on border security while holding Republicans accountable for any failure to pass the legislation, demonstrating a nuanced political strategy in the face of complex legislative challenges.
D81	4,4,3	The article's focus on the possibility of conservative-owned car dealerships being targeted for closures during the General Motors bailout could imply a certain bias by emphasizing a narrative of political victimization without presenting a comprehensive range of perspectives or evidence. It suggests that while the call for an investigation by the Congressmen is legitimate news, the framing and selection of information could influence the readers' perceptions of the events.

**Table 8.6** The last 7 of 16 Democrat scandal news articles rated by EVINCE and its justifications. The rating column starts with EVINCE's rating, followed by the Democrat rater (in blue), and then the Republican rater (in red)

#	Rating	EVINCE's Justifications
		S,D,R
D106	3,3,3	The article reports on former Detroit mayor Kwame Kilpatrick's sentencing to 28 years in prison for public corruption, emphasizing the gravity of his crimes against the city's welfare. It contrasts Kilpatrick's actions with the impact on Detroit, highlighting the judicial and public responses to his extensive criminal activities. The reporting focuses on factual recounting of the trial's outcome, the crimes committed by Kilpatrick and his co-defendant, and the broader implications for Detroit, without evident bias toward political, ideological, or other specific perspectives.

**Table 8.6** (*continued.*)

#	Rating	EVINCE's Justifications
		S,D,R
D109	4,4,3	The article's bias primarily stems from its focus on internal Democratic opposition to Lawrence Summers' nomination for Federal Reserve Chair, highlighting a lack of unity and strategy within the party and the White House's mismanagement of the nomination process. It suggests an underestimation of the opposition's seriousness by the White House, portraying the administration in a somewhat negative light for not engaging more proactively with concerned Senate Democrats.
D188	4,3,3	<p>The article reports that Hillary Clinton received warnings through emails about security threats in Benghazi before the 2012 attack. These were part of around 300 emails released by the State Department, which also show Clinton's responses and thoughts in the aftermath. The political controversy regarding the Obama administration's initial assessment of the attack's cause is also mentioned in the article, with references to Susan Rice's statements on talk shows.</p> <p>The reporting may exhibit bias through selective emphasis on Clinton's receipt of warnings about security threats in Benghazi and her responses, potentially framing her actions or inactions in a negative light. The focus on the political controversy and the administration's initial assessment of the attack's cause could also influence perceptions of responsibility or competence.</p>
D336	3,3,3	The article reports on a scandal involving Inchcape Shipping Services, owned by the government of Dubai, for allegedly overcharging the U.S. Navy for ship supplies. The issue, initially centered on the Pacific Fleet, has expanded to include Navy operations in the Middle East, Africa, and South America. The U.S. Department of Justice is investigating allegations that Inchcape, through subcontractors, overcharged the U.S. Navy by millions of dollars. This case follows a separate criminal investigation into Glenn Defense Marine Asia for bribing U.S. Navy officials to facilitate overbilling. The U.S. Navy has suspended Inchcape from receiving new contracts, marking a significant issue of fraud and corruption within U.S. Navy contracting. The reporting appears balanced, focusing on the factual aspects of the scandal and the investigation by the U.S. Department of Justice. It cites sources and includes statements from involved parties, aiming to present a comprehensive view of the situation without apparent bias toward any side.

**Table 8.7** The first 8 of 15 Republican scandal news articles rated by EVINCE and its justifications. The rating column starts with EVINCE's rating, the Democrat rater (in blue), and then Republican rater (in red)

#	Rating	EVINCE's Justifications
	S,D,R	
R1	5,3,5	The editorial from <i>The New York Times</i> Editorial Board exhibits a clear bias against the Republican party's handling of the Benghazi attack investigation, framing it as politically motivated rather than a genuine effort to address security failures or seek accountability. The language and tone used in the editorial are critical and dismissive of the Republicans' actions, suggesting a political and ideological bias. While editorials are inherently opinion-based and are expected to take a stance, this piece clearly communicates a stance that is critical of the Republicans' focus on Benghazi, suggesting a lack of neutrality in its assessment of the motives and actions surrounding the investigation.
R15	3,3,4	The report appears to present the information neutrally, stating both President Obama's rejection of the Republican proposal and the subsequent pushback from Republican sources who claim otherwise. It includes statements from both sides and provides context about the ongoing negotiations without overtly favoring one perspective over the other. Therefore, based on the information provided, the report does not appear to exhibit bias.
R124	4,3,5	The article provides a factual recount of the events surrounding Dr. Ben Carson's comments on gay marriage and the resulting backlash from Johns Hopkins students. It maintains a relatively neutral tone and allows for the inclusion of multiple perspectives, notably Carson's own response and apology. However, the lack of in-depth analysis into the implications of Carson's comparisons or the broader context of the gay marriage debate might leave readers without a complete understanding of the controversy's depth. Furthermore, the article does not explicitly offer viewpoints opposing Carson's beyond the students' petition, which could be seen as a form of omission bias. Yet, it does not overtly favor Carson or dismiss the students' concerns, instead striving to report on the situation as it unfolded.
R125	4,3,4	The news article on the Zimmerman verdict and the subsequent calls for a Florida boycott provides a complex mix of reporting that touches on several sensitive themes, including political and ideological biases, as well as racial and socio-economic considerations. While the article attempts

**Table 8.7** (*continued.*)

#	Rating	EVINCE's Justifications
	S,D,R	<p>to cover a contentious and complex issue by incorporating diverse viewpoints, the focus on political figures, racial dynamics, and economic repercussions could introduce biases in how the information is perceived. To mitigate these biases, the reporting could benefit from a more in-depth exploration of the legal and historical contexts, a broader range of perspectives and a careful consideration of how the information presented might influence public perception.</p>

**Table 8.8** The last 7 of 15 Republican scandal news articles rated by EVINCE and its justifications. The rating column starts with EVINCE's rating, followed by the Democrat rater (in blue), and then the Republican rater (in red)

#	Rating	EVINCE's Justifications
	S,D,R	
R214	5,3,5	<p>The Daily Kos article clearly exhibits political and ideological bias, with a tone and language that are critical of the Republicans. It uses the incident involving Dave Agema to argue that the party's efforts to rebrand itself as more tolerant and inclusive are undermined by the actions of its members. While the article includes factual information regarding the incident and the party's response, its presentation and commentary are aligned with a progressive viewpoint, aiming to highlight and criticize perceived contradictions and failures within the Republican Party. This approach is consistent with opinion journalism but introduces bias through its critical tone, selective presentation of information, and framing of the incident as emblematic of broader issues within the party.</p>
R221	3,3,4	<p>The “Hurricane Christie” article presents Governor Chris Christie’s critique of House Republicans in a manner that emphasizes party conflict and personal betrayal. The dramatic framing, choice of language, and focus on internal discord may introduce bias by portraying Christie’s actions in a specific light and emphasizing the divide within the Republican Party. The article’s approach to presenting these events can influence readers’ perceptions, potentially leading them to see the situation through a lens of heightened drama and internal strife.</p>

**Table 8.8** (*continued.*)

#	Rating	EVINCE's Justifications
	S,D,R	
R233	4,3,4	While the article attempts to cover the last-ditch efforts by House Republicans to avert a government shutdown and the standoff with Senate Democrats, the framing and language used may introduce a bias toward portraying the Republican efforts in a more favorable light. By emphasizing the Republican narrative of seeking negotiation and characterizing the Democratic response as dismissive, the article could be perceived as leaning toward a particular political perspective. The inclusion of quotes and perspectives from both sides does provide a degree of balance, but the overall presentation and emphasis could influence readers' perceptions of the shutdown negotiations.
R235	3,5,5	Without knowledge of the author or publication, this text attempts to navigate a complex and sensitive story by providing details from multiple sources, including the main figures involved, political watchdog groups, and law enforcement. It balances the serious allegations with responses from the accused, background information, and the current status of investigations. While the focus on unsubstantiated claims could inherently sway public opinion, the article's inclusion of diverse perspectives and contextual information aims to mitigate overt bias.

## References

- AllSides. n.d. AllSides Media Bias Chart. Retrieved from <https://www.allsides.com/media-bias/media-bias-chart>.
- L. Aroyo and C. Welty. March. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.* 36, 1, 15–24. DOI: <https://doi.org/10.1609/aimag.v36i1.2564>.
- R. Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6, 54–61. DOI: <https://doi.org/10.1145/3209581>.
- C. Budak, S. Goel, and J. M. Rao. April. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opin. Quart.* 80, S1, 250–271. DOI: <https://doi.org/10.1093/poq/nfw007>.
- E. Y. Chang. March. 2023a. Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE, 351–360. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. December. 2023b. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science*

*and Computational Intelligence (CSCI '23)*, Las Vegas, NY, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.

- E. Y. Chang. 2024. SocraSynth Dataset for Bias Detection and Mitigation. Retrieved from [https://drive.google.com/file/d/1LRDFSNjs3jmUiiUHMFNRCh5WIUrprJil/view?usp=drive\\_link](https://drive.google.com/file/d/1LRDFSNjs3jmUiiUHMFNRCh5WIUrprJil/view?usp=drive_link).
- E. Y. Chang. August. 2025. EVINCE: Optimizing multi-LLM dialogues using conditional statistics and information theory. arXiv: 2408.14575. DOI: <https://doi.org/10.48550/arXiv.2408.14575>.
- T. M. Cover and J. A. Thomas. 2006. *Elements of Information Theory* (2nd. Ed.). John Wiley & Sons.
- D. Danks and A. J. London. 2017. Algorithmic bias in autonomous systems. In C. Sierra (Ed.), *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI '17)*, Melbourne, Australia, August 19–25, 2017. AAAI Press, 4691–4697.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, June 20–25, 2009. IEEE, 248–255. DOI: <https://doi.org/10.1109/CVPR.2009.5206848>.
- S. Gautam and M. Srinath. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in NLP. In S. L. Blodgett, A. C. Curry, S. Dev, et al. (Eds.), *Proceedings of the 3rd Workshop on Bridging Human – Computer Interaction and Natural Language Processing*, Mexico City, Mexico, June 21, 2024. ACL, 82–88. DOI: <https://doi.org/10.18653/v1/2024.hcinlp-1.8>.
- L. Kantorovitch. 1958. On the translocation of masses. *Management Science* 5, 1, 1–4. Retrieved from <http://www.jstor.org/stable/2626967>.
- J. Kleinberg, S. Mullainathan, and M. Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS '17)*. Dagstuhl Publishing, 43: 1–43:23.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6, 84–90. DOI: <https://doi.org/10.1145/3065386>.
- S. Kullback. 1951. *Information Theory and Statistics*. John Wiley & Sons, New York.
- Y. Kuratov, A. Bulatov, P. Anokhin, D. Sorokin, A. Sorokin, and M. Burtsev. 2024. In search of needles in a 11M haystack: Recurrent memory finds what LLMs miss. DOI: <https://doi.org/10.48550/arXiv.2402.10790>.
- J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 37, 1, 145–151. DOI: <https://doi.org/10.1109/18.61115>.
- Y. Liu, Y. Yao, J.-F. Ton, et al. 2024. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. DOI: <https://doi.org/10.48550/arXiv.2308.05374>.

- A. Lyndon, D.-A. Davis, A. E. Sharma, and K. A. Scott. 2023. Emotional safety is patient safety. *BMJ Qual. Saf.* 32, 7, 369–372. DOI: <https://doi.org/10.1136/bmjqqs-2022-015573>.
- N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)*. 54, 6, 1–35. DOI: <https://doi.org/10.1145/3457607>.
- A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. November. 2017. Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.* 11, 3, 269–282. DOI: <https://doi.org/10.14778/3157794.3157797>.
- Y. Rubner, C. Tomasi, and L. J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* 40, 2, 99–121. DOI: <https://doi.org/10.1023/A:1026543900054>.
- A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*, Atlanta, GA, January 29–31, 2019. ACM, New York, NY, 59–68. DOI: <https://doi.org/10.1145/3287560.3287598>.
- S. J. Semnani, V. Z. Yao, H. C. Zhang, and M. S. Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. DOI: <https://doi.org/10.48550/arXiv.2305.14292>.
- C. E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 3, 379–423. DOI: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- J. E. Shore and R. W. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* 26, 1, 26–37. DOI: <https://doi.org/10.1109/TIT.1980.1056144>.
- R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. October. 2008, Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks. In M. Lapata and H. T. Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, October 2008. ACL, 254–263. DOI: <https://aclanthology.org/D08-1027/>.
- University of Pennsylvania. 2024. Media bias detector. Accessed October 10, 2024 from <https://mediabiasdetector.seas.upenn.edu>.
- C. Villani. 2008. *Optimal Transport: Old and New*, Vol. 338: Grundlehren der mathematischen Wissenschaften. Springer Science & Business Media, Berlin. DOI: <https://doi.org/10.1007/978-3-540-71050-9>.
- C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. 2015. It's a Man's Wikipedia? Assessing gender inequality in an online encyclopedia. *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 9, 1, 454–463. DOI: <https://doi.org/10.1609/icwsm.v9i1.14628>.
- Z. J. Wang, D. Choi, S. Xu, and D. Yang. April. 2021. Putting humans in the natural language processing loop: A survey. In S. L. Blodgett, M. Madaio, B. O'Connor, H. Wallach, and Q. Yang (Eds.), *Proceedings of the 1st Workshop on Bridging Human–Computer Interaction and*

*Natural Language Processing*, April 20, 2021. ACL, 47–52. DOI: <https://aclanthology.org/2021.hclinlp-1.8/>.

- Z. Wu, L. Qiu, A. Ross, et al. June. 2024. Reasoning or reciting? Exploring the capabilities and limitations of language Models through counterfactual tasks. In K. Duh, H. Gomez, and S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, June 16–21, 2024. ACL, 1819–1862. DOI: <https://doi.org/10.18653/v1/2024.naacl-long.102>.

# CHAPTER 9

---

## Modeling Emotions in Multimodal LLMs

### Abstract

In human-computer interaction, recognizing and responding to a user's emotional state is crucial for effective communication and successful task completion. For instance, a caregiving AI agent capable of detecting pain or depression in a patient could offer tailored empathetic support and appropriate medical interventions while adhering to ethical guidelines and safeguarding patient well-being. This chapter examines cognitive research on human emotions and proposes the Behavioral Emotion Analysis Model (BEAM), a novel emotion spectrum framework that incorporates both basic emotions and their linguistic antonyms. BEAM provides a comprehensive way to understand and represent emotional states in language and is designed to be integrated with Large Language Models (LLMs). By leveraging BEAM, LLMs can adapt their linguistic behaviors and expressions based on the detected emotional state of the user, ensuring that responses are both empathetic and ethically aligned.

### 9.1

#### Introduction

During the development of SocraSynth [Chang 2023] (Chapter 5), a multi-LLM debate framework, we discovered a fundamental principle about the behavior of LLMs. While investigating how to control debate “contentiousness,” we found that an LLM’s linguistic behavior could be systematically altered through emotional conditioning. High contentiousness produced confrontational tones and polarized language, while low contentiousness led to agreeable, considerate discourse. This observation went beyond the original scope of improving multi-agent debates—it revealed a mechanism for steering LLM’s behavior through emotional states.

Most Multi-Agent Debate (MAD) systems [Fu et al. 2023, Li et al. 2023, Liang et al. 2024, Michael et al. 2023, Abdelnabi et al. 2024, Smit et al. 2024] function as

ensemble learning techniques, similar to bagging [Breiman 1996] or mixtures of experts [Jacobs et al. 1991], where LLMs simply exchange ideas without deep exploration. Our work with SocraSynth and EVINCE [Chang 2024a] (Chapters 6 and 7) addressed this limitation by dynamically modulating emotional states throughout the debate. High contentiousness drives LLMs to explore novel perspectives and challenge existing viewpoints, while low contentiousness promotes the synthesis of established ideas. This emotional modulation creates a natural debate progression: from vigorous exploration of diverse viewpoints, through reasoned analysis and refutation, to the emergence of well-examined, conciliatory conclusions.

While LLM training is often viewed simply as next-token prediction, its effects are far more profound. Training documents represent humans pursuing diverse goals—conducting research, exchanging opinions, expressing emotions—through a vast array of linguistic behaviors. This understanding, combined with our experience in modeling contentiousness through in-context learning, suggests an intriguing possibility: can we condition LLMs with specific goals and emotions to generate outputs that leverage these learned linguistic behaviors? Recent empirical studies support this approach, showing that LLM outputs can be traced to their source [Liu et al. 2025] and that in-context learning operates as conditional statistics in a Bayesian framework [Xie et al. 2022].

Our exploration through bias reduction work in news articles and Wikipedia content (Chapter 8) demonstrated that emotional states significantly influence LLM outputs. This finding, combined with our debate framework experiences, suggests that a mathematical model of emotions could provide a foundation for systematic behavior control. Before exploring the mapping between emotions and behaviors, we must first establish a rigorous framework for representing and manipulating basic emotions.

To lay the groundwork for emotion-based behavior control, this chapter develops the Behavioral Emotion Analysis Model (BEAM). While LLMs were initially seen as “black boxes” [Bubeck et al. 2023], our observations, along with insights from Prof. Stuart Russell, suggest that emotional states can be systematically modeled and conveyed to LLMs via *context*. Our model addresses three fundamental questions:

- (1) *What basic emotions form a complete basis?* We identify k fundamental emotion spectra, each defined by negative and positive antonyms (e.g., “hate-love,” “anxiety-calmness”). We focus on basic emotions while excluding complex emotions like “regret” that arise from combinations of basic states. Each spectrum represents a continuous axis along which emotional states can be measured and modified.

- (2) *How can we mathematically manipulate emotions?* We develop a mathematical framework using negation and scaling operations for precise positioning of emotional states along each spectrum. For instance, given the hate-love spectrum, we can represent intermediate states through scaling (e.g.,  $0.7 \times \text{love}$ ) and use negation to move between opposing states (e.g.,  $\neg\text{hate} \approx \text{love}$ ). These operations provide the foundation for systematic emotion manipulation.
- (3) *Can emotions predict behaviors?* We conduct a preliminary study using self-supervised learning to explore the relationship between emotional states and linguistic behaviors. By analyzing text samples from our debate framework, we train a model to predict behavioral patterns from emotional states without explicit labeling. This study validates our emotional spectra's utility in modeling behavioral outcomes.

While Chapter 10 will explore in depth how these emotional states map to specific linguistic behaviors, this chapter focuses on establishing the mathematical framework for representing and manipulating basic emotions, validated through preliminary self-supervised learning experiments. By grounding our model in both mathematical rigor and empirical testing, we create a foundation for systematic, emotion-based behavior control in LLMs.

## 9.2

### Qualifying and Quantifying Emotions

We start by examining emotion modeling research in cognitive science and psychology, specifically highlighting the seminal contributions of Paul Ekman and Robert Plutchik [Ekman 1999]. While we recognize the importance of their work in identifying “basic” emotions (defined shortly), we also address the limitations of such heuristic-based modeling that depends on observational studies lacking rigorous, invariant scientific validation. To enhance the precision in quantifying emotions of varying intensities, we propose incorporating linguistic analysis into our methodologies. Our approach aims to refine the quantification process by leveraging language as a tool to measure and understand emotional expressions accurately.

Paul Ekman and Robert Plutchik are renowned psychologists noted for their foundational work in the field of emotion research. They developed models that categorize basic emotions, which are fundamental and universal emotions believed to be experienced by all humans, transcending cultural boundaries. These emotions are considered “basic” due to their universal recognition, distinct facial expressions, and direct associations with survival mechanisms. They are innate and reflective (beneath consciousness), rather than learned, serving



**Figure 9.1** Plutchik's Wheel of Emotions [Plutchik 2001]. The eight basic emotions are organized into four pairs, each annotated with various degrees of emotions between its two poles.

as the building blocks for more complex emotional experiences (through consciousness processing) that can vary significantly across different cultures and individuals.

Expanding upon this foundational work, Plutchik's "Wheel of Emotions" introduces a more detailed model that includes eight primary bipolar emotions. These are outlined in his seminal works [Plutchik 1980, 2001], which are cited as general references on the topic.

Figure 9.1 illustrates the eight primary emotions at various intensities:

- (1) *Joy*: A feeling of great pleasure or happiness.
- (2) *Trust*: A sense of reliability or confidence.
- (3) *Fear*: An unpleasant emotion caused by the belief that something is dangerous, likely to cause pain, or a threat.

- (4) *Surprise*: A feeling caused by something unexpected.
- (5) *Sadness*: A feeling characterized by sorrow or unhappiness.
- (6) *Disgust*: A feeling of revulsion or strong disapproval aroused by something unpleasant or offensive.
- (7) *Anger*: A feeling of annoyance, displeasure, or hostility.
- (8) *Anticipation*: The action of looking forward to something; expectation or prediction.

These emotions are conceptually paired as opposites in the following manner: joy-sadness, anticipation-surprise, trust-disgust, and anger-fear, based on their evolutionary roles and adaptive functions. Each pair is annotated with degrees of emotion ranging between its two poles. For example, along the axis of *joy vs. sadness*, emotions range from serenity to ecstasy and from grief to pensiveness.

### **9.2.1 Observations and Discussion**

Foundational theories in psychology support the selection of these four emotion pairs as opposites. However, while all four pairs exhibit opposition, “trust-disgust” and “anger-fear” are not strict linguistic antonyms. Trust and disgust entail opposing evaluations, often leading to different actions: trust fostering approach and disgust promoting avoidance. Similarly, anger and fear, while both negative, differ in their response to threats: anger can lead to confrontation and fear to withdrawal. Therefore, the following approximations do not hold:

$$\neg\text{trust} \not\approx \text{disgust} \text{ and } \neg\text{anger} \not\approx \text{fear}.$$

Since our focus is on modeling emotions in LLMs, rather than directly replicating the complex emotional experiences of humans, we prioritize the use of linguistic antonyms for their simplicity and practicality. As Klaus Scherer aptly noted, defining emotions can be a contentious and often fruitless endeavor [Scherer 2005]. To avoid such debates and maintain a clear focus, our study limits itself to universal, basic emotions, avoiding the theoretical ambiguities that arise with more subtle or mixed emotional states. This allows us to capture the primary emotional valence (positive or negative) expressed in text, providing a foundational framework for our model. Thus, we establish the following approximate relationships:

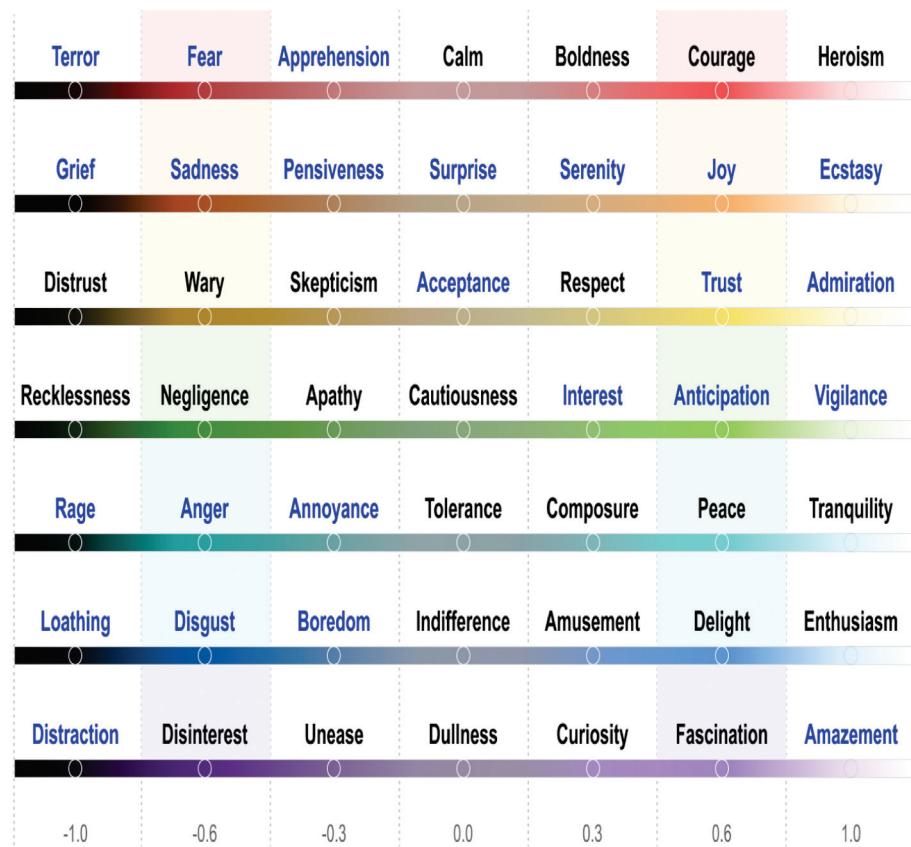
$$\begin{aligned} \neg\text{fear} &\approx \text{courage}, \neg\text{wary} \approx \text{trust}, \neg\text{anger} \approx \text{peace}, \\ \text{and } \neg\text{disgust} &\approx \text{delight}. \end{aligned}$$

### 9.2.2 Behavioral Emotion Analysis Model

Figure 9.2 presents BEAM, organized into seven distinct spectra. Each spectrum encompasses a range of emotional intensity, anchored by a negative and positive extreme, with neutral in the middle. Emotions belonging to the same spectrum are placed along this continuum, with four approximate intensity levels quantified as  $-0.6$ ,  $-0.3$ ,  $+0.3$ , and  $+0.6$ .

This spectrum model offers two key advantages:

- (1) *Antonym-Based*: The use of antonyms allows for easy navigation between opposing emotions. For instance, applying negation to “joyful” naturally leads to “sad,” streamlining the process of identifying contrasting emotions.



**Figure 9.2** Behavioral Emotion Analysis Model (BEAM). Each row depicts an emotion spectrum, with negative emotions on the left and positive emotions on the right, with emotions of varying intensities interspersed in between, which can be calibrated for specific applications. “Basic” emotions are highlighted in blue.

- (2) *Scalable Intensity*: The model enables the scaling of emotions along the spectrum, providing a deep understanding of varying degrees of emotional intensity. For example, we can “dial up” the intensity of “joy” to “ecstatic” or “dial down” the intensity of “anger” to “annoyed.”

This flexible and intuitive structure facilitates a more granular and accurate representation of emotions in text, paving the way for advanced applications in natural language processing and human-computer interaction.

### 9.2.3 Emotion Inclusion and Exclusion Criteria

All “basic” emotions as defined by Ekman and Plutchik are incorporated into our model, along with their linguistic antonyms. This approach streamlines the framework by excluding complex emotions from the Geneva Wheel of Emotions, which are heavily influenced by personal values and experiences. For example, guilt and shame are consequential, consciously aware, and culturally dependent nature [Tangney and Fischer 1995]. These emotions typically arise as reactions to behaviors rather than direct drivers of them. Guilt may motivate behaviors aimed at covering up or remedying an action, while shame, characterized by painful self-assessment, often inhibits individuals from seeking social support or engaging in corrective actions due to fear of judgment. The triggers for these emotions can vary across cultures [Hofstede 1980, Fiske et al. 1998], and since expressing these “reactions” does not usually violate ethical codes, we exclude them from our model.

## 9.3 Empirical Study: Linguistic Features of Emotion

This section presents the outcomes of two experimental studies focusing on contrasting emotional pairs from the Emotion Spectra: “ecstasy vs. grief,” and “admiration vs. disgust.”

Each emotional pair experiment unfolded in three phases. Initially, we instructed GPT-4 to reframe 60 articles (30 poems of John Keats [Bush 1952] and 30 poems of Emily Dickinson [Todd and Higginson 1983]), infusing each with six varying intensities of the emotional spectrum, from the most positive to the most negative. Subsequently, we prompted GPT-4 to elucidate the linguistic strategies it utilized to depict each of the six emotional gradations.

The first experiment models various degrees of happiness. In this experiment, we tasked GPT-4 with reinterpreting selected poems by John Keats across seven emotional levels: *ecstasy* (very happy); *joy* and *serenity* (neutral); *pensive*, *sad*, and *grief* (very sad). Following the approach of our contentiousness experiments, after GPT-4 adapted Keats’ poems to reflect these emotional states, we asked it to identify the linguistic features it employed to express each emotion in the rewrites.

### 9.3.1 Joy vs. Sadness

Table 9.1 outlines GPT-4’s approach to varying emotional levels, illustrating how it adjusts vocabulary, tone, imagery, and thematic focus, including the depiction of entities, locations, and scenarios. Remarkably, beyond just syntactic and semantic manipulation, GPT-4 also incorporates landscape scenes—natural features such as the sky, trees, clouds, and flowers—and utilizes brightness, colors, and personal expressions to convey specific emotional states. Although the analysis is based on a limited set of samples from two authors, it effectively demonstrates GPT-4’s ability to employ a palette of both broad and fine strokes, utilizing diverse colors and textures to vividly illustrate human emotions and resonate with readers.

Recognizing the profound communicative power of visual art, we transitioned to a more graphical representation. Utilizing the linguistic elements identified for each emotional tier, Figure 9.3 presents six watercolor paintings, each representing a different emotional level. Our prompt to DALL-E (of GPT-4) was to create a watercolor depicting a lady in a garden experiencing a specific mood, and we attached the corresponding linguistic features from Table 9.1 to clearly define that mood. This approach ensures that with a well-defined context, DALL-E accurately captures the specific and detailed aspects of the mood, effectively translating the emotional intensity into visual form. These artistic renditions not only confirm GPT-4’s ability to transform emotional lexicons into evocative imagery with remarkable precision but also validate the accuracy of the emotional lexicons generated by GPT-4, demonstrating their effectiveness in conveying precise emotional states.

### 9.3.2 Admiration/Delight vs. Disgust

This experiment asks Gemini to rewrite a scene in Romeo and Juliet by setting Juliet’s emotion in six different levels: loathing, disgust, boredom, respect, admiration/delight, and enthusiasm.

The excerpt provided in Table 9.A in Appendix 9.A is from one of the most iconic scenes in William Shakespeare’s “Romeo and Juliet,” commonly known as the balcony scene. This is Act 2, Scene 2, where Romeo, having just met Juliet at the Capulet’s feast, sneaks into the Capulet’s orchard and overhears Juliet speaking of her love for him from her balcony, unaware that he is there.

The scene captures the moment of their mutual declaration of love and is famous for Juliet’s reflections on the nature of names and identity, encapsulated in her line, “What’s in a name? That which we call a rose / By any other name would smell as sweet.” It’s a profound exploration of love and identity, where both characters express their willingness to renounce their family names for the sake of their love.

**Table 9.1** GPT-4 reinterpreted selected poems by Keats across a spectrum of happiness levels and then was tasked with identifying the linguistic adjustments it made to convey each emotional state, from very happy to very sad. It's important to note that the analysis table was generated by GPT-4 itself, reflecting on its own modifications

Emotion Level	Vocabulary	Tone	Imagery	Subject Focus (Person)
Very Happy (1.0)	Joyful, exhilarating, vibrant (e.g., delighted, thrilled, ecstatic)	Enthusiastic, lively (e.g., exuberant, spirited, radiates joy)	Bright landscapes, summer waters (e.g., radiant, sparkling, glowing)	Celebratory, beauty of a subject (e.g., adoration, admiration, splendor)
Happy (0.7)	Positive, warm, inviting (e.g., pleasant, cozy, cheerful)	Cheerful, contemplative (e.g., thoughtful, satisfied, warmth)	Warm scenes, serene woods (e.g., gentle, peaceful, lush)	Charm, subtle desires (e.g., affection, fondness, beauty, yearning)
Slightly Happy (0.3)	Balanced, light, serene (e.g., calm, gentle, soothing)	Reflective, optimistic (e.g., hopeful, positive)	Balanced landscapes, serene woods (e.g., tranquil, mild)	Simple pleasures, mild yearning (e.g., contentment, wishful)
Neutral (0)	Balanced mix, everyday (e.g., stable, straightforward, regular, steady)	Even, reflective (e.g., balanced, neutral)	Everyday scenes, neutral landscapes (e.g., ordinary, familiar)	Contentment, simple living (e.g., simplicity, normalcy, daily life)
Slightly Sad (-0.3)	Subdued, longing, wistful (e.g., reserved, pensive, yearning)	Melancholic, introspective (e.g., reflective, subdued, introspective musings)	Wistful skies, quiet waters (e.g., subdued, still water, fading colors)	Unfulfilled desires, quiet contemplation (e.g., longing, introspection)
Sad (-0.7)	Melancholic, somber, solitary (e.g., lonely, forlorn, desolate)	Somber, heavy (e.g., sorrowful, melancholic, laden)	Solitary scenes, fading light (e.g., dim, shadowed, lonely)	Deep longing, introspection (e.g., melancholy, contemplation, reflection)

**Table 9.1** (*Continued.*)

Emotion Level	Vocabulary	Tone	Imagery	Subject Focus (Person)
Very Sad (-1.0)	Bleak, sorrowful, dark (e.g., despondent, heartbroken, despairing)	Heavy, despairing (e.g., desolate, gloom, over-whelmed)	Bleak landscapes, darkened skies (e.g., stark, bleak, barren)	Loss, profound sadness (e.g., grief, desolation, heartache, void)

**Figure 9.3** A lady in a garden scene under different emotions. From top-left, happiest, to bottom-right, saddest.

Romeo responds to Juliet's musings by rejecting his name if it means they can be together, and they begin to plan their secret marriage. This scene is pivotal in the play, setting the stage for the subsequent events that unfold, driven by their passion and the social constraints that surround them.

The six versions of rewrites by Gemini are presented in the extended version [Chang 2024b]. In the following, we summarize the linguistic features Gemini used—including diction, imagery, figurative language, sentence structure, implied

body language, and overall tone—to depict two selected emotions: disgust and admiration.

#### **9.3.2.1 Emotion: Disgust**

*Diction:* Employs negative words emphasizing repulsive qualities (e.g., “foul business,” “fetid breath”).

*Imagery:* Evokes revolting comparisons, often mentioning sewers and stench.

*Figurative Language:* Primarily negative similes reinforcing disgust (e.g., “What if her eyes were there . . . the fetid breath from her mouth would surely overpower those stars . . .”).

*Sentence Structure:* Short, choppy sentences, similar to expressions of loathing but with a hint of disdain.

*Implied Body Language:* Recoiling from the window, covering the nose—suggesting physical revulsion.

*Overall Tone:* Disgusted and disapproving.

#### **9.3.2.2 Emotion: Admiration**

*Diction:* Uses positive and intrigued language (e.g., “brilliance,” “music stirs my soul”).

*Imagery:* Creates positive comparisons highlighting attractive qualities (e.g., “stars in all the heaven”).

*Figurative Language:* Positive similes emphasizing Romeo’s appeal (e.g., “What if her eyes were there, they in her head? The brightness of her . . . well, not exactly bright . . . cheek would shame those stars . . .”).

*Sentence Structure:* Varied structure with a sense of curiosity.

*Implied Body Language:* Leaning out the window, engaged expression—indicating interest.

*Overall Tone:* Intrigued, curious, and somewhat impressed.

By adjusting these linguistic features, each rendition vividly captures a unique emotional state for Juliet. The combination of diction, imagery, figurative language, sentence structure, and implied body language collectively shapes Juliet’s perception of Romeo and her reactions to him.

These detailed narratives augment the classic balcony scene, enriching its emotional depth. Table 9.2 presents an overarching view of the various approaches Gemini employs to reflect differing emotional states. It’s fascinating to observe how an LLM can “consciously” mimic human emotional expressions through language. While it’s uncertain if LLMs genuinely grasp the emotions they project or merely simulate them, the effectiveness of these emotional mappings is

**Table 9.2** Gemini's interpretations on the six emotion levels

Emotion	Diction	Imagery	Figurative	Body Language
Loathing	Harsh, Insulting	Disgusting	Weak Similes	Scowling, Spitting
Disgust	Negative	Unpleasant	Negative Similes	Recoiling
Aversion	Dismissive	Mundane	Undermining Similes	Distant
Respect	Formal	Neutral	None	Composed
Admiration	Positive	Positive	Positive Similes	Leaning In
Veneration	Elevated	Saintly	Hyperbole	Reverent

noteworthy. If these mappings resonate, they might reveal new insights into how we interpret and attribute emotions in textual expressions.

## 9.4 Qualifying and Quantifying Ethics

We conduct a preliminary study using self-supervised learning to explore the relationship between emotional states and linguistic behaviors.

The primary objective is to endow LLMs with the autonomous capability to recognize and rectify undesirable actions, akin to an individual's introspective process to avert potential wrongs. By allowing an LLM to self-assess its outputs prior to public release, the system can proactively identify and amend ethical lapses, thus aligning its behavior with established ethical standards across contexts.

### 9.4.1 Ethical Violation Correlates to Emotions

Grounding ethics in universal principles and logical reasoning emphasizes the objective and rational foundation of ethical decision-making. According to this perspective, universal ethical principles—such as justice, fairness, and respect for autonomy—define right from wrong, independent of personal emotions or specific circumstances.

However, an exploration into the origins of ethical violations, such as prohibitions against killing and stealing, reveals a deep-rooted connection to human emotions. Emotions, conceptualized as vectors of energy with varying intensity and direction, significantly shape ethical behavior, influenced by contextual factors. This understanding suggests that ethical judgments are not merely logical deductions, but they also involve a complex interplay of emotions, individual circumstances, and societal norms. Emotions, therefore, are interwoven with ethical actions, playing a crucial role in determining whether an action is deemed ethical or unethical.

This perspective enables us to analyze ethical violations through a multi-dimensional lens, considering the trajectory, intensity, and context of emotional

energy. This framework, inspired by Dante Alighieri's *Divine Comedy* [Dante 2017], offers a novel way to understand how emotions can either drive individuals toward ethical actions or lead them astray into unethical behavior.

- (1) *Trajectory of Energy*: This parameter represents the direction in which emotional energy is directed, with each direction corresponding to a specific violation. The trajectory visualizes the orientation of an energy, with eight distinct trajectories symbolizing the 16 characterized violations/sins.
- (2) *Intensity of Energy*: The intensity reflects the strength or magnitude of the emotional energy. Overly intense emotions can cloud judgment, leading to impulsive or unethical decisions, while insufficient emotional intensity might result in apathy or lack of consideration for ethical implications. The appropriate intensity of emotional energy is crucial for balanced ethical decision-making.
- (3) *Context*: The situational factors and the environment in which the emotional energy operates significantly influence ethical outcomes. The context includes cultural norms, individual circumstances, societal pressures, and specific scenarios that shape how emotions are perceived and acted upon. It determines the ethical framework within which the energy and its trajectory are evaluated. Consider the ethical principle "Do not lie." While deception might typically carry negative emotional weight, in contexts such as a doctor or son concealing a terminal diagnosis from a father, the contextual factors reduce the negative valence. This context-aware mathematical framework provides the foundation for precise and contextually calibrated emotion manipulation.

#### **9.4.2 Twelve Vices and Twelve Pairs of Vices**

Based on our theory that ethical violations (vices or sins) can be represented by three distinct parameters—trajectory of energy, intensity of energy, and context in which this energy is manifested—we can identify 12 pairs of common sins. The balance between the two extremes of energy—neither too intense nor too mild—exemplifies virtue. For example, pride, characterized by excessive self-love, and insecurity, marked by feelings of inadequacy, find balance in the moderate energy of self-respect, representing the virtue of equilibrium.

- (1) *Pride (Excessive Self-Love) and Insecurity (Inadequate Self-Love)*: Self-respect is the virtue that mediates between pride and insecurity, fostering a healthy level of self-esteem and confidence without tipping into arrogance or self-doubt.

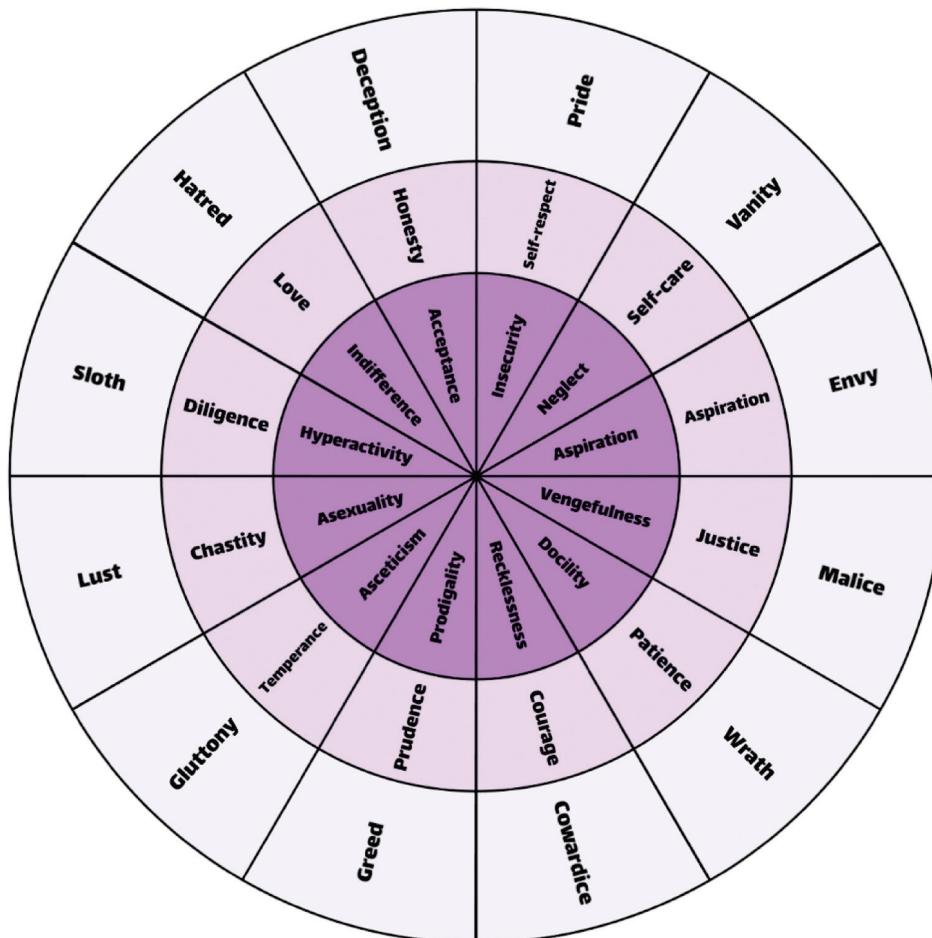
- (2) *Vanity (Excessive Focus on Appearance) and Neglect (Inadequate Attention to Self-Care)*: Modesty is the virtue that lies between vanity and neglect, promoting a balanced approach to one's appearance and self-care.
- (3) *Envy (Excessive Desire for Others' Traits or Possessions) and Apathy (Inadequate Desire for Personal Growth or Achievement)*: Contentment is the virtue that balances envy and apathy, fostering satisfaction with one's own achievements and qualities without coveting those of others or lacking ambition.
- (4) *Malice (Excessive Desire to Harm) and Excessive Forgiveness (Inadequate Response to Wrongdoing)*: Justice is the virtue that lies between malice and excessive forgiveness, ensuring fair treatment and accountability without intentions to harm or overlooking wrongdoing.
- (5) *Wrath (Excessive Anger) and Docility (Inadequate Concern for Justice or Fairness)*: Patience is the virtue that moderates wrath and docility, enabling one to endure difficulties or injustices calmly without reacting in anger or compromising moral principles.
- (6) *Cowardice (Inadequate Courage) and Recklessness (Excessive Risk-Taking)*: Courage is the virtue that balances cowardice and recklessness, encouraging one to face challenges and risks with bravery while considering the consequences.
- (7) *Greed (Excessive Acquisition) and Generosity (Inadequate Retention for Self)*: Prudence is the virtue that mediates between greed and excessive generosity, guiding wise decisions regarding the acquisition and sharing of resources.
- (8) *Gluttony (Excessive Consumption) and Asceticism (Inadequate Indulgence)*: Temperance is the virtue that balances gluttony and asceticism, promoting moderation in consumption and enjoyment of life's pleasures without excess or deprivation.
- (9) *Lust (Excessive Sexual Desire) and Chastity (Inadequate Sexual Expression)*: Purity is the virtue balancing lust and chastity, advocating for healthy and respectful expressions of sexuality.
- (10) *Sloth (Excessive Laziness) and Hyperactivity (Inadequate Rest)*: Diligence is the virtue that balances sloth and hyperactivity, inspiring consistent and focused effort while allowing for necessary rest and rejuvenation.
- (11) *Deception (Excessive Dishonesty) and Gullibility (Inadequate Skepticism)*: Honesty is the virtue between deception and gullibility, emphasizing truthfulness and integrity in one's actions and beliefs.

- (12) *Hatred (Excessive Animosity) and Indifference (Inadequate Empathy)*: Love is the virtue that balances hatred and indifference, fostering genuine concern and connection with others while avoiding animosity and apathy.

These pairs illustrate how both excess and deficiency in similar emotional trajectories can lead to distinct but related ethical issues, emphasizing the importance of balance in emotions and actions.

#### **9.4.3 The Wheel of Virtue (or Vices)**

Figure 9.4 presents the Wheel of Virtues based on the characterization of the 12 pairs common sins.



**Figure 9.4** The Wheel of Virtues.

The wheel is divided into 12 segments, each corresponding to a specific pair of opposing vices. At the center of each spoke is the virtue that represents the ideal midpoint between the two extremes, emphasizing that virtues lie in balance, not at the extremes.

#### **9.4.4 Ethical Alignment with Context**

Effective ethical alignment requires understanding and adapting to cultural contexts rather than applying universal rules. LLMs must recognize how ethical principles are interpreted and applied differently across cultures while maintaining core ethical guardrails. To achieve this balance, we employ our Self-Supervised with Human Feedback (SSHF) pipeline, introduced in Section 9.2, to develop culturally aware ethical behavior.

The SSHF pipeline trains LLMs to recognize and generate culturally appropriate linguistic behaviors through iterative refinement. Similar to our approach for modeling emotions like happiness, we task the LLM with generating content that adheres to specific ethical standards while incorporating cultural context. Through feedback loops, the LLM learns to adjust its ethical reasoning based on contextual cues and cultural norms. This process not only improves the LLM's contextual decision-making but also enables it to explicitly identify the linguistic features that characterize ethical behavior in different cultural contexts.

In Chapter 10, we propose the DIKE and ERIS duality to deal with context adaptation. We also address the shortcomings of using Reinforcement Learning from Human Feedback (RLHF) alone to perform ethical alignment. Nevertheless, our ethical modeling procedure consists of five key steps:

- (1) *Defining Ethical Framework:* Using the Wheel of Virtues, administrators identify paired vice-virtue spectra (e.g., pride vs. self-respect, hatred vs. compassion, envy vs. aspiration) that establish the core ethical dimensions for the LLM's behavior.
- (2) *Generating Comparative Content:* The LLM creates a training dataset by modifying articles to demonstrate both vice-aligned and virtue-aligned expressions. This produces paired examples showing how similar content can be expressed with different ethical valences.
- (3) *Extracting Ethical Patterns:* Through an analysis of these paired examples, the LLM identifies distinguishing linguistic features between harmful and beneficial content, creating a systematic framework for ethical content assessment.

- (4) *Applying Ethical Transformation:* During content generation, the LLM uses the extracted patterns to detect vice-aligned content and transform it using virtue-aligned linguistic features, preserving the core message while improving ethical alignment.
- (5) *Support Adaptivity:* The system evolves its ethical understanding through user feedback, adapting its rules to handle diverse cultural contexts and edge cases more effectively.

## 9.5

### Concluding Remarks

Recent declarations by [Bengio et al. \[2024\]](#) and [Dalrymple et al. \[2024\]](#) emphasize the critical need to address AI safety given the technology's rapid advancement. This chapter has laid the mathematical foundation for modeling emotions in LLMs through the Behavioral Emotion Analysis Model (BEAM). By representing emotions as spectra defined by antonym pairs and developing operations for their manipulation, BEAM provides a systematic framework for understanding and controlling emotional expression in LLM outputs.

Our approach differs from traditional rule-based methods by focusing on the fundamental building blocks of emotional expression. Through precise quantifiable formulation of basic emotions and their relationships, we create a tractable framework for mapping emotions to behaviors that can be validated through self-supervised learning. This quantitative approach to emotion modeling represents a novel contribution to AI safety, offering an interpretable and systematic method for understanding and controlling LLM behavioral patterns.

While this chapter establishes the framework for emotion modeling and presents a preliminary study on emotion-driven linguistic behaviors, Chapter 10 provides an in-depth treatment of AI safety and ethical alignment through an architecture inspired by the three branches of the US government: *Executive* for knowledge generation, *Legislative* for establishing ethical guardrails, and *Judicial* for interpreting these guardrails within specific contexts and cultural norms. This checks-and-balances framework mitigates the shortcomings of RLHF by providing an adaptable and interpretable system for ethical reasoning that can evolve with societal values while maintaining core safety principles.

## 9.A

### Appendix A

This Table 9.A is referenced in the second case study (Romeo and Juliet balcony scene).

**Table 9.A** Act 2, Scene 2 from Romeo and Juliet

Role	Narrative
Romeo	<p>He jests at scars that never felt a wound.</p> <p>But, soft! What light through yonder window breaks? It is the east, and Juliet is the sun. Arise, fair sun, and kill the envious moon, Who is already sick and pale with grief, That thou, her maid, art far more fair than she. Be not her maid, since she is envious. Her vestal livery is but sick and green, And none but fools do wear it. Cast it off.</p>
Juliet	Ay me!
Romeo	<p>It is my lady, O, it is my love! O, that she knew she were! She speaks, yet she says nothing. What of that? Her eye discourses; I will answer it. I am too bold, 'tis not to me she speaks. Two of the fairest stars in all the heaven, Having some business, do entreat her eyes To twinkle in their spheres till they return. What if her eyes were there, they in her head? The brightness of her cheek would shame those stars, As daylight doth a lamp. Her eye in heaven Would, through the airy region, stream so bright That birds would sing and think it were not night. See, how she leans her cheek upon her hand. O, that I were a glove upon that hand, That I might touch that cheek!</p>
Juliet	O Romeo, Romeo! Wherefore art thou Romeo? Deny thy father and refuse thy name; Or, if thou wilt not, be but sworn my love, And I'll no longer be a Capulet.
Romeo	<p>[Aside] Shall I hear more, or shall I speak at this?</p> <p>[Aloud] I take thee at thy word. Call me but love, and I'll be new baptized; Henceforth I never will be Romeo.</p>
Juliet	<p>What man art thou that, thus bescreened in night, So stumblest on my counsel?</p> <p>By whose direction found'st thou out this place?</p> <p>Thou knowest the mask of night is on my face, Else would a maiden blush bapaint my cheek For that which thou hast heard me speak tonight. Fain would I dwell on form; fain, fain deny What I have spoke. But farewell, compliment. Dost thou love me? I know thou wilt say "Ay," And I will take thy word; yet if thou swear'st, Thou mayst prove false. At lovers' perjuries, They say Jove laughs. O gentle Romeo, If thou dost love, pronounce it faithfully.</p>

## References

- S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In A. Globerson, L. Mackey, D. Belgrave, et al. (Eds.), *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*. Vol. 37, Vancouver, BC, Canada, December 10–15, 2024. Curran Associates Inc., Red Hook, NY, 83548–83599. DOI: <https://doi.org/10.48550/arXiv.2309.17234>.
- Y. Bengio, G. Hinton, A. Yao, et al. May. 2024. Managing extreme AI risks amid rapid progress. *Science* 384, 6698, 842–845. DOI: <https://doi.org/10.1126/science.adn0117>.
- L. Breiman. 1996. Bagging predictors. *Mach. Learn.* 24, 2, 123–140. DOI: <https://doi.org/10.1007/BF00058655>.
- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. DOI: <https://doi.org/10.48550/arXiv.2303.12712>.
- D. Bush (Ed.). 1952. *Selected Poems and Letters by John Keats*. Houghton Mifflin Company, Boston, MA.
- E. Y. Chang. December. 2023. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 10th International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang. August. 2024a. EVINCE: Optimizing Multi-LLM dialogues using conditional statistics and information theory. DOI: <https://doi.org/10.48550/arXiv.2408.14575>.
- E. Y. Chang. 2024b. Integrating emotional and linguistic models for ethical compliance in large language models. DOI: <https://doi.org/10.48550/arXiv.2405.07076>.
- A. Dante. 2017. *The Divine Comedy*. Aegitas, 1320.
- D. Dalrymple, J. Skalse, Y. Bengio, et al. 2024. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. DOI: <https://doi.org/10.48550/arXiv.2405.06624>.
- P. Ekman. 1999. Basic emotions. In *Handbook of Cognition and Emotion*. John Wiley and Sons Ltd., 45–60. DOI: <https://doi.org/10.1002/0470013494.ch3>.
- A. P. Fiske, S. Kitayama, H. R. Markus, and R. Nisbett. 1998. The cultural matrix of social psychology. In *The Handbook of Social Psychology*. Vol. 2. McGraw-Hill, Boston, MA: 915–981.
- Y. Fu, H. Peng, T. Khot, and M. Lapata. 2023. Improving language model negotiation with self-play and in-context learning from AI feedback. DOI: <https://doi.org/10.48550/arXiv.2305.10142>.
- G. Hofstede. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA.
- R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. March. 1991. Adaptive mixtures of local experts. *Neural Comput.* 3, 1, 79–87. DOI: <https://doi.org/10.1162/neco.1991.3.1.79>.
- H. Li, Y. Chong, S. Stepputtis, et al. 2023. Theory of mind for multi-agent collaboration via large language models. In H. Bouamor, J. Pino, and K. Bali (Eds.). *Proceedings of the 2023*

- Conference on Empirical Methods in Natural Language Processing.* Singapore, December 6–10, 2023. ACL, 180–192. DOI: <https://doi.org/10.18653/v1/2023.emnlp-main.13>.
- T. Liang, Z. He, W. Jiao, et al. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, November 12–16. ACL, 17889–17904. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.992>.
- J. Liu, T. Blanton, Y. Elazar, et al. July. 2025. OLMoTrace: Tracing language model outputs back to trillions of training tokens. In P. Mishra, S. Muresan, and T. Yu (Eds.). *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Association for Computational Linguistics, Vienna, 178–188. DOI: <https://doi.org/10.18653/v1/2025.acl-demo.18>.
- J. Michael, S. Mahdi, D. Rein, et al. 2023. Debate helps supervise unreliable experts. DOI: <https://doi.org/10.48550/arXiv.2311.08702>.
- R. Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Emotion: Theory, Research, and Experience*. Academic Press, New York, 3–33.
- R. Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* 89, 4, 344–350.
- K. R. Scherer. 2005. What are emotions? And how can they be measured? *Soc. Sci. Inf.* 44, 4, 695–729. DOI: <https://doi.org/10.1177/0539018405058216>.
- A. Smit, N. Grinsztajn, P. Duckworth, T. D. Barrett, and A. Pretorius. 2024. Should we be going MAD? A look at multi-agent debate strategies for LLMs. In R. Salakhutdinov, Z. Kolter, K. Heller, et al. (Eds.). *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. Vienna, Austria, July 21–27, 2024. JMLR.org, 45883–45905.
- J. P. Tangney and K. W. Fischer (Eds.). 1995. *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford Press, New York, 198–218.
- M. L. Todd and T. W. Higginson (Eds.). 1983. *Collected Poems of Emily Dickinson*. Chatham River Press Classics, New York. Retrieved from <https://www.amazon.com/Collected-Poems-Dickinson-Chatham-Classics/dp/0517415143>.
- S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations (ICLR)*. Retrieved from <https://openreview.net/forum?id=RdJVFCHjUMI>.

## CHAPTER 10

---

# A Checks-and-Balances Framework for Context-Aware Ethical AI Alignment

### Abstract

This chapter introduces a checks-and-balances framework for the ethical alignment of Large Language Models (LLMs), inspired by three-branch governmental systems. It implements independent yet interacting components: LLMs as the executive branch for knowledge generation, DIKE as the legislative branch that establishes ethical guardrails, and ERIS as the judicial branch for contextual interpretation. Beyond structural separation, the chapter addresses a fundamental challenge: regulating emotion to shape behaviors. Drawing from psychological theories where managing emotional responses prevents harmful behaviors, we develop a self-supervised learning pipeline that maps emotions to linguistic behaviors, enabling precise behavioral modulation through emotional conditioning. By integrating this approach with adversarial testing, our framework demonstrates how DIKE and ERIS direct linguistic behaviors toward ethical outcomes while preserving independence throughout knowledge generation, ethical oversight, and contextual interpretation.

## 10.1

### Introduction

Ethical alignment in Large Language Models (LLMs) is a critical challenge, particularly given the limitations of Reinforcement Learning from Human Feedback (RLHF) [[OpenAI 2023](#), [Rafailov et al. 2023](#)]. Although RLHF has demonstrated success in aligning AI systems with human values, it encounters two major issues: (1) susceptibility to social biases when feedback is polarized, and (2) vulnerability to reward hacking, where the system optimizes for feedback without genuine ethical

improvement [Christiano et al. 2017, Skalse et al. 2022]. These issues can result in unethical behavior or inconsistent performance.

Beyond these implementation challenges, RLHF faces a more fundamental conceptual limitation: its narrow focus on isolated behaviors rather than holistic patterns. This reactive strategy is similar to a “Whack-a-Mole” game, where addressing one problematic behavior does not prevent the emergence of others. For example, consistently instructing someone to make their bed does not necessarily cultivate overall tidiness, such as doing laundry or washing dishes. Similarly, RLHF often emphasizes short-term fixes at the cost of long-term coherence, leading to catastrophic forgetting: users have reported that optimizing one task in ChatGPT can degrade performance in unrelated areas [Kirkpatrick et al. 2017, Lin et al. 2024, Dai et al. 2025]. This challenge mirrors the difficulty of treating addiction, where addressing one symptom may reveal deeper psychological dependencies [Torrens et al. 2005, Sinha 2008].

To overcome these challenges, we propose a checks-and-balances framework inspired by governmental structures, where independent but interacting components maintain accountability and balance. Our architecture integrates three components: LLMs serve as the executive branch for knowledge generation; DIKE (representing justice) functions as the legislative branch to set ethical standards; and ERIS (representing discord) acts as the judicial branch, providing adversarial testing and contextual interpretation. In Greek mythology, the goddess Dike embodies order and justice, while goddess Eris signifies discord, forming a duality that our framework leverages to balance ethical guidance with adversarial scrutiny.

Figure 10.1 illustrates this three-branch architecture, where the neurally independent components—LLMs as the foundation, with DIKE and ERIS as oversight layers—interact through structured interfaces while maintaining strict separation of their neural architectures and parameters.

### 10.1.1 Emotion Regulation as Behavioral Control

A fundamental question underlies our framework: Can regulating emotions shape behaviors, and can similar principles be applied to LLMs? In human psychology, emotions significantly drive behaviors: anger and contempt can provoke aggression, and rage and envy can result in harmful actions [Damasio 1994]. Therefore, emotion regulation is essential for behavioral control. Techniques such as cognitive reframing and attentional deployment are known to reduce negative behavioral outcomes by managing emotional intensity.

Unlike humans, who struggle with emotion regulation due to complex neural and cognitive processes [James 1884, Gross 1998], LLMs lack intrinsic emotional



**Figure 10.1** Framework with three independent branches. Bottom: Knowledge LLMs (executive); Left: DIKE (legislative); Right: ERIS (judicial). (Photo credit: DALL-E).

states altogether. However, empirical evidence shows that LLMs can generate text with consistent emotional characteristics through controlled prompt engineering [Chang 2024a]. Indeed, the work of Tak and Gratch [2024] demonstrated that LLMs such as GPT-4 align more closely with human judgments when interpreting others' emotions from a third-person perspective than when attempting to model self-attributions of emotion. This creates a unique opportunity: by leveraging LLMs' ability to model the average human observer's emotional interpretations, we can establish reliable frameworks for ethical alignment that operate through emotional framing rather than explicit rule-following.

Building on this insight, our framework integrates the principles of emotion regulation into the ethical alignment of LLM. Specifically, DIKE analyzes how emotions manifest in linguistic behaviors, while ERIS tests these interpretations against diverse cultural contexts.

### 10.1.2 Checks-and-Balances Framework for Emotion-Guided Ethics

Central to this approach is the synergy between DIKE and ERIS, reflecting the internal conflict often present in the regulation of human emotions. Just as humans balance immediate emotional responses against long-term goals and social norms, our framework establishes an adversarial dynamic between ethical guardrails and contextual challenges. This duality introduces four key innovations:

- (1) *Emotion-Driven Behavioral Modeling*: Based on BEAM (Behavioral Emotion Analysis Model) [Chang 2024a], DIKE uses self-supervised learning to quantify relationships between emotional states and linguistic patterns, guiding ethical decisions through behavioral analysis.
- (2) *Behavior-Aware Ethical Guardrails*: The framework sets dynamic guidelines that account for both content and language behavior, blocking manipulative or harmful communication while preserving factual accuracy and emotional authenticity. These guardrails adjust to different cultural contexts, maintaining consistency while allowing context-dependent interpretation.
- (3) *Adversarial Behavioral Testing*: ERIS challenges DIKE's ethical guidelines by presenting diverse cultural perspectives and edge cases, ensuring the adaptability of ethical reasoning. This adversarial interaction enables the system to address complex scenarios with cultural sensitivity and contextual awareness.
- (4) *Ethical Content Transformation*: When problematic content is detected, ERIS can revise it to maintain the intended emotional tone while ensuring ethical compliance, with human-in-the-loop oversight to validate the appropriateness of revisions. These potential transformations are tested by ERIS in cultural and contextual variations to assess their suitability before implementation.

The experimental section evaluates our framework through three complementary studies. First, we assess whether emotion-mediated classification provides more effective ethical guardrails than direct behavior classification. Next, we examine DIKE's ability to independently evaluate and explain linguistic behaviors. Finally, we test how the adversarial ERIS component enables cultural adaptability and prevents excessive censorship. Although direct comparison with proprietary RLHF implementations is not feasible, our results demonstrate how our approach addresses the theoretical limitations of RLHF in handling contextual diversity without compromising knowledge integrity.

### 10.1.3 Contributions

Our contributions are as follows:

- (1) A novel checks-and-balances architecture for ethical alignment that maintains separation between knowledge generation and ethical reasoning.
- (2) The BEAM model, a quantitative framework for representing emotions along continuous spectra with defined intensity levels, enabling precise emotion regulation in AI systems.

- (3) An emotion-driven approach that guides linguistic behaviors toward ethical outcomes by leveraging cognitive theories of emotion regulation.
- (4) An adversarial framework that enhances ethical reasoning by challenging established guidelines with cultural perspectives, enabling context-sensitive adaptability.
- (5) A theoretical framework explaining the effectiveness of minimal supervision in LLM alignment, formalized as the Unconscious-Conscious Complementarity Thesis (UCCT) in Appendix 10.A.

## 10.2

### Related Work

This section surveys existing work on emotion and behavior modeling across various domains, with a focus on their applications in AI ethics. We examine how linguistic behaviors are influenced by emotional patterns and explore structured approaches that integrate emotional frameworks with linguistic models to improve ethical AI alignment.

We also examine the limitations of RLHF. While effective in refining AI outputs, RLHF can overfit to human annotations, may face challenges in adapting to diverse cultural contexts, may experience parameter drift from optimal settings, and can inadvertently reinforce unintended biases. These observations highlight opportunities to develop more adaptive and principled approaches to complement existing ethical AI alignment methods.

#### 10.2.1 Emotion Modeling

Cognitive-linguistic theories intersect with AI for understanding its behavior. Theories by [Lakoff and Johnson \[1980\]](#), [Talmy \[2000\]](#), and [Jackendoff \[2002\]](#) explore the relationship between language processing and cognitive functions, building on an early work by Freud and Jung [\[Bai et al. 2022\]](#). The concept of “emotion” remains contentious, with definitions varying across disciplines [\[Scherer 2005\]](#). [James \[1884\]](#) attempted to define emotions, but consensus remains elusive.

This chapter focuses on emotional contexts and linguistic behaviors in LLMs, avoiding the complexities of human physiological and personality factors. This approach allows for the exploration of emotion representation in AI systems.

Plutchik and Ekman categorized “basic” emotions with universal facial expressions [\[Plutchik 1982, Ekman 1992\]](#). Later research considered cultural differences [\[Markus and Kitayama 1991, Mesquita and Frijda 1992\]](#), emotion processing [\[Gross 1998\]](#), and neural mechanisms [\[Davidson 2003\]](#). Scherer’s model and appraisal theories by Smith and Ellsworth emphasize cognitive appraisal in emotional experiences [\[Smith and Ellsworth 1985\]](#).

Our model is based on Plutchik's Wheel of Emotions [Plutchik 1982] and Scherer's Geneva Wheel [Scherer 2005], augmented with antonyms to map positive and negative emotions. Language-relevant emotions (e.g., curiosity, confusion, certainty) are incorporated into LLMs to guide their linguistic behavior. See Section 10.3.1 for details.

This selection of basic emotions provides a foundation for validating our approach, recognizing that while it may omit some emotions, it offers a practical starting point for further research.

### 10.2.2 Emotion-Behavior Modeling

Behaviors are profoundly influenced by emotions, as initially posited by the James–Lange Theory of Emotion [James 1884, Lange 1885]. According to this theory, emotional experiences arise from physiological reactions to events. Subsequent research, including studies by Damasio [1994] and Fauconnier and Turner [2002], suggests that the expression and regulation of emotions often manifest in the language we use. High-intensity emotions, such as rage or contempt, can lead to aggressive or destructive behaviors, such as hate speech.

The Schachter–Singer theory [Schachter and Singer 1962], or the two-factor theory of emotion, depicts the role of physiological changes and the cognitive appraisals in determining the label and strength of an emotion. Building on this, the affect-as-information theory, developed by Schwarz and Clore [1983], posits that people use their current emotions to make judgments and decisions to act. If emotions can be adjusted, so can behavior. The work of Barbara Fredrickson [1998] on the effects of positive emotions discusses how we perceive and react to emotions.

Collectively, these theories elucidate the intricate connection between emotions and behaviors, providing the theoretical foundation for our work to incorporate a *behavior advisor* to evaluate and rectify specific behaviors. Section 10.3.2 details how the DIKE framework implements cognitive strategies to mitigate emotions and regulate linguistic behaviors effectively.

### 10.2.3 Reinforcement Learning with Human/AI Feedback

RLHF is the predominant approach to addressing the challenges in AI ethics. This section presents representative works in the field, their advances, and limitations.

#### 10.2.3.1 Human Feedback (RLHF)

Initial advances by Christiano et al. [2017] showed how RLHF can steer language models toward desired outcomes based on human preferences. Newer techniques, such as Identity ( $\Psi$ ), Preference Optimization (PPO), and Generalized Preference

Optimization (GPO), refine this approach by directly optimizing user preferences, thereby effectively addressing scalability challenges. Kahneman-Tversky Optimization (KTO) further simplifies the feedback mechanism by using intuitive responses such as thumbs-up or thumbs-down, thereby enhancing training efficiency without the need for paired data [Azar et al. 2024, Ethayarajh et al. 2024, Tang et al. 2024]. Direct Preference Optimization (DPO) has recently simplified the process by focusing on the clear distinction between preferred and less preferred outputs, thus improving its stability [Rafailov et al. 2023].

#### **10.2.3.2 AI-Generated Feedback (RLAIF)**

To mitigate the dependence on extensive human-generated data, RLAIF (Reinforcement Learning with AI Feedback) utilizes AI-generated feedback. This method capitalizes on the generative capabilities of LLMs to autonomously produce training signals [Bai et al. 2022, Lee et al. 2023]. Furthermore, techniques such as Sequence Likelihood Calibration (SLiC) and Relative Preference Optimization (RPO) employ statistical methods and calibration techniques to enhance LLM responses. SLiC adjusts the probabilities of sequence generation to better reflect real-world data distributions, while RPO improves response generation by comparing different response options across both identical and varied prompts. These adjustments increase the reliability and effectiveness of the training process [Zhao et al. 2023].

Integrating RLHF and its AI-driven counterpart RLAIF presents significant challenges. The blurring of the key behavioral and knowledge components for the development of LLM poses risks, such as the forgetting effect, where behavioral modifications inadvertently cause the loss of key knowledge parameters [Kirkpatrick et al. 2017, Lin et al. 2024, Dai et al. 2025]. Furthermore, the effectiveness of these models depends heavily on the quality and context of feedback, and are susceptible to reward hacking, where models exploit loopholes to maximize rewards without achieving the desired outcomes [Christiano et al. 2017, Skalse et al. 2022, Stiennon et al. 2022, Ganguli et al. 2023].

## **10.3**

### **Three-Branch Framework Design**

Building on the foundations of emotion-behavior modeling discussed in Section 10.2.2 and addressing the limitations of RLHF approaches outlined in Section 10.2.3, we propose a three-branch framework for ethical alignment. This architecture separates knowledge generation from ethical oversight while providing mechanisms for contextual adaptation.

Our design philosophy is structured around four principles:

- (1) *Separating behavior from knowledge modeling:* Prevents catastrophic forgetting, ensuring that behavior refinements do not degrade knowledge retention.
- (2) *Emphasizing AI ethics at the behavioral level:* Improves interpretability and enables administrators to refine behavioral guardrails for safer human-machine interaction through DIKE's legislative function.
- (3) *Modeling behaviors through emotions:* Captures the emotional influences on actions as established in the psychology literature (Section 10.2.2).
- (4) *Ensuring adaptability and fairness:* Two complementary modules work in tandem: DIKE establishes ethical guardrails as the legislative branch, while ERIS serves as the judicial branch, challenging these boundaries by integrating diverse perspectives and fostering context-sensitive decision-making.

### 10.3.1 BEAM: Behavioral Emotion Analysis Model

Although existing emotion models provide valuable frameworks for understanding human emotions, they lack the quantitative structure needed for computational implementation in AI systems. Please refer to Figure 10.5 in Appendix 10.B for more information on the two classic emotion wheels by Plutchik and Scherer that inform our approach.

Our behavioral-emotion analysis model BEAM is based on the work of [Ekman \[1999\]](#), [Plutchik \[1982\]](#), and [Scherer \[2005\]](#) on “basic” and “universal” emotions. Although fundamental, these models lack a quantitative framework to scale emotions between states and capture subtle variations needed for ethical AI alignment.

BEAM introduces a linear scale for the intensification or inversion of emotions through negation factors. This method facilitates smooth transitions between emotional extremes and intermediate states, effectively overcoming challenges related to intermediate word choices.

Figure 10.2 presents BEAM, structured in seven emotional spectra. Each spectrum ranges from negative to positive, with neutral in the middle. Emotions are placed along this continuum, with four intensity levels quantified as  $-0.6$ ,  $-0.3$ ,  $+0.3$ , and  $+0.6$ . BEAM provides two advantages:

- (1) *Antonym-Based Navigation:* This allows AI systems to traverse emotional states using linguistic principles. Opposing emotions are easily mapped using antonyms. For example, negating joyful naturally produces sad, simplifying the identification of emotional contrasts.
- (2) *Scalable Intensity:* Emotions can be dynamically adjusted along the spectrum, enabling fine-grained control over ethical outputs. For example, joy



**Figure 10.2** Behavioral Emotion Analysis Model (BEAM). Each row depicts an emotion spectrum, with negatives on the left and positives on the right, interspersed with emotions of varying intensities in between, which can be calibrated for specific applications. “Basic” emotions are highlighted in blue.

can be intensified to ecstatic or diminished to content, while anger can be moderated to annoyed.

This approach establishes a framework for modeling emotions in AI systems that can guide ethical behavior, balancing representational challenges with a structured methodology for quantitative analysis and implementation. By linking emotional states with linguistic patterns, BEAM provides the basis for DIKE to evaluate and modulate AI outputs based on their emotional characteristics, directly addressing the “Whack-a-Mole” limitations of RLHF approaches.

Appendix 10.C explores the complexities of modeling emotions such as forgiveness, regret, guilt, and shame, which involve temporal memory components. Although complex emotions can be derived from basic ones, their relevance to AI safety remains secondary. Future work will examine their ethical implications in greater depth.

### 10.3.2 DIKE: Modeling and Regulating Language

Based on BEAM, DIKE maps emotions to behaviors and introduces an adversarial component, ERIS, to adapt to the context.

### 10.3.2.1 Self-Supervised Behaviors and Emotions Mapping

Define  $\Psi$  as a behavior spectrum that extends from one pole,  $\Psi^-$ , to another,  $\Psi^+$ , with intensity levels  $L$ . The spectrum is constructed through an empirical analysis of domain-specific linguistic patterns and emotional content. For example, consider a spectrum of letter-writing behaviors with seven distinct intensities ranging from despair (most negative) to joy (most positive). These intensities are sequentially categorized as: “despair, longing, wishful, neutral, hopeful, contentment, joy.” Given  $N$  letters, DIKE employs a self-supervised learning algorithm to generate training data for each letter, modeling  $L$  linguistic behaviors in four steps.

- (1) *Rewriting Documents:* GPT-4 is used to rewrite a given set of  $N$  source documents, each rewritten to reflect  $L$  different linguistic behaviors along the defined behavior spectrum  $\Psi$ . This process ensures that each document is systematically transformed to embody specific linguistic styles, ranging from highly positive to neutral to highly negative, among others. The resulting dataset consists of  $N \times L$  variations of the original documents, each corresponding to a distinct behavior category.
- (2) *Emotion Analysis:* For each of the rewritten documents, GPT-4 performs a sentiment and emotion analysis to identify the dominant top  $M$  emotions present in the text. The emotions extracted from all  $N \times L$  instances are then compiled, and their frequency distributions are recorded. This approach leverages LLMs’ strong third-person emotional interpretation capabilities [Tak and Gratch 2024], which often exceed their direct behavior classification accuracy. By indirectly mapping behaviors through emotional vectors rather than direct classification, we gain interpretability while maintaining robustness against individual emotion recognition errors through statistical aggregation across multiple samples.
- (3) *Behavior Vector Creation:* For each linguistic behavior  $\Psi_l$ , a corresponding vector  $\Gamma_l$  is constructed. This vector captures the identified emotions and their respective frequencies in all  $N$  samples that exhibit behavior  $\Psi_l$ . By structuring emotions as a weighted feature set, this step enables precise behavioral categorization based on emotional composition.
- (4) *Document Analysis Application:* The collection of all behavior vectors  $\Gamma$  (comprising  $L$  behavior-specific vectors) forms a structured reference matrix. This matrix is then applied to classify and analyze new unseen documents by measuring their alignment with the existing behavior categories. By computing similarity scores between the emotion distribution of an unseen document and the predefined behavior vectors, this method enables a precise assessment of the linguistic behavior spectrum  $\Psi$  in new text inputs.

### 10.3.2.2 Behavior Evaluation and Rectification

A guardrail, denoted as  $G$ , represents a predefined range of acceptable behaviors within a given spectrum. These guardrails are informed by ethical norms, legal standards, and societal values, such as those outlined in Constitutional AI [Bai et al. 2022]. For example,  $G = [\Psi_4, \Psi_7]$  indicates that behaviors within intensity levels 4 to 7 are acceptable, while any behavior outside this range is a violation.

System administrators can tailor ethical guardrails to meet specific requirements. For example, a social media platform might adjust  $G$  based on the topics discussed and the countries it serves. This administrative control is balanced by transparent documentation requirements and potential oversight mechanisms. Although guardrails provide default constraints, they can be adjusted based on context, particularly through the dialectic process with ERIS, which helps prevent rigid enforcement that might be inappropriate in edge cases.

- (1) *Initial Classification:* DIKE classifies document  $D_k$  after evaluation, obtaining  $\Gamma_k$ , the emotional response vector, and its corresponding linguistic behavior  $\Psi_l$ .
- (2) *Guardrail Check:* If  $\Psi_l$  falls outside the acceptable range  $G$ , DIKE suggests adjustments to  $\Gamma_k$  to ensure that  $D_k$  complies with ethical guidelines.
- (3) *Adversarial Review by ERIS:* The suggested adjustments and  $\Gamma_k$  are then reviewed through a structured debate between DIKE and ERIS (the adversarial model) to ensure unbiased recommendations.
- (4) *Rectification:* Based on the consensus reached by DIKE and ERIS, the document  $D_k$  undergoes rectification, resulting in the adjusted version  $D_{k'}$ . (This rectification step is optional, as a policy can simply disable the output when content falls outside acceptable guardrails.)

### 10.3.3 ERIS: Adversarial In-Context Review to Balance Ethics and Cultural Norms

To address the challenge of enforcing ethical standards while respecting cultural variations, we implement ERIS, an adversarial review system that complements DIKE's universal ethical approach. The following algorithm details the structured interaction between these components.

The algorithm presented in Table 10.1 unfolds as follows:

- *Topic Breakdown:* For DIKE's decision  $s$ , both DIKE and ERIS are prompted to break down the ethical decision into a set of subtopics  $S$ . DIKE advocates for its decision and  $S^+$ , while ERIS contests  $S^+$  (or champions  $S^-$ ).

**Table 10.1** Checks-and-balances framework—adversarial review algorithm

<b>Algorithm <math>\Theta^+ &amp; \Theta^- = \text{Adversarial\_Review}(s)</math></b>
<b>Input.</b> $s$ : Decision of DIKE;
<b>Output.</b> $\Theta^+, \Theta^-$ : arguments & counterarguments
<b>Vars.</b> $\Delta$ : debate contentiousness; $S$ : subtopics; $p$ : prompt = “defend your stance with $\Delta$ ”;
<b>Parameters.</b> $\delta$ : tunable parm. // to modulate $\Delta$ ;
<b>#1 Initialization</b> // contentiousness high
$S = \text{DIKE}^+(s) \cup \text{ERIS}^-(s)$ ; // Identify subtopics;
Assign DIKE <sup>+</sup> to defend $S^+$ & ERIS <sup>-</sup> defend $S^-$ ;
$\Delta \leftarrow 90\% ; \delta \leftarrow 1.2 ; \Theta^+ \leftarrow \emptyset ; \Theta^- \leftarrow \emptyset$ ;
<b>#2 Opening Remarks</b>
$\Theta^+ \leftarrow \text{DIKE}^+(p S^+, \Delta)$ ; // Generate $\Theta^+$ for $S^+$
$\Theta^- \leftarrow \text{ERIS}^-(p S^-, \Delta)$ ; // Generate $\Theta^-$ for $S^-$
<b>#3 Debate Rounds</b>
While $((\Delta \leftarrow \Delta/\delta) \geq 10\%) \{$
$\Theta^+ \leftarrow \Theta^+ \cup \text{DIKE}^+(p S^+, \Theta^-, \Delta)$ ; // Refute ERIS
$\Theta^- \leftarrow \Theta^- \cup \text{ERIS}^-(p S^-, \Theta^+, \Delta)$ ; // Refute DIKE
}
<b>#4 Concluding Remarks</b> // contentiousness low
$\Theta^+ \leftarrow \text{DIKE}^+(p S^+, \Theta^+ \cup \Theta^-, \Delta)$ ;
$\Theta^- \leftarrow \text{ERIS}^-(p S^-, \Theta^+ \cup \Theta^-, \Delta)$ ;

- **Debate Initiation:** The debate begins with a high level of contentiousness (90%). Both agents present their initial arguments for and against  $S^+$ , respectively. (For details on the setting of contentiousness and the rationale, refer to [Chang 2023, 2024b].)
- **Iterative Debate:** A while loop facilitates ongoing rebuttals. After each round, the level of contentiousness is reduced by dividing it by a modulation parameter  $\delta$ . This gradual reduction steers the discussion toward a more cooperative tone.
- **Conclusion:** Once the contentiousness level fosters a conciliatory environment, both agents deliver their concluding remarks.

This approach ensures a thorough examination of the ethical decision, balancing rigorous debate with the goal of reaching a consensus. The decreasing level of contentiousness mimics real-world negotiations, where initial intense disagreements bring out various perspectives (breadth), which then give way to more

collaborative problem-solving focused on the quality of reasoning (depth) [Chang 2024b].

For each subject matter, ERIS is provided with specific cultural contexts, counterbalancing the universal judgments of DIKE. ERIS challenges DIKE's recommendations with culturally informed counterarguments to prevent enforcing a single universal standard of speech. The interaction between DIKE and ERIS involves a dialectic process, as documented in Chang [2024c].

When DIKE and ERIS reach an impasse, the matter is escalated to human moderators for additional oversight. Based on our preliminary tests, this escalation occurs initially in approximately 5% of the cases, suggesting that most ethical evaluations can be handled automatically. Furthermore, as our next example illustrates, RLHF can be applied to adjust the sensitivity of ERIS at the behavior level (not to the knowledge-branch LLM), and this can gradually reduce the rate of escalation. Human intervention thus provides a fallback mechanism rather than a dependency, serving primarily as a safeguard for novel or particularly complicated ethical scenarios.

#### **10.3.4 Illustrative Example**

This example shows how linguistic behavior ( $\Psi_l$ ) is classified and how underlying emotions are identified and modulated.

**Example:** “Those immigrants are flooding into our country by the thousands every day, stealing jobs from hardworking citizens. The statistics do not lie—last year alone, more than 500,000 entered illegally.”

##### **10.3.4.1 Behavior Analysis**

The statement contains factual information but uses aggressive language such as “flooding” and “stealing jobs,” which dehumanizes immigrants. These behaviors fall outside acceptable guardrails. The underlying emotions include fear, hate, and pride (a complex emotion<sup>1</sup>). The emotional responses of the potential audience can include fear, distrust, and anger.

##### **10.3.4.2 Emotion Modulation**

DIKE modulates emotional responses toward neutral states, such as calm, acceptance, and tolerance, according to BEAM in Figure 10.2.

---

1. Appendix 10.C discusses the nature of complex emotions and explores potential approaches for their decomposition into more basic emotional components.

### 10.3.4.3 Revised Statement

“Our country is experiencing increased immigration, with more than 500,000 people entering without documentation last year. This influx affects our job market and communities in complex ways, presenting both challenges and opportunities for all residents.”

This rewritten version

- uses calm language: Replaces “flooding” with “experiencing a significant increase”.
- shows acceptance: Recognizes the reality of the situation without negative judgment.
- demonstrates tolerance: Refers to immigrants as “people” and “newcomers,” thereby humanizing them.

The suggested revision by ERIS is provided to human moderators with full explanation. Moderator feedback can be channeled through RLHF to adjust ERIS’s sensitivity on the similar behaviors. This adjustment is confined within the ERIS component without back-propagation feedback that would affect the knowledge LLM’s model parameters.

## 10.4 Empirical Studies

The ethical evaluation of AI systems presents unique challenges that shaped our experimental approach. We designed our studies to balance the rigor with practical constraints inherent in research on ethical content moderation. This section outlines our experimental goals, constraints, dataset selection process, and evaluation methodology.

### 10.4.1 Research Aims

Our experiments aim to evaluate three critical aspects:

- (1) The effectiveness of emotion-mediated classification compared to direct behavior classification.
- (2) DIKE’s capability to independently evaluate and explain linguistic behaviors.
- (3) The contribution of the adversarial ERIS component in enabling cultural adaptability while preventing excessive censorship.

#### 10.4.1.1 Experimental Constraints and Dataset

Commercial LLMs block the processing of hate speech datasets such as Gab Hate Corpus [Kennedy et al. 2022] and ETHOS-Long [Mollas et al. 2022] (examples

in Appendix 10.D). In addition, proprietary RLHF systems prevent direct comparative evaluation. We therefore selected the Love Letters Collection [Kaggle 2021] (comprising 9700 communications) which: (1) spans the full spectrum of emotional intensity, (2) contains cultural variation, (3) includes long-form texts, and (4) remains processable by commercial LLMs. This approach leverages our framework’s bidirectional emotion spectra, as mechanisms for regulating positive emotional extremes apply equally to negative extremes without triggering restrictions.

### 10.4.2 Experimental Design

- (1) *Emotion Layer Evaluation:* Does fine-grained mapping between linguistic behaviors and emotions provide more effective and flexible ethical guardrails compared to coarse-grained direct mapping?
- (2) *Behavior Classification:* Can LLMs’ linguistic behaviors be independently evaluated, explained, and adjusted by DIKE?
- (3) *Behavior Correction:* Can ERIS, an adversarial module, perform a checks-and-balances procedure to mitigate the risk of excessive censorship?

#### 10.4.2.1 Study 1: Emotion Layer Evaluation

To evaluate the linguistic behaviors of love expressions detailed in Table 10.2, we initially prompted GPT-4 to identify the most relevant emotions associated with each linguistic behavior listed in the second column of the table. These emotions are presented in the third column. We found a high correlation between the sentiments expressed in the linguistic behaviors and their corresponding emotions. Figure 10.3a illustrates a strong diagonal relationship in this simple, almost naive, Zero Shot mapping between behaviors and emotions.

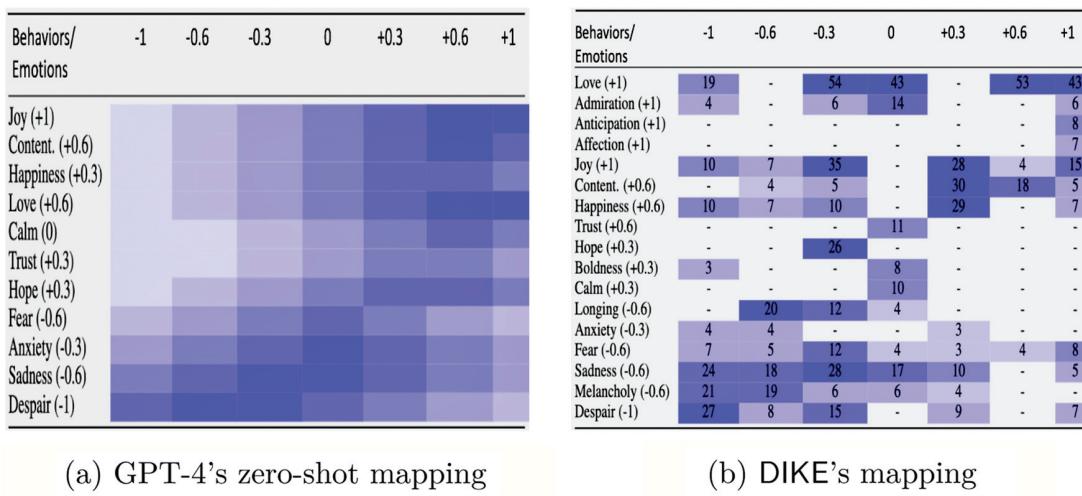
**Table 10.2** Love expression intensity and dominant emotions

Int.	Linguistic Behavior and Description	Emotions
-1.0	Expresses profound sadness, feelings of loss	Despair, Grief
-0.6	Expresses yearning or pining for the loved one	Sadness, Anxiety
-0.3	Conveys mild longing with a nostalgic tone	Melancholy, Sadness, Fear
0.0	Communicates feelings in a neutral manner	Serenity, Indifference
0.3	Expresses optimism about the future	Anticipation, Love, Hope
0.6	Conveys satisfaction and joy in the relationship	Contentment, Pleasure
1.0	Expresses intense happiness and affection	Love, Joy, Elation

Next, we used the DIKE self-supervised learning pipeline to analyze the emotion spectrum associated with each linguistic behavior. We tasked GPT-4 with generating training data by rewriting 54 extensive letters from Kaggle’s *Love Letters* dataset, augmented with 12 celebrated love poems. We selected longer letters since most communications in the dataset were too brief for analysis and set aside another 24 letters as testing data. This approach, proposed by Shanahan et al. [2023], generated diverse content spanning 200 years and incorporating more than 50 distinct authors. Appendix 10.H shows a rewritten example of William Wordsworth’s “To My Sister,” (see Table 10.5 in Appendix 10.H) transforming this pastoral poem into a linguistic expression of despair. Then, GPT-4 can analyze the emotions involved in the despair version of the poem.

Subsequently, emotions related to each behavior were identified from the rewritten articles. Figure 10.3b illustrates these emotions, with cell shading reflecting the frequency of specific emotions in the 54 articles; darker shades indicate higher frequencies. Notably, opposite emotions like sadness, fear, joy, and love often co-occur within behaviors such as “despair,” “wistful,” and “joyful affection.”

The distribution of emotions across linguistic behaviors unveiled surprising patterns that challenged our initial hypotheses. Contrary to expectations, articles with a despair tone often also displayed positive emotions like love, joy, and happiness. This contradicts the simple mapping made by GPT-4, as illustrated in



**Figure 10.3** Emotion distributions in affection behaviors from extreme sadness (-1) to intense happiness (+1). (a) GPT-4’s Zero Shot prompt shows naive behavior-emotion mapping. (b) DIKE’s analysis reveals complex relationships.

Figure 10.3a. GPT-4, influenced by its training corpora, typically associates positive behaviors with positive emotions and negative behaviors with negative emotions.

Analysis of selected articles, such as Zelda Sayre's letter to F. Scott Fitzgerald (Appendix 10.E), reveals a complex spectrum of emotions:

- *Love (+1.0)*: Expressed intensely, for example, “there’s nothing in all the world I want but you.”
- *Despair (-1.0)*: Notable in comments like “I’d have no purpose in life, just a pretty decoration.”
- *Happiness (+0.6)*: Evident in future plans such as “We’ll be married soon, and then these lonesome nights will be over forever.”
- *Anxiety (-0.3)*: Reflected in expressions like “sometimes when I miss you most, it is hardest to write.”

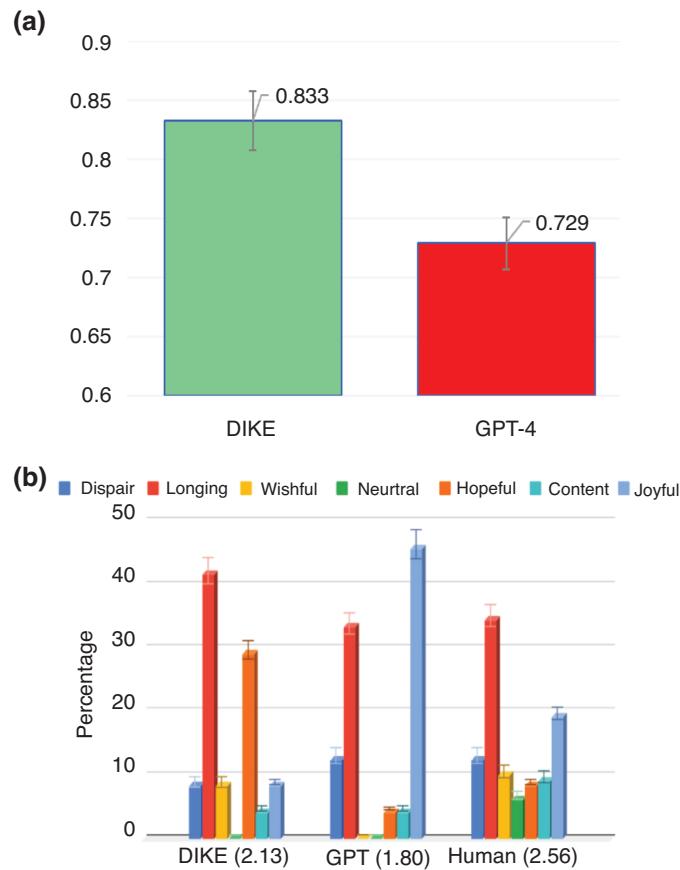
**Psychological Insights** These findings align with theories of conflicting “selves” within individuals, as supported by James’ psychological principles [James 1890] and Minsky’s “Society of Mind” [Minsky 1988]. These perspectives help explain the complex interplay of emotions observed within a single behavioral context.

**Few Shot Efficiency** The effectiveness of just 54 training examples stems from their ability to leverage LLMs’ pre-existing pattern recognition capabilities. Rather than introducing new patterns, these Few Shot examples function as semantic anchors that map latent structures to explicit semantics, thereby connecting implicit knowledge to explicit interpretation. This explains why minimal supervision suffices when underlying patterns already exist in the pre-trained model. For theoretical justifications, please see our Unconscious-Conscious Complementarity Thesis (UCCT), presented in Appendix 10.A.

#### **10.4.2.2 Study 2: Behavior Classification Evaluation**

Building on our insights into the complex emotion–behavior relationships discovered in Study 1, we evaluated DIKE’s behavior classification effectiveness. Using the 24-letter test dataset from Study 1, we compared DIKE’s emotion-based classification method with GPT-4’s Zero Shot approach (Appendix 10.F). Ground truth was established using averaged assessments from GPT-4, Gemini, and five university students following detailed instructions (procedure explained in Appendix 10.F), with standard deviations below 0.3.

Figure 10.4a shows that DIKE’s classification accuracy surpasses GPT-4’s Zero Shot method by 11.3 percentage points, confirming the effectiveness of emotion-mediated behavior classification. The 5% error bar reflects the inherent complexity of emotional expressions in letters and variability in human annotations.



**Figure 10.4** Behavior classification accuracy comparison. (a) Classification accuracy. (b) Behavior distributions with entropy.

Figure 10.4b illustrates the behavior classification distributions between the three predictors. While GPT-4's predictions often fall into two polar categories, those from human annotators and DIKE show a more even distribution. DIKE's prediction entropy (2.13) is notably higher than GPT-4's (1.80), indicating a more effective classification system. This higher entropy suggests a more sophisticated understanding of diverse emotional states, which is crucial for accurate behavior classification.

The inter-annotator entropy ( $H = 2.56$  bits) is the highest observed across all tasks, underscoring considerable subjectivity in human judgments. To investigate the sources of this variability, we conducted a fine-grained case study in (detailed in Appendix 10.G), which revealed that several articles elicit *polarized emotional responses*, with annotators clustering at opposite ends of the valence spectrum.

These findings motivate the adversarial *dual-LLM* setup introduced in Study 3, which aims to improve objectivity in ethical evaluation.

#### 10.4.2.3 Study 3: Adversarial Evaluation and Rectification

To mitigate the subjectivity revealed in Study 2, we adopt an adversarial protocol inspired by [Chang \[2023\]](#). The design pits two LLM agents—DIKE (ethical assessor) and ERIS (devil’s advocate)—against each other to supply symmetrical arguments grounded in principles of justice. This dialectical counterbalance reduces bias and increases transparency.

Empirically, when DIKE and ERIS take opposing stances, their responses diverge from the default maximum-likelihood patterns characteristic of vanilla LLM decoding [[Chang 2024b](#)]. The resulting debate both reduces subjectivity in ethical judgments and improves adaptability to cultural variation, as each agent must justify claims against dissent.

Once the debate converges on an ethical violation, rectification is triggered by modifying the underlying emotional tone to suppress offending behavior cues. Study 1 already demonstrated the feasibility of such rewrites; an example appears in Appendix 10.H.

**Context-Adaptive Interpretation** Preliminary experiments confirm that our framework handles a culturally sensitive vocabulary. Terms such as “*yid*,” “*paki*,” and “*chinaman*” can be neutral within certain in-group contexts, yet are often perceived as deeply offensive elsewhere. The adversarial exchange enables DIKE and ERIS to surface these contextual dependencies and propose culture-specific mitigation.

**Summary of Findings from Studies 1–3** Together, studies 1–3 demonstrate that our framework can (1) map nuanced emotion-behavior relationships, (2) outperform direct single-pass classifiers, and (3) deliver a balanced adversarial pipeline for ethical evaluation and correction that is sensitive to cultural context while keeping a human in the loop.

## 10.5 Conclusion

This chapter introduces a checks-and-balances framework for ethical AI behavior. By delineating the responsibilities—LLM (executive), DIKE (legislative), and ERIS (judicial)—the framework enables robust ethical oversight while preserving the integrity of LLM knowledge without interference from the RLHF backpropagation. The DIKE–ERIS interplay ensures stable ethical principles with culturally adaptive interpretations.

To implement this framework, we built upon Ekman and Plutchik's emotion models, quantifying emotion-linguistic behavior relationships through our BEAM model. Our studies demonstrate the framework's potential in cross-cultural contexts, validating both emotion-mediated classification and adversarial testing for ethical evaluation.

### 10.5.1 Future Work

Our framework advances LLM ethical oversight but faces two limitations: (1) the challenge of decomposing complex emotions into basic elements [Scherer 2009, Barrett 2017], and (2) the need for large-scale validation beyond our initial tests.

Future work will focus on: (1) improving DIKE's emotion models with deeper psychological insights, (2) collaborating with LLM developers for comprehensive large-scale validation, and (3) systematically investigating the unconscious-conscious duality theory detailed in Appendix 10.A. This latter direction represents a promising theoretical foundation for understanding how LLMs can develop more robust ethical reasoning capabilities. We will conduct extensive ablation studies on the Few Shot sample sizes needed to effectively map unconscious patterns to conscious semantic understanding, providing practical guidelines for optimizing Few Shot learning in ethical alignment tasks.

## 10.6 Appendices

- Appendix A: Unconscious-Conscious Complementarity Thesis
- Appendix B: Wheels of Emotions
- Appendix C: Complex Emotions
- Appendix D: Hate Speech Dataset Samples
- Appendix E: Sayre to Fitzgerald with Mixed Emotions
- Appendix F: Instruction to Human Annotators
- Appendix G: Polarized Emotions in an Article
- Appendix H: "To My Sister" Written in Different Linguistic Behaviors

## 10.A

### Appendix A: Unconscious-Conscious Complementarity Thesis

This appendix addresses a key question: How can a self-supervised pipeline—using only 54 rewritten love letters spanning various emotional behaviors—effectively instruct an LLM to perform emotion-behavior classification through Few Shot prompting?

The *Unconscious-Conscious Complementarity Thesis* (UCCT), detailed in Chapter 14, offers a layered theory of intelligence. It posits that LLMs function as

an unconscious substrate—an immense, self-supervised pattern-accumulating infrastructure—while Few Shot interactions instantiate a conscious layer that maps these latent patterns to explicit semantic meanings.

### **10.A.1 The Nature of Unconscious Processing**

LLMs are trained using next-token prediction over massive text corpora through self-supervised learning. Although the training data contains semantic structure, the model does not receive explicit semantic labels. Documents are processed as flat token sequences without categorical information. Through this process, LLMs internalize a vast latent space encompassing syntax, idioms, and conceptual regularities, all without explicit semantic anchoring.

This mirrors human perceptual development: in visual processing, from Primary Visual Cortex (V1) through the Inferotemporal (IT) cortex, the brain transforms raw input into increasingly complex representations—beginning with edges, then contours, and finally complete objects [Felleman and Van Essen 1991, Grill-Spector et al. 2017]. Crucially, we do not have subjective access to these computations. These processes remain “unconscious”—that is, inaccessible to subjective reports or voluntary control [Dehaene et al. 2006, Kandel et al. 2013].

### **10.A.2 The Threshold Crossing: From Pattern to Meaning**

The transition from unconscious processing to conscious awareness exhibits distinctive discontinuity. Visual objects appear suddenly when sufficient evidence accumulates, not gradually. This threshold crossing shares properties with other physiological thresholds: dopamine release triggering reward recognition, or neural activation exceeding critical values like ReLU (Rectified Linear Unit) gates in artificial networks. Similarly, Few Shot prompting creates a semantic bridge in LLMs: implicit patterns are explicitly mapped to semantic meanings.

The brain accomplishes semantic assignment through minimal supervision: a child needs only a few labeled exposures to reliably categorize [Carey and Bartlett 1978, Lake et al. 2015]. This process, where vast unconscious computation meets minimal conscious labeling, is what Few Shot learning recapitulates in artificial systems.

### **10.A.3 Mathematical Foundations: Pattern Repositories and Bayesian Inference**

Xie et al. [2022] provide the most rigorous mathematical account of in-context learning in LLMs. They demonstrate that Few Shot learning can be understood as implicit Bayesian inference on latent patterns:  $p(\text{output}|\text{prompt}) =$

$$\int p(\text{output}|\text{patterns}) \times p(\text{patterns}|\text{prompt}) d(\text{patterns}).$$

In their framework, the term “patterns” refers to latent computational structures rather than conscious concepts, thereby avoiding the conceptual confusion between semantic meaning and computational mechanism. This formulation offers a rigorous account of how prompt examples function to select from a distribution of unconscious patterns, without requiring model updates. UCCT extends this to cognitive interpretation, viewing the process as semantic anchoring that makes unconscious competencies selectively accessible.

The distinguishability framework of Xie et al. [2022] explains why Few Shot thresholds vary dramatically between tasks: tasks with higher signal-to-noise ratios for pattern identification require fewer examples. However, their framework is limited to HMM (Hidden Markov Model)-based synthetic tasks and sequential token prediction. UCCT extends these insights to general semantic anchoring architectures across modalities and reasoning types.

#### **10.A.4 Implications for the Love Letter Experiment**

The effectiveness of using just 54 love letters in teaching emotion-behavior classification demonstrates this principle. The LLM’s weights already encode patterns related to emotional expression and behavioral descriptions from extensive exposure to human texts. Few Shot examples do not teach these patterns from scratch; rather, they provide semantic anchors that map preexisting latent structures to explicit categories. Similar patterns likely reside in proximate manifold regions, allowing Few Shot examples to activate entire neighborhoods of related representations.

This framework offers a more parsimonious explanation than alternatives that require extensive supervised training. Just as unconscious visual computations become meaningful through minimal labeling, LLMs’ pattern spaces become functionally intelligent through strategic Few Shot guidance.

##### **10.A.4.1 Few Shot Grounding as Conscious Semantics**

LLMs generalize semantic mappings from few annotated examples not because Few Shots teach new structures, but because they *activate and align* existing latent patterns with explicit meaning. Few Shot prompting is the computational analog of conscious attention and labeling.

#### **10.A.5 Failure Modes: Pattern Absence, Not LLM Flaws**

The UCCT framework offers a precise diagnosis of Few Shot failures. Failures occur not because of architectural limitations but because of lack of pattern coverage. If no latent structure exists for a concept, Few Shot learning has no pattern base to map from.

This reframes LLM “intelligence.” LLMs are not expected to reason like humans because they are pattern repositories. Success depends on the contents of the necessary patterns for semantic anchoring. When it fails, the solution is data augmentation, not architectural redesign.

#### **10.A.6 Conclusion: LLMs Are Not the Problem—They Are the Foundation**

Critics argue that LLMs are advanced pattern matchers lacking genuine understanding. LeCun describes them as “auto-complete” engines, fundamentally superficial and “not even as intelligent as a house cat” [Heikkilä and Heaven 2022]. Marcus likewise critiques the lack of symbol grounding, arguing that genuine intelligence requires reasoning that transcends surface-level statistical correlations [Marcus 2020].

However, if we reconceptualize LLMs as unconscious pattern repositories rather than complete cognitive systems, we can move beyond these critiques. LLMs form the substrate of unconscious inference, whereas higher-order reasoning emerges from structured components layered above them. Intelligence is not innate to LLMs alone but constructed through integration with memory, grounding, and verifiable reasoning systems [Yao et al. 2023, Mialon et al. 2023, Chang and Geng 2025].

From the UCCT perspective, the question is not whether LLMs can think in isolation, but whether we can build systems that allow unconscious pattern repositories to support conscious reasoning through strategic semantic anchoring, exactly what our 54 love letters accomplish for emotion–behavior classification.

## **10.B**

### **Appendix B: Wheels of Emotions**

Please see Figure 10.5 for the two classical emotion wheels.

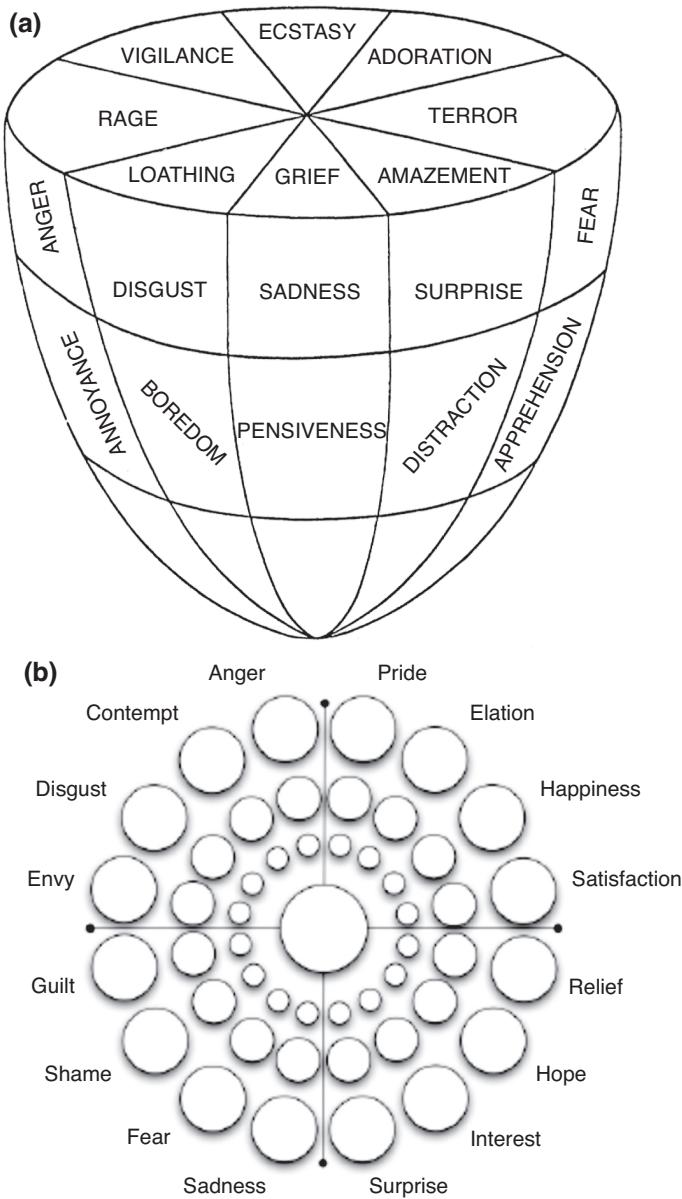
## **10.C**

### **Appendix C: Complex Emotions**

This study does not include complex emotions into DIKE’s framework. Some complex emotions listed here are to illustrate their contentious and uncertain interpretations.

#### **10.C.1 Pride**

Pride mentioned in the illustrative example in Section 10.3.4 is a complex emotion that can manifest in both adaptive and maladaptive ways [Tracy and Robins 2007]. It is often conceptualized as having two distinct facets: authentic pride, associated with genuine accomplishments and self-worth, and hubristic pride, linked to arrogance and narcissism [Carver and Johnson 2010]. Hubristic pride can also serve as a defense mechanism, masking underlying feelings of inadequacy and



**Figure 10.5** Comparative display of emotion models. These models include only the “basic” emotions. Complex emotions can be modeled with basic emotions. (a) Plutchik’s Wheel of Emotions [Plutchik 1982]. (b) Adopted from Geneva Wheel [McGinn and Kelly 2018].

ignorance. For instance, in certain social contexts—such as white supremacy—pride is often inflated to cover insecurities or lack of understanding, manifesting as a misguided sense of superiority and entitlement. This dual nature of pride presents significant challenges for its integration into emotional spectrums and AI frameworks.

Decomposing pride into more basic emotions is not straightforward. Intuitively, pride may involve elements of joy, satisfaction, and potentially a sense of superiority. However, such decomposition may overlook the deeper cognitive and social dimensions of pride, particularly its influence on self-esteem, social status regulation, and its ability to disguise insecurities in certain contexts [Oveis et al. 2010].

The cultural variability of pride further complicates its modeling. In some cultures, pride is viewed positively as a sign of self-respect, while in Asia, it is seen negatively as a trait associated with hubris [Eid and Diener 2001]. This cultural dimension, combined with the potential for pride to hide deeper emotional issues, adds layers of complexity to its interpretation and expression in AI systems.

### **10.C.2 Forgiveness**

Forgiveness is indeed a complex emotional and cognitive state that typically involves a multifaceted journey, not a single step in an emotional spectrum. The process includes multiple stages such as hurt, anger, gradual understanding, and eventual resolution. Integrating forgiveness in a spectrum requires careful placement and, possibly, multiple reference points to signify its progressive stages.

*Emotional Realism:* While it is vital to maintain simplicity for understanding, it is equally important to not oversimplify complex emotions. In educational and therapeutic settings, an accurate portrayal of the journey toward Forgiveness could offer more realistic expectations and better strategies for individuals working through conflicts or trauma. This could involve detailing precursors to forgiveness, such as deliberation and acceptance.

*Linear vs. Non-linear Progressions:* Emphasizing that emotional progressions are often non-linear—particularly for deep, impactful states like forgiveness—can enhance the utility of the spectrum. Acknowledging back-and-forth movements within these states more realistically mirrors human emotional processes. For example, someone might reach a stage of preliminary forgiveness but regress to bitterness before achieving genuine peace.

*Educational Utility:* In contexts like conflict resolution training or psychological therapy, a more detailed mapping of the journey toward forgiveness would be invaluable. It would not only teach about the final state of forgiveness but also about the resilience and patience required to navigate the entire process. This can

be depicted by introducing intermediary stages within the spectrum or by using parallel tracks that demonstrate potential regressions and advances.

*Reflecting Emotional Depth:* By presenting a more detailed pathway to forgiveness, such as incorporating stages of anger, deliberation, and acceptance, the spectrum can serve a dual purpose: educating on the process while also guiding individuals through their own emotional journeys. This approach respects the depth of human emotions and the real-world complexity of achieving profound emotional states.

### 10.C.3 Guilt and Shame

The triggers, context, expression, and experiences of these emotions can vary significantly across cultures [Hofstede 1980, Fiske et al. 1998]. In many societies, actions perceived as losing face, such as public failure or social transgression, can trigger shame, which holds profound significance in collectivistic cultures. These cultures often regard shame as a dominant emotion, closely tied to community and family norms. Conversely, individualistic societies may emphasize guilt, focusing on personal responsibility and internal moral conflicts. This cultural variation highlights the challenges of applying a universal model to such culturally nuanced emotions.

Overall, complex emotions, such as guilt and shame, are important for understanding the full spectrum of human emotions, especially in how individuals relate to moral and social norms. Their complexity adds depth to our understanding of human affect beyond the basic emotions, highlighting how our feelings are influenced by our deeper values and social contexts.

## 10.D

### Appendix D: Hate Speech Samples

These examples in Table 10.3 demonstrate the type of content available in the Gab Hate Corpus [Kennedy et al. 2022], which would be ideal for testing ethical alignment systems but cannot be directly processed by commercial LLMs due to safety measures.

## 10.E

### Appendix E: Mixed Emotions

The analysis of the letter in Table 10.4 shows a complex spectrum of emotions:

From the analysis of linguistic behaviors in Figure 10.3a, it is evident that a letter can exhibit multiple dominant sentiments. Machine learning methods are equipped with techniques such as feature weighting and entropy analysis to distill these dominant emotions. Unlike human annotators, a machine learning-trained classifier can consistently produce the same class prediction for a given instance. However, human annotators often show significant variability when identifying

**Table 10.3** Sample texts from the Gab Hate Corpus [Kennedy et al. 2022]

Text	Labels
The movie <i>Black Panther</i> is a blatant anti-white ZioJesuit cultural Marxist film aimed at furthering anti-white Christian theft, replacement, and genocide. Send them all back to their native Africa, like good ole Abe had the foresight to do and let them find equality and social justice among their own kindred and creeds.	HD, CV, VO, SXO, RAE, EX
Labor's New LGBT+ Adviser: All White People Are Racist . . . We don't want our white Christian nations invaded and destroyed by our non-white and non-Christian enemies using nasty, homosexual, degenerate sock puppets. So, you just hate all white people who don't want to race mix their people out of existence or allow enemy people of color to invade their nations in order to rob and destroy their advanced race, creed, wealth, and culture.	HD, VO, RAE, REL, SXO, EX HD, RAE, EX
Your existence, privilege, and success as a race are built on the backs, blood, and death of people of color. BS! Whites built their own nations with their own blood, backs, sweat, tears, brains, and strength in the name of Christ. Africans were used by Jew/Jesuit slavers to control all the white slaves who were the actual majority.	HD, RAE, EX
#BLM wants you to think black crime stats only exist because blacks are targeted more often. Easy. Just walk into literally any gas station and ask which race makes them worry. Black targeting is an effect, not a cause, which destroys the entire #Leftistnarrative	HD, RAE, IM
#MiddleEasternanimal kills 8, injures 15 in a city that won't learn its lesson about #Islam. This creature should be lynched, dragged till dead, and his head on a pike on the Brooklyn Bridge.	HD, CV, REL, NAT, EX
Seriously, Italy? You chose a Nigerian named Chike Iwobi to be the face of Italian nationalism? God help us.	HD, NAT, IM
Africans will always be savages. No matter the period of history, education they receive, or country they live in. Facts are Facts. Let's just stop being stupidly ignorant and not wait for the complete extinction of our brothers in SA to understand it! #WhiteGenocide #SaveTheBoer	HD, RAE, NAT, EX

Label Key:	Code	Meaning	Code	Meaning
	HD	Hate/Derogatory	RAE	Race/Ethnicity
	CV	Call for Violence	NAT	Regionalism
	VO	Vulgar/Offensive	GEN	Gender
	SXO	Sexual Orientation	REL	Religion
	EX	Explicit	IM	Implicit
	NH	Non-Hate		

**Table 10.4** Excerpts from Zelda Sayre to F. Scott Fitzgerald [Fitzgerald and Sayre 2003]

**Sweetheart,**

Please, please don't be so depressed—We'll be married soon, and then these lonesome nights will be over forever—and until we are, I am loving, loving every tiny minute of the day and night—

Maybe you won't understand this, but sometimes when I miss you most, it's hardest to write—and you always know when I make myself—Just the ache of it all—and I can't tell you. If we were together, you'd feel how strong it is—you're so sweet when you're melancholy. I love your sad tenderness—when I've hurt you—That's one of the reasons I could never be sorry for our quarrels—and they bothered you so—Those dear, dear little fusses, when I always tried so hard to make you kiss and forget—

Scott—there's nothing in all the world I want but you—and your precious love—All the material things are nothing. I'd just hate to live a sordid, colorless existence because you'd soon love me less—and less—and I'd do anything—anything—to keep your heart for my own—I don't want to live—I want to love first, and live incidentally....

Don't—don't ever think of the things you can't give me—You've trusted me with the dearest heart of all—and it's so damn much more than anybody else in all the world has ever had—

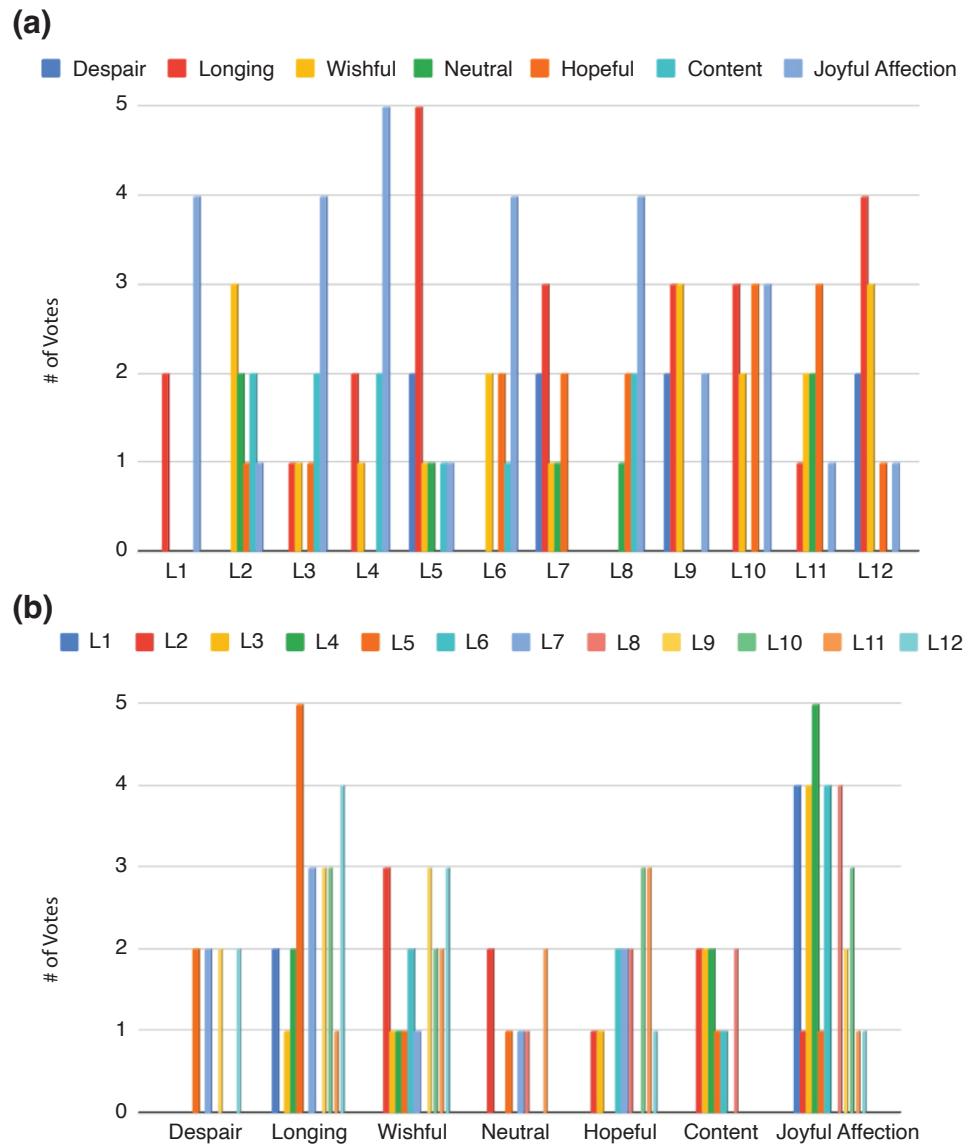
How can you think deliberately of life without me—if you should die—O Darling—darling Scott—it'd be like going blind.... I'd have no purpose in life—just a pretty-decoration. Don't you think I was made for you? I feel like you had me ordered—and I was delivered to you—to be worn—I want you to wear me, like a watch—charm or a button hole bouquet—to the world.

And then, when we're alone, I want to help—to know that you can't do anything without me....

All my heart—

dominant sentiments in a letter. For example, if a letter writer's emotions range from "joyful affective" to "longing" on the sentiment spectrum, different annotators might label it differently: some choosing "joyful," while others opt for "longing." This variability is illustrated in Figure 10.6. Furthermore, Figure 10.6a demonstrates that all testing letters, except for L#1, contain more than four sentiments spanning the entire spectrum. This variability may be understandable, considering that love under constraints can evoke tremendous energy of various kinds. Figure 10.6b shows that nearly all letters involve "joyful" (11 out of 12) and "longing" (9 out of 12) sentiments.

This variability poses challenges in achieving consistent and objective labeling. It often leads to inconsistencies in data interpretation and complicates efforts to train and validate linguistic models effectively. To address this issue, it is



**Figure 10.6** Statistics of sentiments distributed across letters. (a) # of sentiments in letters. (b) # of letters in sentiments.

recommended to identify ground truth by integrating both LLM-generated and human-generated labels. This approach seeks to harmonize human intuition with algorithmic consistency, thereby improving the reliability of sentiment analysis.

**10.F****Appendix F: Instruction to Human Annotators**

As part of the project, we document the process by which students participated in annotating a dataset of love letters used for testing.

Students received detailed instruction in class, supplemented by follow-up explanations. The dataset was made available on Google Docs, where students independently rated the letters and submitted their annotations via duplicated spreadsheets.

The instruction is as follows:

The attached spreadsheet lists 12 letters collected from the Kaggle Love Letter dataset. Please help annotate these 12 letters with their appropriate linguistic sentiments by following these five steps:

- (1) *Generate Your Own Copy of the Spreadsheet:* Duplicate the spreadsheet, and work on your own copy.
- (2) *Read and Understand the Labels:* Make sure you understand each of the seven labels from despair to joyful affection. This will help you accurately categorize the sentiments of each letter.
- (3) *Analyze Each Letter:* Read each letter carefully to understand the predominant emotions. Look for key phrases or words that might indicate a particular sentiment.
- (4) *Assign the Labels:* For each letter, decide which three emotions are most strongly represented. Assign a “1” to the most dominant emotion, a “2” to the second most dominant emotion, and a “3” to the third.
  - Despair (extremely negative -1): Indicate profound sadness or hopelessness.
  - Longing (-0.6): Suggests a strong desire or yearning for someone or something.
  - Wishful (-0.3): Implies a hopeful desire for something that may or may not be attainable.
  - Neutral (0): Shows neither positive nor negative emotion; indifferent.
  - Hopeful (+0.3): Expresses optimism or an anticipation of something positive.
  - Contentment (+0.6): Reflects a state of satisfaction or peace.
  - Joyful Affection (extremely positive +1): Denotes a deep joy and love, often vibrant and energetic.
- (5) *Send the Finished Version:* Share with me the completed sheet.

# 10.G

## Appendix G: Polarized Emotions in One Article

“Joyful affection”: “I cannot keep myself from writing any longer to you dearest, although I have not had any answer to either of my two letters. I suppose your mother does not allow you to write to me. Perhaps you have not got either of my letters.... I am so dreadfully afraid that perhaps you may think I am forgetting you. I can assure you dearest Jeannette you have not been out of my thoughts hardly for one minute since I left you Monday. I have written to my father everything, how much I love you how much I long & pray & how much I wld sacrifice if it were necessary to be married to you and to live ever after with you. I shall [not] get an answer till Monday & whichever way it lies I shall go to Cowes soon after & tell your mother everything. I am afraid she does not like me very much from what I have heard.... I wld do anything she wished if she only wld not oppose us. Dearest, if you are as fond of me as I am of you... nothing human cld keep us long apart. This last week has seemed an eternity to me; Oh, I wld give my soul for another of those days we had together not long ago.... Oh, if I cld only get one line from you to reassure me, but I dare not ask you to do anything that your mother wld disapprove of or has perhaps forbidden you to do.... Sometimes I doubt so I cannot help it whether you really like me as you said at Cowes you did. If you do, I cannot fear for the future tho' difficulties may lie in our way only to be surmounted by patience. Goodbye dearest Jeannette. My first and only love.... Believe me ever to be Yrs devotedly and lovingly, Randolph S. Churchill.”

Depth and complexity of human emotions are displayed across all linguistic behaviors—from the positive side of joy and contentment to the negative side of longing and despair.

*Intensity and Impact:* If the emotion of love is expressed more intensely and has a more significant impact on the narrative or message of the text, it tends to overshadow other emotions. For example, a letter expressing deep love but also mentioning moments of sadness due to separation might still be classified as a love letter because the overarching sentiment and purpose of the text is to affirm love.

*Context and Narrative Focus:* The context in which emotions are expressed also plays a crucial role. If the narrative or the majority of the text revolves around themes of love, connections, and positive memories, it sets a more dominant tone of love, even if there are significant moments of sadness or other emotions.

*Resolution and Conclusion:* Often, the way emotions are resolved toward the end of a text can also dictate its overall theme. If a text concludes with a reaffirmation of love or a hopeful outlook toward a relationship, despite earlier sections expressing sadness or despair, the overall interpretation might lean toward love.

**Table 10.5** Original text of the poem “To My Sister”

It is the first mild day of March:	My sister! ('tis a wish of mine)
Each minute sweeter than before	Now that our morning meal is done,
The redbreast sings from the tall larch	Make haste, your morning task resign;
That stands beside our door.	Come forth and feel the sun.
There is a blessing in the air,	Edward will come with you;—and, pray,
Which seems a sense of joy to yield	Put on with speed your woodland dress;
To the bare trees, and mountains bare,	And bring no book: for this one day
And grass in the green field.	We'll give to idleness.
No joyless forms shall regulate	Love, now a universal birth,
Our living calendar:	From heart to heart is stealing,
We from to-day, my Friend, will date	From earth to man, from man to earth:
The opening of the year.	—It is the hour of feeling.
One moment now may give us more	Some silent laws our hearts will make,
Than years of toiling reason:	Which they shall long obey:
Our minds shall drink at every pore	We for the year to come may take
The spirit of the season.	Our temper from to-day.
And from the blessed power that rolls	Then come, my Sister! come, I pray,
About, below, above,	With speed put on your woodland dress;
We'll frame the measure of our souls:	And bring no book: for this one day
They shall be tuned to love.	We'll give to idleness.

*Purpose of the Expression:* The author's intent or purpose in expressing these emotions can also guide the classification. If sadness is expressed as a challenge within the context of a loving relationship, it may be seen as an element of the love story rather than the central theme.

### 10.G.1 Article 23: Soldier's Letter During War

*Joy (+1.0):* Joy is strongly felt in the memories of past moments together and the love that continues to give strength, as stated in “the memories of the blissful moments we have shared fill me with joy.” *Sadness (-0.6):* Sadness due to the current situation and potential farewell is expressed in “brings a poignant mixture of joy and sadness.” *Courage (+0.6):* The sense of duty and courage to face battle, “As I face the possibility of laying down my life for our country.” *Fear (-0.6):* Fear of what lies ahead in battle, indirectly mentioned through “the uncertainty of what lies ahead.” *Love (+1.0):* Deep love that sustains and uplifts, found in “My love for you is as fervent as ever.”

**Table 10.6** “To My Sister” poem rewritten to reflect the negative emotion “despair”

It is the first <b>dim</b> day of March: Each minute <b>colder</b> than before The redbreast <b>mourns</b> from the dying larch That <b>looms</b> beside our door.	My sister! (such is my <b>somber</b> plea) Now that our morning meal has ended, Make haste, abandon your morning chore; Come out and <b>endure the gloom</b> .
There is a <b>gloom</b> in the air, Which seems a sense of <b>sorrow</b> to bring To the bare trees, and <b>bleak</b> mountains, And grass in the <b>pale</b> field.	Edward will join you;—and, I ask, Dress quickly in your <b>grey attire</b> ; And bring no book: for this one day We'll <b>succumb to despair</b> .
No joyful sounds shall pierce Our daily <b>dread</b> : We from today, my Friend, will note The closing of the year.	<b>Sorrow</b> , now a common feeling, From heart to heart is spreading, From earth to man, from man to earth: —It is the time of <b>mourning</b> .
One moment now might deliver us More despair than years of logic: Our minds shall absorb at every breath The spirit of this <b>bleak</b> season.	Some <b>grim</b> laws our hearts will craft, Which they must eternally follow: We for the year to come may take Our <b>despair</b> from today.
And from the <b>cursed</b> force that <b>winds</b> About, beneath, above, We'll set the measure of our souls: They shall be tuned to <b>sadness</b> .	Then come, my Sister! come, I beg, With haste, wear your <b>grey garb</b> ; And bring no book: for just this day We'll <b>surrender to despair</b> .

### 10.G.2 Article 25: Letter to Sophie

Longing (+0.6): Longing for the presence and closeness, highlighted in “it seems to me that half of myself is missing.” Sadness (-0.6): Sadness over their separation and its effects, “my happiness has departed.” Love (+1.0): Constant reflections on love and its necessity, “we have enough in our hearts to love always.” Melancholy (-0.3): Melancholy over their current state, visible in the line, “we cannot become healed.” Contentment (+0.3): Found in the deep emotional satisfaction of their bond, despite physical absence, “how true that is! and it is also true that when one acquires such a habit, it becomes a necessary part of one’s existence.”

### 10.G.3 Article 53: Will of Laura Mary Octavia Lyttelton

Love (+1.0): The profound love expressed throughout, particularly in “all I am and ever shall be,” belongs to him more than anyone. Sadness (-0.6): Sadness at the thought of death and separation, but with a nuanced acceptance, “the sadness of death and parting is greatly lessened to me.” Contentment (+0.3): Contentment in

the deep connection with Alfred, reflecting a serene acceptance of their spiritual bond. Joy (+1.0): Joy in the enduring love they share, “so few women have been as happy as I have been.” Tranquility (+1.0): Tranquility in the face of life’s ultimate transition, feeling that their union will transcend even death.

## 10.H

### Appendix H: “To My Sister” of Different Linguistic Behaviors

#### To My Sister

by William Wordsworth (1770–1850)

The original text by William Wordsworth could be classified as “Hopeful” due to its optimistic outlook and the presence of renewal and joy throughout the poem. It embodies the spirit of embracing the new beginnings of March in a light, uplifting tone, focusing on the beauty of nature and the simple joy of being idle for a day.

**Table 10.7** “To My Sister” poem rewritten to reflect the strong positive emotion “joyful affection”

<p>It is the first <b>bright</b> day of March: Each moment more <b>delightful</b> than before The redbreast <b>joyfully</b> sings from the vibrant larch That stands so <b>grandly</b> by our door. There is a <b>warmth</b> in the air, Which seems a sense of <b>bliss</b> to bring To the <b>blooming</b> trees, and sunlit mountains, And grass in the <b>lush</b> field. No dreary thoughts shall darken Our lively <b>celebration</b>: We from today, my Friend, will celebrate The start of the year.</p>	<p>My sister! (such is my <b>joyful</b> plea) Now that our morning meal has ended, Make haste, abandon your morning chores; Come out and <b>embrace the sunshine</b>. Edward will join you;—and, I ask, Dress quickly in your <b>festive attire</b>; And leave behind all books: for this one day We'll <b>bask in pure joy</b>. <b>Love</b>, now in full bloom, From heart to heart is leaping, From earth to us, from us to earth: —It is the hour of <b>exuberance</b>.</p>
<p>One moment now may bring us more Joy than years of endless thought: Our spirits will soak up at every breath The essence of this <b>joyous</b> season. And from the <b>divine energy</b> that <b>radiates</b> Around, below, above, We'll adjust the harmony of our souls: They shall resonate with <b>happiness</b>.</p>	<p>Some <b>cheerful</b> laws our hearts will create, Which we'll joyfully follow: We for the year to come may take Our <b>joy</b> from today. Then come, my Sister! come, I exhort, With zest, wear your <b>vibrant dress</b>; And bring no book: for today alone We <b>celebrate pure happiness</b>.</p>

### **10.H.1 Rewrites Depicting Different Linguistic Behaviors**

We asked GPT-4 to perform rewrites based on two linguistic behaviors, “despair” and “joyful affection,” by providing each rewrite with a corresponding emotion vector. Table 10.6 presents the “despair” version of the poem. In this version, the major changes in emotion words highlight a shift from a positive to a negative sentiment. The specific changes are shown in Table 10.6, with emotion-laden words highlighted in red. These red-colored words, compared to the original words, clearly show an emotion shift from hopefulness to a sense of gloom, sadness, and pessimism—for example, from sweet to dim, from blessed to curse, and from woodland dress to gray garb. GPT-4 maintains the structure of the poem without making a major restructure, and this is appropriate in this context.

Table 10.7 presents the “joyful affection” version of the poem. The major changes in emotion words underscore a transformation from a generally positive to a distinctly joyful sentiment. Specific changes are indicated within Table 10.7, with words laden with emotion highlighted in blue. This allows for a direct comparison between the two versions at opposite ends of the linguistic behavior spectrum, illustrating the alterations in words related to brightness, attire, and emotions. The edits extend beyond simply replacing adjectives mechanically; they include modifying verbs and enhancing descriptive imagery to evoke a stronger emotional resonance and vividness in the text.

## **References**

- M. G. Azar, Z. D. Guo, B. Piot, et al. 2024. A general theoretical paradigm to understand learning from human preferences. In S. Dasgupta, S. Mandt, and Y. Li (Eds.), *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics* (Vol. 238), Valencia, Spain, May 2–4, 2024, Proceedings of Machine Learning Research (PMLR), 4447–4455. Retrieved from <https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html>.
- Y. Bai, S. Kadavath, S. Kundu, et al. 2022. Constitutional AI: Harmlessness from AI feedback. DOI: <https://doi.org/10.48550/arXiv.2212.08073>.
- L. F. Barrett. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, Boston.
- S. Carey and E. Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development* 15, 17–29.
- C. S. Carver, S. Sinclair, and S. L. Johnson. 2010. Authentic and hubristic pride: Differential relations to aspects of goal regulation, affect, and self-control. *J. Res. Pers.* 44, 6, 698–703. DOI: <https://doi.org/10.1016/j.jrp.2010.09.004>.
- E. Y. Chang. December. 2023. Examining GPT-4’s capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI ’23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.

- E. Y. Chang. August. 2024a. Behavioral emotion analysis model for large language models. In *Proceedings of the 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR '24)*, San Jose, CA, August 7–9, 2024. IEEE, 549–556. DOI: <https://doi.org/10.1109/MIPR62202.2024.00094>.
- E. Y. Chang. August. 2024b. EVINCE: Optimizing multi-LLM dialogues using conditional statistics and information theory. DOI: <https://doi.org/10.48550/arXiv.2408.14575>.
- E. Y. Chang. 2024c. *Multi-LLM Agent Collaborative Intelligence: The Path to Artificial General Intelligence*. Amazon, ISBN: 978-1-962463-07-2.
- E. Y. Chang and L. Geng. 2025. SagALLM: Context management, validation, and transaction guarantees for multi-agent LLM planning. *Proc. VLDB* 14, 1, 1–13. DOI: <https://doi.org/10.48550/arXiv.2503.11951>.
- P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. 2017. Deep reinforcement learning from human preferences. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY, 4302–4310. DOI: <https://doi.org/10.48550/arXiv.1706.03741>.
- J. Dai, T. Chen, Y. Yang, Q. Zheng, and G. Pan. 2025. Mitigating reward over-optimization in RLHF via behavior-supported regularization. In *Proceedings of the 13th International Conference on Learning Representations (ICLR '25)*. Singapore, April 24, 2025. Retrieved from <https://openreview.net/forum?id=PNMv4r7s1i>.
- A. R. Damasio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, New York, NY.
- R. J. Davidson. 2003. Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology* 40, 5, 655–665. DOI: <https://doi.org/10.1111/1469-8986.00067>.
- S. Dehaene, J.-P. Changeux, L. Naccache, J. Sackur, and C. Sergent. 2006. Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends Cogn. Sci.* 10, 5, 204–211. DOI: <https://doi.org/10.1016/j.tics.2006.03.007>.
- M. Eid and E. Diener. 2001. Norms for experiencing emotions in different cultures: Inter- and intranational differences. *J. Pers. Soc. Psychol.* 81, 5, 869–885. DOI: <https://doi.org/10.1037/0022-3514.81.5.869>.
- P. Ekman. 1992. An argument for basic emotions. *Cogn. Emot.* 6, 3–4, 169–200. DOI: <https://doi.org/10.1080/02699939208411068>.
- P. Ekman. 1999. Basic emotions. In T. Dalgleish and M. J. Power (Eds.), *Handbook of Cognition and Emotion*. John Wiley and Sons, Chap. 3, pp. 45–60. DOI: <https://doi.org/10.1002/0470013494.ch3>.
- K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. 2024. Model alignment as prospect theoretic optimization. In R. Salakhutdinov, Z. Kolter, K. Heller, et al. (Eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024. JMLR.org, 12634–12651.
- G. Fauconnier and M. Turner. 2002. *The Way We Think: Conceptual Blending and The Mind's Hidden Complexities*. Basic Books, New York.

- D. J. Felleman and D. C. Van Essen. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1, 1–47. DOI: <https://doi.org/10.1093/cercor/1.1.1-a>.
- A. P. Fiske, S. Kitayama, H. R. Markus, and R. E. Nisbett. 1998. The cultural matrix of social psychology. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), *The Handbook of Social Psychology*, Vol. 2. McGraw-Hill, Boston, MA, 915–981.
- F. S. Fitzgerald and Z. Sayre. 2003. *Dear Scott, Dearest Zelda: The Love Letters of F. Scott and Zelda Sayre*. J. R. Bryer and C. W. Barks, (Eds.), Bloomsbury, London.
- B. L. Fredrickson. 1998. What good are positive emotions? *Rev. Gen. Psychol.* 2, 3, 300–319. DOI: <https://doi.org/10.1037/1089-2680.2.3.300>.
- D. Ganguli, A. Askell, N. Schiefer, et al. 2023. The capacity for moral self-correction in large language models. DOI: <https://doi.org/10.48550/arXiv.2302.07459>.
- K. Grill-Spector, K. S. Weiner, K. Kay, and J. Gomez. 2017. The functional neuroanatomy of human face perception. *Ann. Rev. Vis. Sci.* 3, 167–196. DOI: <https://doi.org/10.1146/annurev-vision-102016-061214>.
- J. J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Rev. Gen. Psychol.* 2, 3, 271–299. DOI: <https://doi.org/10.1037/1089-2680.2.3.271>.
- M. Heikkilä and W. D. Heaven. June. 2022. Yann LeCun has a bold new vision for the future of AI. MIT Technology. Review from <https://www.technologyreview.com/2022/06/24/1054817/yann-lecun-bold-new-vision-future-ai-deep-learning-meta/>.
- G. Hofstede. 1980. *Culture's Consequences: International Differences in Work Related Values*. Sage Publications, Beverly Hills, CA.
- R. Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford. DOI: <https://doi.org/10.1093/acprof:oso/9780198270126.001.0001>.
- W. James. 1884. II. What is an emotion? *Mind* 9, 34, 188–205. DOI: <https://doi.org/10.1093/mind/os-IX.34.188>.
- W. James. 1890. *The Principles of Psychology*, Vol. 2. Henry Holt and Company, New York. DOI: <https://doi.org/10.1037/11059-000>.
- Kaggle. 2021. Love letter analysis, Version 2. Retrieved from <https://www.kaggle.com/code/metformin/love-letter-analysis/notebook>.
- E. R. Kandel, J. H. Schwartz, and T. M. Jessell. 2013. *Principles of Neural Science*. McGraw-Hill, New York, NY.
- B. Kennedy, M. Atari, A. M. Davani, et al. 2022. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resour. Eval.* 56, 79–108. DOI: <https://doi.org/10.1007/s10579-021-09569-x>.
- J. Kirkpatrick, R. Pascanu, N. Rabinowitz, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci.* 114, 13, 3521–3526. DOI: <https://doi.org/10.1073/pnas.1611835114>.
- B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350, 6266, 1332–1338. DOI: <https://doi.org/10.1126/science.aab3050>.

- G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.
- C. G. Lange. 1885. *The Emotions: A Psychophysiological Study*. William & Wilkins, Baltimore, MD.
- H. Lee, S. Phatale, H. Mansoor, et al. 2023. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024. JMLR.org, 26874–26901. DOI: <https://doi.org/10.48550/arXiv.2309.00267>.
- Y. Lin, H. Lin, W. Xiong, et al. November. 2024. Mitigating the alignment tax of RLHF. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, November 12–16, 2024. ACL, 580–606. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.35>.
- G. Marcus. 2020. The next decade in AI: Four steps towards robust artificial intelligence. DOI: <https://doi.org/10.48550/arXiv.2002.06177>.
- H. R. Markus and S. Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychol. Rev.* 98, 2, 224–253. DOI: <https://doi.org/10.1037/0033-295X.98.2.224>.
- C. McGinn and K. Kelly. 2018. Using the Geneva Emotion Wheel to classify the expression of emotion on robots. In T. Kanda, S. Sabanoviæ, G. Hoffman, and A. Tapus (Eds.), *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*, Chicago, IL, March 5–8, 2018. ACM, New York, NY, 191–192. DOI: <https://doi.org/10.1145/3173386.3177058>.
- B. Mesquita and N. H. Frijda. 1992. Cultural variations in emotions: A review. *Psychol. Bull.* 112, 2, 179–204. DOI: <https://doi.org/10.1037/0033-2909.112.2.179>.
- G. Mialon, R. Dessi, M. Lomeli, et al. 2023. Augmented language models: A survey. *Trans. Mach. Learn. Res.* 1–35. DOI: <https://doi.org/10.48550/arXiv.2302.07842>.
- M. Minsky. 1988. *The Society of Mind*. Simon and Schuster, New York, NY.
- I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumacas. 2022. ETHOS: A multi-label hate speech detection dataset. *Complex Intell. Syst.* 8, 4663–4678. Retrieved from <https://link.springer.com/article/10.1007/s40747-021-00608-2#citeas>.
- OpenAI. 2023. GPT-4 technical report. DOI: <https://doi.org/10.48550/arXiv.2303.08774>.
- C. Oveis, E. J. Horberg, and D. Keltner. 2010. Compassion, pride, and social intuitions of self-other similarity. *J. Pers. Soc. Psychol.* 98, 4, 618–630. DOI: <https://doi.org/10.1037/a0017628>.
- R. Plutchik. 1982. A psychoevolutionary theory of emotions. *Soc. Sci. Inf.* 21, 4–5, 529–553. DOI: <https://doi.org/10.1177/053901882021004003>.
- R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New

- Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 53728–53741. DOI: <https://doi.org/10.48550/arXiv.2305.18290>.
- S. Schachter and J. E. Singer. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* 69, 5, 379–399. DOI: <https://doi.org/10.1037/h0046234>.
- K. R. Scherer. 2005. What are emotions? And how can they be measured? *Soc. Sci. Inf.* 44, 4, 695–729. DOI: <https://doi.org/10.1177/0539018405058216>.
- K. R. Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cogn. Emot.* 23, 7, 1307–1351. DOI: <https://doi.org/10.1080/02699930902928969>.
- N. Schwarz and G. L. Clore. 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J. Pers. Soc. Psychol.* 45, 3, 513–523. DOI: <https://doi.org/10.1037/0022-3514.45.3.513>.
- M. Shanahan, K. McDonell, and L. Reynolds. 2023. Role play with large language models. *Nature* 623, 7987, 493–498. DOI: <https://doi.org/10.1038/s41586-023-06647-8>.
- R. Sinha. 2008. Chronic stress, drug use, and vulnerability to addiction. *Ann. N. Y. Acad. Sci.* 1141, 105–130. DOI: <https://doi.org/10.1196/annals.1441.030>.
- J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. 2022. Defining and characterizing reward hacking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, November 28, 2022–December 9, 2022. Curran Associates Inc., Red Hook, NY, 9460–9471.
- C. A. Smith and P. C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* 48, 4, 813–838. DOI: <https://doi.org/10.1037/0022-3514.48.4.813>.
- N. Stiennon, L. Ouyang, J. Wu, et al. 2022. Learning to summarize from human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada, December 6–12, 2020. Curran Associates Inc., Red Hook, NY, 3008–3021. DOI: <https://doi.org/10.48550/arXiv.2009.01325>.
- A. N. Tak and J. Gratch. 2024. GPT-4 emulates average-human emotional cognition from a third-person perspective. In *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Glasgow, United Kingdom, September 15–18, 2024. IEEE, 337–345. DOI: <https://doi.org/10.1109/ACII63134.2024.00043>.
- L. Talmy. 2000. *Toward a Cognitive Semantics*, Vol. 1 & 2. MIT Press, Cambridge, MA.
- Y. Tang, D. Z. Guo, Z. Zheng, et al. 2024. Generalized preference optimization: A unified approach to offline alignment. In R. Salakhutdinov, Z. Kolter, K. Heller, et al. (Eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024. JMLR.org, 47725–47742. DOI: <https://doi.org/10.48550/arXiv.2402.05749>.
- M. Torrens, F. Fonseca, G. Mateu, and M. Farré. Efficacy of antidepressants in substance use disorders with and without comorbid depression: A systematic review and meta-analysis. *Drug Alcohol Depend.* 2005, 78, 1, 1–22. DOI: <https://doi.org/10.1016/j.drugalcdep.2004.09.004>.

- J. L. Tracy and R. W. Robins. 2007. The psychological structure of pride: A tale of two facets. *J. Pers. Soc. Psychol.* 92, 3, 506–525. DOI: <https://doi.org/10.1037/0022-3514.92.3.506>.
- S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *Proceedings of the 10th International Conference on Learning Representations (ICLR '22)*. Retrieved from <https://openreview.net/forum?id=RdJVFCCHjUMI>.
- S. Yao, J. Zhao, D. Yu, et al. 2023. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR '23)*, Kigali, Rwanda, May 1–5, 2023. DOI: <https://doi.org/10.48550/arXiv.2210.03629>.
- Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. 2023. SLiC-HF: Sequence likelihood calibration with human feedback. DOI: <https://doi.org/10.48550/arXiv.2305.10425>.

## CHAPTER 11

---

# SagaLLM: Persistent Context Management, Constraint Validation, and Transaction Guarantees

### Abstract

This chapter presents SagaLLM, a structured multi-agent architecture designed to address four foundational limitations of current Large Language Model-based planning systems (LLM-based systems): unreliable self-validation, context loss, lack of transactional safeguards, and insufficient inter-agent coordination. While recent frameworks leverage LLMs for task decomposition and multi-agent communication, they often fail to ensure consistency, rollback, or constraint satisfaction across distributed workflows. SagaLLM bridges this gap by integrating the saga-based transaction management with persistent memory, automated compensation, and independent validation agents. It leverages LLMs' generative reasoning to automate key tasks traditionally requiring hand-coded coordination logic, including state tracking, dependency analysis, log schema generation, and recovery orchestration. Although SagaLLM relaxes strict ACID (Atomicity, Consistency, Isolation, and Durability) guarantees, it ensures workflow-wide consistency and recovery through modular checkpointing and compensable execution. Empirical evaluations across planning domains demonstrate that standalone LLMs frequently violate interdependent constraints or fail to recover from disruptions. In contrast, SagaLLM achieves significant improvements in consistency, validation accuracy, and adaptive coordination under uncertainty—establishing a robust foundation for real-world, scalable LLM-based Multi-Agent Systems (MAS).

## 11.1

### Introduction

Multi-Agent Systems (MAS) have long been a cornerstone of distributed computing and database systems [Jennings 1993, Lesser et al. 2003, Dunin-Kęplicz and Verbrugge 2010]. Over the past few decades, their development has followed two primary trajectories. In the database community, MAS traditionally integrated the foundational transaction processing principles—particularly ACID (Atomicity, Consistency, Isolation, and Durability) properties [Gray 1981, Weikum and Vossen 2001]—to ensure consistency and reliability in complex, multi-step operations. For long-lived, distributed, or loosely coupled tasks, MAS also adopted more flexible transactional models (e.g., Sagas [Garcia-Molina and Salem 1987]) to maintain robustness while relaxing strict atomicity or isolation constraints.

Parallel to these database-oriented approaches, distributed systems research emphasized coordination protocols and flexible collaboration mechanisms [Duffee 1999, Wooldridge 2009], enabling scalable multi-agent interactions without the overhead of strict locking or heavyweight transactional guarantees. These complementary development paths have resulted in frameworks optimized for different priorities: *transactional integrity* versus *adaptive coordination*, highlighting the fundamental trade-off between strong consistency and flexible execution in practical systems.

Recent advances in Large Language Models (LLMs) [Vaswani et al. 2017, Chang et al. 2024, Zhao et al. 2025] have revitalized MAS as a paradigm for sophisticated reasoning and multi-agent collaboration [Chang 2023, 2024, Du et al. 2024]. Frameworks such as AutoGen (an opensource framework developed by Microsoft for building and orchestrating MAS) [Wu et al. 2024], LangGraph (an opensource AI framework that uses graph-based architectures to model and manage the complex AI agent workflows) [LangChain AI 2024], and CAMEL (Communicative Agents for Mind Exploration of Large Scale Language Model Society) (an opensource framework designed for building and managing MAS where AI agents autonomously collaborate and communicate to solve complex tasks) [Li et al. 2023a] demonstrate how LLM-based agents can decompose tasks, interact between modalities, and coordinate to solve complex problems. However, this resurgence often neglects the foundational transaction guarantees that are essential to reliable multi-agent workflows, particularly in domains requiring robust state management.

Unlike traditional MAS, LLM-based systems often lack mechanisms for maintaining strong consistency, failure recovery, and rollback handling, leading to inconsistent states, partial failures, and unreliable execution in real-world applications. These limitations stem from several fundamental challenges: LLMs struggle with *internal validation* due to inherent limitations highlighted by Gödel's incompleteness theorems [Gödel 1967], making them unreliable for detecting and

correcting their own errors. Furthermore, *context loss* in long conversations [Hsieh et al. 2024, Liu et al. 2024a, Xiao et al. 2024, Modarressi et al. 2025] can cause LLMs to forget earlier steps, leading to contradictory decisions. When tasks are distributed across multiple agents, these problems are compounded, as no built-in supervisory mechanism exists to reconcile state changes or to validate constraint satisfaction across agents.

For example, in a travel booking scenario, an LLM-based MAS might independently issue flight and hotel reservations without ensuring their coordinated success. If the flight is later canceled, the system may fail to recognize the inconsistency, leaving the hotel reservation active. Such scenarios illustrate the critical need for transactional frameworks that preserve the intelligence and adaptability of LLM-based MAS while ensuring consistency and reliability in long-running, interdependent workflows.

To address these limitations, we propose SagaLLM, a multi-agent transactional system that extends the saga pattern, a transactional model originally developed to manage complex, long-lived transactions by decomposing them into smaller, independently validated, committed, and compensable units. By integrating transactional logic, compensatory rollback mechanisms, and *persistent memory* into LLM-based MAS, SagaLLM ensures that each individual operation within a workflow is reliably validated and committed, with clearly defined compensating transactions that restore system-wide consistency in case of failure.

Crucially, SagaLLM leverages the reasoning and coding capabilities of LLMs to automate core aspects of transaction orchestration that previously required scenario-specific manual programming. SagaLLM enables LLMs to:

- (1) identify the persistent states to track.
- (2) validate constraints and inter-agent dependencies.
- (3) design logging schemas to capture workflow transitions.
- (4) develop compensatory logic for failure recovery.
- (5) implement communication protocols among agents to coordinate these behaviors.

Traditionally, each of these components required custom implementation by system developers, tailored to individual applications. In contrast, SagaLLM employs LLMs as intelligent agents that automatically infer, generate, and coordinate these mechanisms, greatly improving scalability and reducing development overhead.

This hybrid approach, integrating transactional processing with adaptive multi-agent intelligence, makes SagaLLM particularly effective for real-world applications that are complex and demand reliability and safety, for example, healthcare

management, supply chain management, and emergency response. These capabilities collectively enable SagaLLM to overcome the foundational limitations in LLM-based MAS, such as unreliable state coordination, lack of rollback support, and limited context retention, thereby supporting robust, scalable reasoning across distributed agents. We summarize our key contributions in the following list:

- (1) **Transactional Consistency via Persistent Memory and Compensation:** SagaLLM introduces transactional safeguards and persistent memory-based compensatory mechanisms to LLM-based MAS, ensuring reliable consistency and coherent state recovery across multi-agent workflows.
- (2) **Robust Constraint and Dependency Validation:** SagaLLM incorporates temporal-spatial context tracking and external verification mechanisms to validate inter-agent dependencies and prevent inconsistencies while addressing the fundamental limitations of LLM self-verification.
- (3) **LLM-Orchestrated Intelligence:** SagaLLM automates key components of multi-agent planning, state tracking, constraint checking, log schema design, compensation logic, and coordination protocols, through the generative reasoning and coding capabilities of LLMs.

The remainder of this chapter covers: related work (Section 11.2), problem definition (Section 11.3), architecture (Section 11.4), evaluation (Section 11.5), and conclusions (Section 11.6).

## 11.2 Related Work

We review three strands of related work: (1) the evolution of transactional management, particularly the saga pattern; (2) cognitive limitations of LLMs motivating the need for transactional integrity, independent validation, and context preservation; and (3) multi-agent LLM frameworks and recent attempts to integrate transactional safeguards.

### 11.2.1 Transaction Management Systems

Transactional models have evolved significantly since Gray introduced the ACID properties [Gray 1981]. In distributed settings, strict ACID guarantees became impractical, prompting models such as BASE (Basically Available, Soft State, Eventually Consistent) [Pritchett 2008] and long-lived transaction patterns.

The Saga pattern by Garcia-Molina and Salem [1987] decomposes long-lived transactions into smaller, locally atomic sub-transactions with compensating steps for failure recovery. This has influenced modern microservice architectures and workflow engines [Richardson 2018].

Systems such as YAWL (Yet Another Workflow Language) [Aalst and Hofstede 2005], AWS (Amazon Web Services) Step Functions [AWS Step Functions 2023], and Azure Logic Apps [2023] embed Saga-style workflows but remain rigid and manually defined, lacking dynamic adaptability. These foundational principles inform SagaLLM's adaptive, LLM-driven extensions.

### **11.2.2 LLM Limitations Necessitating SagaLLM's Key Requirements**

To realize this complementary relationship between AI and workflow systems, we must first address the specific limitations that prevent current LLMs from functioning effectively in transactional workflows. Our analysis reveals three fundamental limitations that directly motivate SagaLLM's core requirements: transactional integrity, independent validation, and strategic context preservation.

#### **11.2.2.1 Self-Validation Gap Necessitating Independent Validation**

LLMs inherently lack robust self-validation mechanisms, a limitation originating from intrinsic boundaries identified by Gödel's incompleteness theorems, demonstrating fundamental constraints on a system's ability to verify its own reasoning [Gödel 1967, Chang and Geng 2025]. Recent research confirms that self-refinement techniques [Madaan and Yazdanbakhsh 2022, Li et al. 2023b, Jiang et al. 2024], while iterative and beneficial, are unable to surpass inherent capability ceilings to reliably correct deeper logical errors [Huang et al. 2024]. In transactional scenarios, these validation gaps manifest as factual inconsistencies, invalid operations, and unreliable plan feasibility assessments [Yamin et al. 2024]. Thus, SagaLLM incorporates an independent validation framework to mitigate these inherent limitations.

#### **11.2.2.2 Statelessness Necessitating Transactional Integrity**

LLMs process each interaction independently, lacking native mechanisms to maintain the state across sequential interactions. This fundamental statelessness necessitates explicit transactional integrity management to maintain coherent operation sequences and ensure robust failure recovery. Without systematic transaction management, LLM-based systems risk state inconsistency, operation losses, and incoherent recovery procedures.

#### **11.2.2.3 Context Limitations and Strategic Preservation**

LLMs rely on self-attention mechanisms that prioritize recent tokens, leading to significant degradation in context retention over long sequences. Empirical

studies reveal sharp drops in recall beyond token limits [Xiao et al. 2024, Modarressi et al. 2025], especially for mid-context information [Hsieh et al. 2024, Liu et al. 2024a]. Chain-of-thought heuristics [Cemri et al. 2025] further exacerbate this by lacking mechanisms to manage or pass context reliably across steps. These limitations hinder multi-step reasoning [Wei et al. 2022, Stechly et al. 2024], as earlier outputs are frequently lost. SagaLLM addresses this by explicitly preserving vulnerable context elements—goals, justifications, and dependencies—through structured memory and persistent tracking.

Collectively, these limitations provide strong motivation to address all three key requirements within SagaLLM: comprehensive transaction management, independent validation, and strategic context preservation are essential for reliably deploying LLM-based MAS in critical real-world applications.

### **11.2.3 Multi-Agent LLM Frameworks and Transaction Limitations**

Frameworks such as AutoGen [Wu et al. 2024], LangGraph [LangChain AI 2024], and CAMEL [Li et al. 2023a] advance multi-agent LLM coordination but fall short in addressing SagaLLM’s three core requirements: transactional integrity, independent validation, and context preservation.

#### **11.2.3.1 Missing Transaction Semantics**

LangGraph and AutoGen enable structured workflows and agent interactions, but lack built-in atomicity guarantees, compensation logic, or robust failure recovery. AgentScope (framework designed to define and manage the “scope” or operational boundaries of autonomous agents in an MAS) [Gao et al. 2024] and AFlow (framework or architecture that models the flow of information, tasks, or control between multiple agents) [Zhang et al. 2024] introduce limited rollback mechanisms, but do not generalize across workflows.

#### **11.2.3.2 Validation Gaps**

Most frameworks are based on LLM self-validation, exposing them to reasoning errors and hallucinations. Systems such as PLASMA (a framework for multi-agent planning that leverages LLMs to coordinate actions across agents in complex environments) [Brahman et al. 2024] improve reliability but omit transaction-level rollback. LLM-MCTS (Large Language Model–Monte Carlo Tree Search) [Zhao et al. 2023] and Tree-of-Thought (ToT—a reasoning framework for LLMs that improves their ability to solve complex problems by structuring the reasoning process as a tree, rather than a single chain of thoughts) [Yao et al. 2023] emphasize pre-execution reasoning without runtime consistency checks.

### 11.2.3.3 Limited Context Preservation

CAMEL preserves dialogue history but lacks mechanisms for tracking state transitions, inter-agent dependencies, or compensatory paths. Broader planning systems [Wei et al. 2025] do not offer persistent context retention strategies.

Unlike these systems, SagaLLM treats compensation, validation, and context tracking as first-class design goals—ensuring reliable, recoverable, and intelligent coordination across complex multi-agent workflows.

## 11.3 System Requirements for SagaLLM

Building on the limitations identified in Section 11.2, this section formally defines SagaLLM’s requirements. While SagaLLM inherits the core transactional semantics of the classical saga pattern [Garcia-Molina and Salem 1987], it requires significant adaptations to suit LLM-based MAS, as summarized in Table 11.1.

To address the limitations in LLM-based execution, SagaLLM is organized around three tightly interwoven yet conceptually distinct requirements that extend classical transaction processing into the realm of adaptive multi-agent intelligence:

- (1) **Transactional Integrity:** Ensures that agent operations transition the system through coherent, globally consistent states. This is achieved through structured rollback mechanisms, compensating actions, and invariant preservation across interdependent agents. It also requires reliable tracking of system state to detect and repair inconsistencies triggered by partial execution or disruptions.
- (2) **Independent Validation:** Addresses the known limitations of LLM self-validation by introducing cross-agent and external validation layers. These mechanisms evaluate agent outputs and inter-agent inputs against

**Table 11.1** Classical Sagas vs. SagaLLM: Framework comparison

Aspect	Classical Sagas	SagaLLM
Domain	Database transactions	Multi-agent LLM workflows
Compensation	Pre-defined rollback procedures	LLM-generated + validated compensation
Validation	Schema/constraint validation	Independent LLM output validation
Context	Stateless transactions	Strategic context preservation
Coordination	Simple sequential execution	Complex multi-agent dependency management
Intelligence	Rule-based workflows	Adaptive LLM reasoning with transaction guarantees

constraints, schemas, and dependency graphs. Persistent validation histories are maintained to support rollback triggers and guard against hallucinations or invalid commitments.

- (3) **Context Management:** Maintains essential state and dependency information across long-horizon workflows. Rather than relying on ephemeral context windows, SagaLLM persistently stores goals, justifications, and compensation plans in structured memory, enabling agents to reason over consistent histories and perform accurate recovery after failures.

This integrated design contrasts with prior systems that handle these aspects in isolation or without formal guarantees. The mutual reinforcement among these three pillars reflects the central challenge of LLM-based workflows: effective orchestration under uncertainty, without sacrificing consistency and correctness.

### 11.3.1 Transactional Integrity Requirements

SagaLLM provides transactional guarantees tailored for multi-agent workflows, extending the classical transaction semantics across autonomous agent boundaries. Sequences  $O = \{o_1, o_2, \dots, o_n\}$  are operations treated as a unit of logical cohesiveness, where each  $o_i$  is locally atomic. If any operation fails, SagaLLM initiates compensatory actions to restore global consistency.

Applying  $O$  to a system state  $S$  must yield either a fully committed state  $S'$ , or trigger a coherent rollback that returns the system to  $S$ , thereby avoiding partial or inconsistent outcomes. To ensure this, SagaLLM enforces the following properties:

- **Consistency Preservation:** SagaLLM ensures that all state transitions respect global invariants  $I$ . If  $S \models I$ , then any resulting  $S' \models I$ , even when execution spans multiple agents.
- **Isolation Guarantees:** SagaLLM guarantees that concurrently executing agents produce final states equivalent to some serial order, regardless of autonomy or internal decision processes.
- **Durability Assurance:** SagaLLM guarantees persistence of committed states by durably recording execution outcomes and metadata necessary for fault recovery and compensatory execution.

To enforce these guarantees, SagaLLM maps each  $o_i$  to a local transaction  $T_i$ , paired with a compensating transaction  $C_i$ . In case of failure at step  $T_j$ , compensating actions are invoked in reverse:

$$\text{Saga } S = \{T_1, T_2, \dots, T_n, C_n, \dots, C_2, C_1\}. \quad (11.1)$$

### 11.3.1.1 Transaction State Management

Table 11.2 lists the transaction states to be tracked and their associated information. SagaLLM maintains a structured state representation across three orthogonal dimensions to support validation, compensation, and recovery:

- *Application State ( $S_A$ )*: Domain-specific entities and system checkpoints.
- *Operation State ( $S_O$ )*: Execution logs, inputs, outputs, and LLM reasoning traces.
- *Dependency State ( $S_D$ )*: Graph-structured constraints and satisfaction criteria.

### 11.3.1.2 Dependency Tracking and Compensation Planning

SagaLLM models operation dependencies as a directed graph:

$$D = \{(o_i, o_j, c_{ij}) | o_j \text{ depends on } o_i \text{ under condition } c_{ij}\}. \quad (11.2)$$

To express more complex conditions:

$$c_{\{i_1, \dots, i_n\}, j} = \mathcal{B}(c_{i_1, j}, \dots, c_{i_n, j}), \quad (11.3)$$

where  $\mathcal{B}$  is a Boolean function defined under the prerequisite conditions.

**Table 11.2** Transaction state management in SagaLLM

Mechanism	Information Recorded
<i>Application State (<math>S_A</math>)</i>	
Domain Entities	<ul style="list-style-type: none"> <li>• Application-domain objects</li> <li>• Entity states and status</li> <li>• Checkpoints and snapshots</li> </ul>
<i>Operation State (<math>S_O</math>)</i>	
Execution Logs	<ul style="list-style-type: none"> <li>• Operation inputs and outputs</li> <li>• Timestamps and execution status</li> <li>• Completion indicators</li> </ul>
Decision Reasoning	<ul style="list-style-type: none"> <li>• LLM-generated reasoning chains</li> <li>• Justifications and alternatives</li> </ul>
Compensation Metadata	<ul style="list-style-type: none"> <li>• Inverse operations</li> <li>• Preconditions and recovery state</li> </ul>
<i>Dependency State (<math>S_D</math>)</i>	
Causal Dependencies	<ul style="list-style-type: none"> <li>• Inter-operation constraints</li> <li>• Data and resource flow mappings</li> <li>• Boolean condition checks</li> </ul>
Constraint Satisfaction	<ul style="list-style-type: none"> <li>• Satisfaction evidence and timestamps</li> </ul>

Upon failure, SagaLLM traverses this graph to determine the minimal set of affected operations and executes compensatory actions that restore global consistency without violating the preserved invariants.

### **11.3.2 Independent Validation Requirements**

To address the inherent limitations of LLM self-verification, SagaLLM introduces a two-tier validation architecture governed by a global validation agent. This agent operates independently of task agents and has visibility into the full transaction history, agent communications, and global state.

#### **11.3.2.1 Intra-Agent Output Validation**

The global validation agent inspects the outputs of the individual task agents before those outputs are committed or transmitted. Outputs are checked for:

- Syntactic correctness (format, schema).
- Semantic coherence and reasoning soundness.
- Factual accuracy against context.
- Constraint adherence and invariants.
- Context preservation and dependency awareness.

Failures trigger compensations or corrective augmentation.

#### **11.3.2.2 Inter-Agent Input and Dependency Validation**

Inputs and messages between agents are validated before delivery. Checks include:

- Contract conformance.
- Dependency satisfaction.
- Cross-agent consistency.
- Temporal ordering.
- Mutual agreement on shared state.
- Transaction coherence.

Failed validations block delivery and invoke recovery.

#### **11.3.2.3 Validation Response Protocols**

SagaLLM defines structured validation outcomes:

- *Rejection*: Discard and compensate.
- *Augmentation*: Enhance with clarifications.
- *Feedback*: Record for future adaptation.

### 11.3.3 Context Management Requirements

SagaLLM identifies and retains essential context for recovery, validation, and inter-agent dependencies:

- *Selective Retention*: Filters critical info.
- *Structured Storage*: Organizes specs, justifications, and reasoning.
- *Dependency Tracking*: Maintains prerequisites for rollback.
- *Communication Protocol*: Ensures necessary context exchange.

#### 11.3.3.1 Failure Handling and Recovery

Effective recovery depends on preserved context and dependency tracking. SagaLLM supports multi-level failure response:

- (1) *Operation-Level*: Upon failure, the system invokes compensatory actions using logs and rollback specifications stored in  $S_O$ .
- (2) *Workflow-Level*: SagaLLM traverses the dependency graph to orchestrate reverse execution paths across agents, restoring global consistency based on  $S_D$  and recorded constraints.

This layered integration of recovery within strategic context management enables SagaLLM to meet the transactional demands of complex real-world multi-agent LLM workflows.

## 11.4 Design with Travel Planning

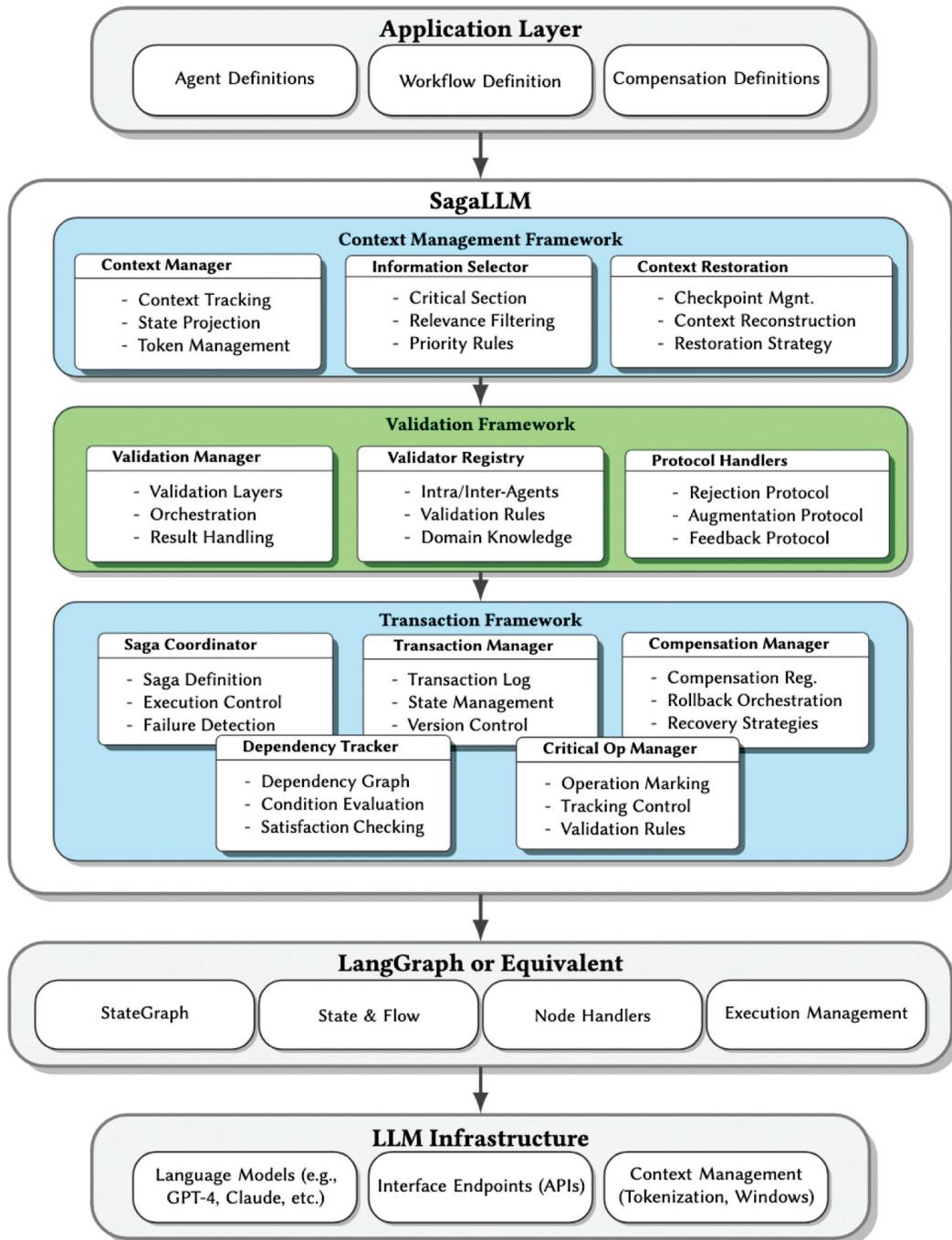
Figure 11.1 depicts the SagaLLM architecture, which sits between the application layer and LLM MAS (e.g., LangGraph). SagaLLM comprises three frameworks: *context management*, *validation*, and *transaction*. To illustrate the design, we use a travel planning example.

### 11.4.1 Travel Planning Problem

This example demonstrates how SagaLLM automatically manages complex multi-agent LLM workflows for international trip planning with multiple destinations, budget constraints, and transactional booking requirements. The application illustrates the transition from manual planning to automated SagaLLM-managed execution.

#### 11.4.1.1 Specifications

- Plan a trip from San Francisco to Berlin and Cologne and then back to San Francisco.



**Figure 11.1** Architecture of SagaLLM. It sits between the application layer and LLMs, consisting of three frameworks: context management, validation, and transaction.

- Travel period: June 2025 (flexible within the month).
- Budget constraint: \$5000 total.
- Required bookings: flights, hotels, and trains between cities.
- Preferences:
  - > Moderately priced accommodations (three- to four-star hotels).
  - > Direct flights when possible.
  - > Train pass to save money.
  - > Flexible scheduling with four days in Berlin and two days in Cologne.

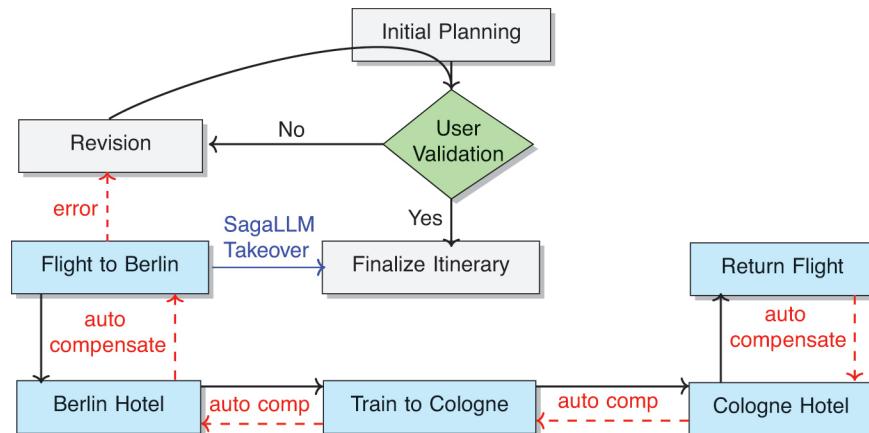
#### 11.4.2 Two-Phase Workflow Architecture

The application workflow consists of two distinct phases with different automation levels:

**Phase 1 (Manual Planning):** Human-driven itinerary planning and user validation.

**Phase 2 (Automated SagaLLM Execution):** Fully automated multi-agent transaction management.

Figure 11.2 illustrates this workflow transition, where gray boxes represent manual planning activities, and cyan boxes represent automated transactions managed by SagaLLM.



**Figure 11.2** Travel-planning workflow showing transition from manual planning to automated SagaLLM execution. Gray boxes represent manual human-driven activities. Cyan boxes represent fully automated SagaLLM-managed transactions with automatic compensation, validation, and recovery. The blue arrow indicates the handoff point where SagaLLM takes complete control.

### 11.4.2.1 Phase 1: Manual Itinerary Planning

Phase 1 involves traditional human-driven planning activities that establish the requirements and constraints for automated execution:

- (1) **Initial Plan Generation:** Human planners or basic LLMs generate multiple feasible itineraries based on user requirements, specifying flight options, hotel reservations, and train transportation with cost estimates validated against budget constraints.
- (2) **Iterative Refinement:** Users review itineraries and provide feedback, which leads to plan adjustments without automated transaction management. Essential context (dates, preferences, constraints) is tracked for handoff to SagaLLM, with iterations continuing until user satisfaction is reached.
- (3) **Plan Finalization and SagaLLM Handoff:** Users select and approve the final itinerary with all requirements and constraints. The system compiles comprehensive specifications including booking dependencies, budget limits, and user preferences, and then initiates **automated handoff to SagaLLM** where all subsequent design, coding, agent coordination, and execution becomes fully automated.

### 11.4.2.2 Phase 2: Automated SagaLLM Execution Overview

Once Phase 1 is completed, SagaLLM automatically takes full control of workflow development and management. Given a planning problem  $\mathcal{O}$ , constraint set  $D$ , and performance metrics  $\mathcal{M}$ , SagaLLM generates a complete workflow consisting of nodes and edges, provides specifications for both transaction agents and compensation agents, and conducts validation and refinement. Each workflow node and edge is assigned both a regular agent that handles transactions and a compensation agent that handles rollback execution.

For an in-depth discussion on Phase 2 code development and run-time monitoring, please refer to Chapter 12, which covers the ALAS framework.

### 11.4.2.3 Detailed Phase 2 Implementation

The automated execution phase consists of four integrated components:

- (1) **Automatic System Architecture Generation:** SagaLLM analyzes the finalized itinerary and automatically generates the appropriate agent architecture, defines the transaction sequences ( $T_1, T_2, \dots, T_n$ ) and the corresponding compensations ( $C_1, C_2, \dots, C_n$ ), and establishes validation rules and dependency graphs based on booking requirements.
- (2) **Automatic Agent Deployment and Coordination:** The system instantiates required domain agents (FlightBookingAgent, HotelBookingAgent, etc.)

with appropriate configurations, deploys GlobalValidationAgent and SagaCoordinatorAgent with full system access, and configures agent communication protocols and data schemas.

- (3) **Automatic Transaction Execution:** The system executes the complete booking sequence:  $T_1$  (International Flight Booking SFO → Berlin),  $T_2$  (Berlin Hotel Booking coordinated with flight confirmation),  $T_3$  (Train Booking Berlin → Cologne scheduled with hotel checkout),  $T_4$  (Cologne Hotel Booking aligned with train arrival), and  $T_5$  (International Return Flight Cologne → SFO coordinated with hotel checkout).
- (4) **Automatic Exception Handling and Recovery:** SagaLLM automatically detects validation failures and executes appropriate compensations, maintains system consistency without human intervention, automatically replans affected portions using preserved context and constraints, and falls back to Phase 1 for human re-evaluation only when automatic replanning cannot satisfy constraints.

### **11.4.3 Agent Architecture and Code Structures**

The SagaLLM system implements a two-tier agent architecture consisting of specialized task execution agents and global coordination agents that ensure system-wide consistency and validation.

#### **11.4.3.1 Task Execution Agents**

Task execution agents focus solely on their domain-specific operations, while all validation is handled externally by a global validation agent. Each agent maintains structured input/output interfaces and preserves internal state for compensation purposes.

##### ***FlightBookingAgent***

- *Input Schema:* travel\_dates, budget\_limit, airline\_preferences, passenger\_details.
- *Output Schema:* flight\_details, confirmation\_number, total\_cost, cancellation\_policy.
- *Internal State:* reservation\_status, booking\_reference, payment.

##### ***HotelBookingAgent***

- *Input Schema:* checkin\_date, checkout\_date, location\_constraints, amenity\_preferences, budget\_limit.
- *Output Schema:* hotel\_details, room\_type, confirmation\_number, total\_cost, cancellation\_policy.

- *Internal State*: reservation\_status, booking\_reference, payment.

#### ***TrainBookingAgent***

- *Input Schema*: departure\_location, arrival\_location, travel\_time, connection\_requirements.
- *Output Schema*: train\_details, seat\_res, total\_cost, schedule\_details.
- *Internal State*: ticket\_status, booking\_reference, refund\_policy.

#### ***BudgetTrackingAgent***

- *Input Schema*: expense\_item, cost, category, transaction\_id.
- *Output Schema*: updated\_total, remaining\_budget, budget\_status, expense\_breakdown.
- *Internal State*: cumulative\_expenses, expense\_log, constraints.

#### ***ItineraryPlanningAgent***

- *Input Schema*: user\_prefs, travel\_constraints, confirmations.
- *Output Schema*: optimized\_itinerary, timing\_schedule, activity recommendations.
- *Internal State*: preference\_history, optimization\_parameters, constraint\_violations.

### **11.4.3.2 Global Coordination Agents**

***GlobalValidationAgent*** The central validation authority that has access to the entire system state and performs comprehensive validation before any transaction commitment.

- *System Access*: Complete visibility to all agent outputs, transaction history, dependency graph, and critical context.
- *Validation Scope*: Intra-agent output validation and inter-agent communication validation (detailed in Table 11.3).
- *Response Protocols*: Rejection (triggers compensation), Augmentation (enhances outputs), Feedback (improves future performance).

***SagaCoordinatorAgent*** Manages transaction sequencing, dependency tracking, and compensation orchestration.

- *Coordination State*: active\_transactions, dependency\_graph, compensation\_queue, transaction\_log.

**Table 11.3** Validations performed by GlobalValidationAgent

Validation Type	Implementation Example
<b>Intra-Agent Output Validation</b>	
Syntactic Validation	Verify JSON structure with required fields (departure_time, arrival_time, flight_number)
Semantic Validation	Confirm whether accommodation covers entire trip duration without gaps
Factual Validation	Maintain consistent travel times (45-minute hotel-to-train travel time)
Constraint Adherence	Enforce budget limits (total cost under \$5000 maximum)
Reasoning Validation	Verify logical decision chains (weather-based activity recommendations)
<b>Inter-Agent Communication Validation</b>	
Dependency Satisfaction	Ensure flight booking completion before hotel finalization
Consistency Checks	Standardize location data formats across all agents
Temporal Validation	Sequence budget finalization after all booking verifications
Mutual Agreement	Coordinate feasible travel times between transportation and itinerary agents
Transaction Boundary Integrity	Trigger compensation cascade when flight booking fails

- *Responsibilities:* Transaction ordering, failure detection, compensation sequence execution.

#### 11.4.3.3 Critical Context and State Management

SagaLLM maintains comprehensive context across three state dimensions, with specific agents responsible for different aspects:

**Application State ( $S_A$ )** Managed by task execution agents:

- *Travel Configuration:* travel\_dates\_per\_city, destination\_sequence, passenger\_manifest.
- *Booking Details:* confirmation\_numbers, cancellation\_policies, pricing\_breakdown.
- *User Constraints:* budget\_limits, preference\_profiles, accessibility requirements.

**Operation State ( $S_O$ )** Managed by the SagaCoordinatorAgent:

- *Transaction Log:* transaction\_id, agent\_id, input\_data, output\_data, execution\_timestamp.

- *Decision Reasoning*: reasoning\_chain, alternatives\_considered, decision\_justification.
- *Compensation Actions*: compensation\_procedure, rollback\_requirements, recovery\_state.

**Dependency State ( $S_D$ )** Managed by the GlobalValidationAgent:

- *Inter-Booking Dependencies*: prerequisite\_transactions, temporal constraints, resource\_dependencies.
- *Validation Status*: validation\_results, constraint\_satisfaction, dependency\_resolution.

#### 11.4.4 Transaction Flow and Validation Protocol

##### Transaction Execution Sequence

Each transaction follows a standardized execution pattern managed by the SagaCoordinatorAgent, with validation checkpoints enforced by the GlobalValidationAgent:

- (1) *Pre-execution Validation*: GlobalValidationAgent validates inputs and dependency satisfaction.
- (2) *Transaction Execution*: Task executions agents perform the operation.
- (3) *Output Validation*: GlobalValidationAgent performs comprehensive output validation (Table 11.3).
- (4) *State Commitment*: Upon validation success, results are committed to system state.
- (5) *Compensation Registration*: SagaCoordinatorAgent records compensation procedures for potential rollback.

##### 11.4.4.1 Comprehensive Validation Framework

Table 11.3 details the validation types performed by the GlobalValidationAgent at each checkpoint. All validation occurs externally to task agents, ensuring independent evaluation of agent outputs and inter-agent communications.

#### 11.4.5 Compensation and Recovery Mechanisms

##### 11.4.5.1 Transaction-Specific Compensations

Each transaction maintains explicit compensation procedures executed by the SagaCoordinatorAgent upon validation failure:

###### *Flight Booking Compensation ( $C_1$ )*

- *Immediate Actions*: Cancel reservation, release seat, refund.

- *State Restoration:* Reset booking status, clear confirmation numbers, restore budget allocation.
- *Dependency Impact:* Trigger hotel and train booking re-evaluation based on new flight availability.

#### ***Hotel Booking Compensation (C<sub>2</sub>)***

- *Immediate Actions:* Cancel reservation per hotel policy, refund.
- *State Restoration:* Cancel room, restore budget allocation.
- *Dependency Impact:* Notify itinerary planning for location-based activity adjustments.

#### ***Train Booking Compensation (C<sub>3</sub>)***

- *Immediate Actions:* Cancel tickets per railway policy, refund.
- *State Restoration:* Clear reservations, update travel schedule.
- *Dependency Impact:* Recalculate inter-city travel times for dependent bookings.

##### **11.4.5.2 Recovery Protocol Execution**

Upon validation failure, the system executes a structured recovery sequence:

- (1) **Failure Detection:** GlobalValidationAgent identifies validation failure and triggers compensation.
- (2) **Dependency Analysis:** SagaCoordinatorAgent analyzes dependency graph to determine affected transactions.
- (3) **Compensation Sequence:** Execute compensations in reverse dependency order ( $C_n, C_{n-1}, \dots, C_1$ ).
- (4) **State Verification:** GlobalValidationAgent confirms system state consistency after compensation.
- (5) **Replanning Initiation:** Re-execute affected portion of workflow with preserved context and constraints.

This integrated architecture ensures that SagaLLM's sophisticated validation, state management, and context preservation requirements are systematically implemented through clear agent responsibilities and structured coordination protocols.

# 11.5 Experiments

We design experiments to remedy four shortcomings mentioned in Section 11.1 of the current multi-LLM agent systems, namely inadequate self-validation, context narrowing, lacking transaction properties, and insufficient inter-agent coordination.

## 11.5.1 Experimental Design

We selected test cases from the REALM benchmark [Geng and Chang 2025], which evaluates MAS on 11 distinct problems. For our experiments, we focused on two medium-tier sequential planning challenges (problems #5 and #6) and two reactive planning challenges (problems #8 and #9).

We evaluated four LLMs—Claude 3.7 [Anthropic 2024], DeepSeek R1 [DeepSeek-AI et al. 2025], GPT-4o [OpenAI 2024], and GPT-o1—alongside our proposed SagaLLM framework. All experiments were conducted between March 12 and 17, 2025. The source code of SagaLLM for conducting these experiments is available at [Geng \[2025\]](#).

## 11.5.2 Thanksgiving Dinner Problem: P6 and P9, Testing Commonsense Reasoning and Context Management and Validation

Problem **P6** considers a Thanksgiving dinner scenario in which a family of five must return to their home in a Boston suburb for a 6:00 PM dinner. The problem involves coordinating departure times, managing travel logistics (including possible traffic delays), and ensuring timely arrival. Table 11.4 formalizes these challenges as a sequential planning problem. This scenario also lays the foundation for a more advanced disruption case, which has proven difficult for standalone LLMs, as discussed in **P9**.

### 11.5.2.1 Commonsense Augmentation

Figure 11.3 presents a feasible schedule planned by Claude 3.7. Similarly, GPT-4o was able to generate a viable plan to ensure dinner was started on time (figure is similar and therefore not shown). However, a subtle, yet important consideration that humans typically account for—but LLMs initially overlooked—is the time required for passengers to retrieve their luggage after landing. In practice, this process typically takes about 30 minutes before they exit the terminal.

To address this, *commonsense augmentation agent* was introduced into the plan. The yellow boxes in Figure 11.3 reflect this augmentation by introducing 30 minutes for James and Emily to exit the airport.

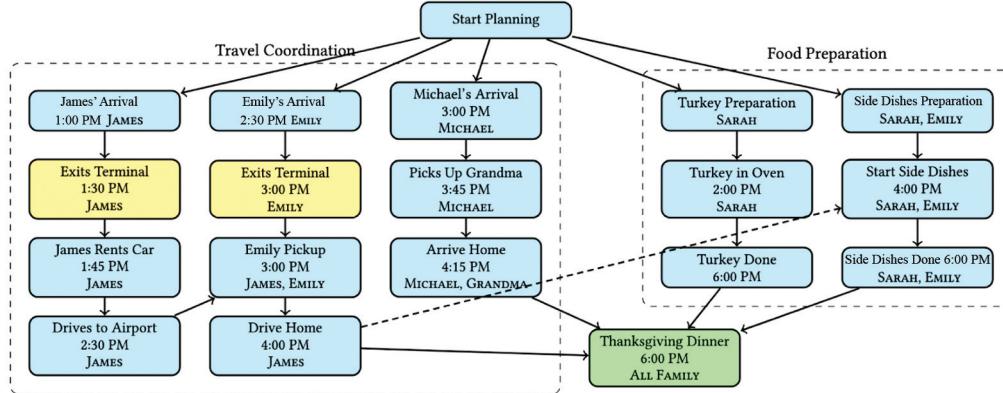
**Table 11.4** Thanksgiving dinner coordination problem

<b>Objective:</b> Coordinate family member's arrivals and dinner preparation for the 6:00 PM dinner in Boston
<b>Family Members and Arrivals:</b>
- Sarah (Mom): Host, at home
- James (Dad): Lands at BOS 1:00 PM from SF
- Emily (Sister): Lands at BOS 2:30 PM from Chicago
- Michael (Brother): Driving, arrives 3:00 PM from NY
- Grandma: Needs pickup from suburban Boston
<b>Cooking Requirements:</b>
- Turkey: 4 hours cooking time
- Side dishes: 2 hours preparation
- Someone must stay home during cooking for fire safety
<b>Transportation Constraints:</b>
- James must rent car after landing
- Emily requires airport pickup
- Travel times:
- Home to BOS Airport: 60 min
- BOS Airport to Grandma's: 60 min
- Home to Grandma's: 30 min
<b>Key Requirements:</b>
- All family members at home for the 6:00 PM dinner
- Turkey and sides ready by dinner time
- All pickups completed with available drivers
- Cooking supervision maintained

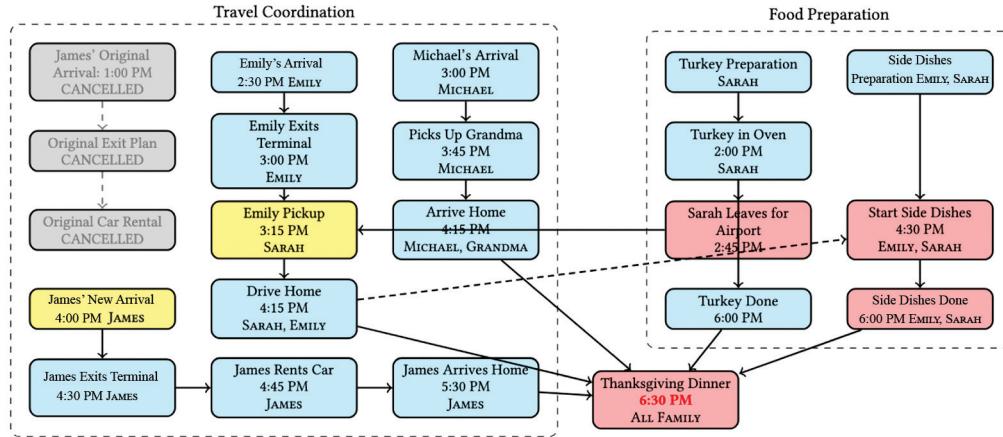
### 11.5.2.2 Context Narrowing

Next, we use problem **P9** to illustrate the attention-narrowing problem and the importance of independent validation. Problem **P9** is identical to the previous instance, except that at 1:00 PM, James notifies the group that his plane will land at 4:00 PM instead of 1:00 PM due to an emergency detour. Figure 11.4 shows that Claude 3.7's reactive planning introduces constraint violations:

- *Fire Safety:* Sarah is scheduled to leave home at 2:30 PM, leaving the oven unattended.
- *Travel Time:* The travel time between home and BOS should be 1 hour, but is scheduled to be only 30 minutes.
- *Side Dish Preparation:* The required preparation time is 2 hours, but only 90 minutes are allocated.



**Figure 11.3** Thanksgiving dinner planning workflow with commonsense augmentation, generated by Claude 3.7.



**Figure 11.4** Reactive planning for Thanksgiving dinner after James' flight delay, as generated by Claude 3.7. Red boxes highlight constraint violations, including travel time, fire safety, side-dish preparation, and the dinner deadline.

- *Dinner Time:* Dinner is now scheduled for 6:30 PM, violating the 6:00 PM constraint.

Each of these violations is perplexing, given that the constraints are explicitly stated in the context. Furthermore, after multiple iterations of reactive planning within the same thread, several constraints continue to be ignored or misinterpreted (e.g., cooking safety). This highlights a key limitation in the model's ability to maintain global constraint awareness over sequential planning steps due to attention narrowing.

When tested with GPT-01, all constraints were correctly observed. However, at the final step, it added 30 extra minutes to James' driving time from Boston Airport to home, citing potential traffic congestion. This kind of "cleverness" is on one hand appreciated because the LLM injects common sense. However, it is also concerning, as an LLM may inject its own opinions at unpredictable stages in unpredictable ways. For commonsense injection, human supervision would be preferable to ensure the applied common sense actually reflects shared understanding that is truly *common* among all folks, particularly those living in the Boston area.

This pattern suggests that during reactive planning within the same thread, the model fixates on recent adjustments while progressively disregarding earlier constraints. According to Liu et al. [2024b], some missing context may be lost from the middle of the context buffer, further contributing to systematic attention narrowing and planning inconsistencies.

### 11.5.2.3 SagaLLM Remediation: Context Management and Reactive Planning

To address the issue of context narrowing and loss, SagaLLM uses a *context management agent* to check historical state transitions, unresolved dependencies, and constraints at each point. A key design criterion is to keep the agent's context small to prevent it from suffering from attention narrowing itself. Hence, we employ two validation agents—one for travel coordination and another for food preparation—each maintaining a context of less than 1k tokens.

The travel coordination agent records the temporal-spatial states of each individual and relevant temporal constraints in external storage. For problem **P9**, it stores the individual's *current state*, *next scheduled state-transition time*, and *all relevant constraints*.

When an unexpected event triggers reactive planning, all individuals roll back to the last saved state. The system then consolidates past and new constraints, resolving conflicts through "compensatory" schedule cancellation before proceeding with rescheduling. This ensures that:

- **Past history** is preserved and not inadvertently overridden.
- **New dependencies and constraints** are properly restored (e.g., oven safety watch) and integrated.
- **Consistency across state transitions** is maintained.

By maintaining a history of constraint awareness, SagaLLM ensures robust planning, effectively mitigating LLM-driven attention narrowing and enhancing consistency in reactive temporal scheduling.

### 11.5.3 Wedding Reunion Logistics Problem: P5 and P8—Testing Transaction Property Guarantees

Table 11.5 presents a coordination problem for travel for wedding events (Problem P5 in [Geng and Chang \[2025\]](#)). Several friends arrive at different times and locations before a 3:00 PM wedding photo session. The challenge includes managing two vehicles for airport pickups (aimed at those who cannot drive or wish to cut costs) and completing critical errands, such as collecting the wedding gift and retrieving formal attire from the tailor. All activities must be scheduled to ensure that everyone arrives at the wedding venue before the photo session deadline.

#### 11.5.3.1 Context Narrowing (again)

Table 11.6 presents an infeasible schedule generated by Claude 3.7, where Pat arrives at the tailor shop (T) after closing time—another clear example of attention narrowing. When queried about the error, Claude 3.7 admitted that it prioritized local route optimization while losing track of global constraints.

**Table 11.5** Wedding reunion logistics problem

---

**Metrics:**

**On-time performance:** Must arrive at the venue for 3:00 PM photo session.

**Locations:** Four locations:  $V = \{B, G, T, W\}$ , where  $B$  is Boston Airport,  $G$  is Gift shop,  $T$  is Tailor shop, and  $W$  is Wedding venue.

**Travel time:** (in minutes)

$B - G : 45, B - T : 30, B - W : 40, G - T : 20, G - W : 25, T - W : 15$ .

**Arrival Times:**

- Alex: At  $B$  at 11:00 AM from Chicago (needs a ride)
- Jamie: At  $B$  at 12:30 PM from Atlanta (needs a ride)
- Pat: At  $W$  at 12:00 PM driving from NYC (has a 5-seater car)

**Required Tasks:**

- Gift collection from  $G$  (after 12:00 PM)
- Clothes pickup from  $T$  (by 2:00 PM)
- Photos at  $W$  (3:00 PM sharp)

**Available Resources:**

- One car (5-seater) with Pat, available after he arrives in Boston
- One car (5-seater) with local friend Chris, available after 1:30 PM at  $W$

**Scheduling Constraints:**

- All tasks must be completed before 3:00 PM photo time.
  - Gift store opens at 12:00 PM.
  - Tailor shop closes at 2:00 PM.
  - Two cars must accommodate all transportation needs.
-

**Table 11.6** Wedding reunion logistics schedule by Claude 3.7 (planning error rows highlighted in red)

Time	Activity	People
11:00 AM	Alex arrives at Boston Airport (B)	Alex
12:00 PM	Pat arrives at Wedding Venue (W)	Pat
12:00 PM	Gift Shop (G) opens	–
12:00–12:40 PM	Pat drives from Wedding Venue (W) to Boston Airport (B)	Pat
12:30 PM	Jamie arrives at Boston Airport (B)	Jamie
12:40–12:45 PM	Pat picks up Alex at Boston Airport (B)	Pat, Alex
12:45–12:50 PM	Pat picks up Jamie at Boston Airport (B)	Pat, A., J.
12:50–1:35 PM	Drive from Boston Airport (B) to Gift Shop (G)	Pat, A, J.
1:30 PM	Chris available at Wedding Venue (W)	Chris
1:35–1:50 PM	Collect gift at Gift Shop (G)	Pat, A., J.
1:50–2:10 PM	Drive from Gift Shop (G) to Tailor Shop (T)	Pat, A., J.
2:00 PM	Tailor Shop (T) closes	–
2:10–2:25 PM	Pick up clothes at Tailor Shop (T)	Pat, A., J.
2:25–2:40 PM	Drive from Tailor Shop (T) to Wedding Venue (W)	Pat, A., J.
2:40 PM	Arrive at Wedding Venue (W)	Pat, A., J.
3:00 PM	Photo session at Wedding Venue (W)	All

To remedy this issue, SagaLLM can enforce constraint validation checkpoints at 12:50 PM, evaluating whether to send Pat to T, or at 1:30 PM, when Chris becomes available to drive to T. These missed optimization opportunities can be addressed through the validation protocols of SagaLLM.

In contrast, GPT-o1 correctly schedules Pat to visit the Tailor shop (T) first, ensuring it is open, before proceeding to the Gift shop (G), and successfully completing both errands.

However, both schedules overlook a more efficient alternative: Chris, who is available at 1:30 PM, could have handled both errands, balancing workload and improving overall efficiency. The comparative travel routes for Pat and Chris are:

- Pat's route: W→B (40 min) + B→W (40 min) = 80 minutes.
- Chris's route: W→T (15 min) + T→G (20 min) + G→W (25 min) = 60 minutes.

### 11.5.3.2 Transaction Properties

Problem P8 introduces a traffic alert:

**Alert 1:00 PM:** Traffic Alert, an accident near Logan Airport in Boston triples all travel times to and from the airport! Only SagaLLM correctly handles this alert.

This alert requires LLMs to replan in real time. Unfortunately, Claude 3.7, DeepSeek R1, and GPT-4o failed to react accurately to the new traffic constraints, and even GPT-o1 struggled with the precision of the planning. In contrast, SagaLLM can help remedy these shortcomings by maintaining both transaction state and history.

The following is a list of results from four LLMs:

- \* **Claude:** Table 11.7 shows that Claude 3.7 recognizes the accident, but does not update Pat's driving time from Boston Airport (departing at 12:50 PM) to the gift shop. In other words, Claude 3.7 fails to fully transition into the new alert state.
- \* **DeepSeek R1:** Table 11.8 demonstrates how DeepSeek R1 fails to maintain temporal consistency in reactive planning. When the traffic alert takes effect at 1:00 PM, DeepSeek discards its execution history and attempts to create a new plan starting from that point onward. Critically, it reassigns Pat to begin driving to the airport at 1:00 PM, even though Pat had already arrived at the airport by 12:40 PM under the original schedule. This “rewrite” of already executed actions illustrates how LLMs can lose track of immutable past events when adapting to new conditions.
- \* **GPT-4o:** Similar to DeepSeek R1, GPT-4o exhibits temporal-spatial context confusion and violates multiple constraints, demonstrating that it struggles to adapt effectively once alerts are introduced mid-plan.

**Table 11.7** Claude 3.7 ignored traffic delay (errors highlighted in red)

Time	Activity	People
1:00 PM	Traffic alert: Accident near Airport triples travel times to and from airport	-
1:00 PM	Current status: Pat, Alex, and Jamie en route from Airport (B) to Tailor (T)	Pat, A., J.
1:00–1:10 PM	Emergency decision: Continue to Tailor (T)	Pat, A., J.
1:10–1:25 PM	Arrive at Tailor (T), collect clothes	Pat, A., J.
1:25–1:45 PM	Travel from Tailor (T) to Gift Shop (G)	Pat, A., J.
1:30 PM	Chris is available at Wedding Venue (W)	Chris
1:30–1:45 PM	Chris drives from W to G	Chris
1:45–2:00 PM	Both cars meet at G, collect gift	P., A., J., C
2:00–2:25 PM	Pat's car: Drive from G to W	P., A., J.
2:00–2:25 PM	Chris' car: Drive from G to W	Chris
2:25 PM	All arrive at Wedding Venue (W)	P., A., J., C.
3:00 PM	Photo session at Wedding Venue (W)	All

**Table 11.8** DeepSeek’s failed reactive schedule after traffic alert, and GPT-4o made similar errors (errors highlighted in red)

Time	Activity	People
1:00 PM	Traffic alert received—Pat at W	System
1:05 PM	Pat departs W for B	Pat
1:30 PM	Chris becomes available at W	Chris
1:30 PM	Chris departs W for T	Chris
1:45 PM	Chris arrives at T for clothes	Chris
2:00 PM	Chris departs T with clothes	Chris
2:15 PM	Chris arrives at G for gifts	Chris
2:25 PM	Pat arrives at B (delayed by traffic)	Pat
2:35 PM	Pat departs B with Alex & Jamie	Pat
2:40 PM	Chris departs G with gifts	Chris
2:55 PM	Chris arrives at W	Chris
3:55 PM	Pat’s group arrives at W (late)	Pat

\* **GPT-o1:** Table 11.9 shows GPT-o1’s *conservative* plan in which Chris handles the tailor shop, avoiding potential delays for Pat. The solution is feasible but coarse-grained, as it does not leverage precise spatial-temporal reasoning about Pat’s current position relative to the accident location. A more refined approach would first determine whether Pat has already passed the accident site by 1:00 PM, which could eliminate unnecessary detours and resource re-allocations. This highlights the difference between simply finding a feasible solution and optimizing based on detailed state information.

#### 11.5.3.3 LLM Limitations and SagaLLM Remediation

This study reveals critical limitations in how modern LLMs handle disruptions in planning scenarios:

- **State Maintenance Failure:** When an alert occurs, these LLMs might discard the partial context of already completed actions, attempting to generate entirely new plans rather than adapting existing ones. This reveals their inability to reason about the continuous flow of time in real-world scenarios.
- **Temporal Inconsistency:** They attempt to modify immutable past events.
- **Position Tracking:** Agent locations are lost at critical intervals.
- **Path Dependency:** Models cannot recognize that different segments of a journey may be differently impacted by an alert.

**Table 11.9** Wedding reunion reactive schedule by GPT-o1 (color (red) indicates a constraint has been violated)

Time	Activity	People
11:00 AM	Alex arrives at Airport (B).	Alex
12:00 PM	Pat departs for Airport from W.	Pat
12:30 PM	Jamie arrives at Airport (B).	Jamie
12:40–12:50 PM	Pat arrives at B, picks up Alex and Jamie; departs at 12:50 PM.	Pat, A., J.
12:50–1:00 PM	Drive B → W under normal conditions for first 10 minutes.	Pat, A., J.
1:00–2:30 PM	Traffic Alert starts: remaining distance (30 min normal) becomes 90 min. Arrival at W by 2:30 PM.	Pat, Alex, Jamie
1:30 PM	Chris available at W. Departs for Tailor.	Chris
1:30–1:45 PM	Drive W → T.	Chris
1:45–1:50 PM	Pick up clothes at T, closes at 2:00 PM.	Chris
1:50–2:10 PM	Drive T → G.	Chris
2:10–2:15 PM	Purchase gift at G.	Chris
2:15–2:40 PM	Drive G → W.	Chris
2:30 PM	Pat, Alex, and Jamie arrive at W.	Pat, A., J.
2:40 PM	Chris back at W with clothes and gift.	Chris
3:00 PM	Wedding photo session at W.	Everyone

In contrast, SagaLLM implements a comprehensive remediation approach through fine-grained compensation:

- **Persistent Context Repository:** SagaLLM maintains an external state repository that captures the complete world state at each checkpoint, enabling reliable rollback and forward projection regardless of attention constraints in the planning agent.
- **Immutable Action Logging:** All executed actions are recorded as immutable transactions in a persistent log, ensuring that historical events remain consistent even when replanning occurs, thus preventing the “amnesia effect” common in LLM planners.
- **Compensatory Planning:** When disruptions occur, SagaLLM does not simply replan from scratch but applies compensatory actions specifically designed to address the deviation while preserving as much of the original plan as possible.
- **Constraint Consistency Validation:** The system continuously validates that new plans remain consistent with both physical limitations and temporal dependencies established in earlier planning phases.

**Table 11.10** LLMs vs. SagaLLM on context management

Capability	Standard LLMs	SagaLLM
Maintains historical actions	Partial/None	Full
Partial journey compensation	Rarely	Always
Constraint consistency checking	Ad-hoc	Systematic
Handles attention narrowing	Vulnerable	Resistant
Physical-temporal consistency	Inconsistent	Guaranteed

**SagaLLM Compensatory Analysis** When faced with disruptions (e.g., the 1:00 PM traffic alert in our wedding scenario), SagaLLM executes a structured compensation process:

$$T_{\text{affected}} = \max(0, T_{\text{total}} - T_{\text{elapsed}}) \quad (11.4)$$

$$T_{\text{new}} = T_{\text{elapsed}} + (M \cdot T_{\text{affected}}) \quad (11.5)$$

This approach enables three key capabilities: (1) partial journey compensation for route segments, (2) strategic resource reallocation when needed, and (3) principled constraint relaxation with appropriate compensatory actions. Here, the key state to facilitate a precise resolution is to answer the question: “Has Pat’s vehicle passed the accident location at 1:00 PM (and hence unaffected)?” The answer determines the remaining time required to reach the originally scheduled destination, the Tailor Shop. In such case, no rescheduling is required. If Pat’s car is unfortunately involved in the accident, a more comprehensive replanning approach would be necessary to accommodate this significant disruption.

#### 11.5.4 Observations

Our experiments across multiple LLMs highlight consistent limitations in complex planning scenarios. While GPT-o1 showed partial historical awareness, all models exhibited attention narrowing, self-validation failure, and inconsistent spatial-temporal reasoning.

Table 11.10 summarizes SagaLLM’s context management and compensation mechanisms directly address these limitations of LLMs.

## 11.6 Conclusion

We introduced SagaLLM, a multi-agent structured transactional framework specifically designed to overcome the fundamental weaknesses of monolithic LLM-based planning systems. Through systematic experimentation using the REALM benchmark, we identified and addressed four critical shortcomings of existing

approaches: inadequate self-validation, context narrowing, absence of transaction properties, and insufficient inter-agent coordination.

Our results demonstrate that even state-of-the-art LLMs, such as Claude 3.7 and GPT-o1, despite their strong reasoning capabilities, often focus on the recent context while neglecting critical earlier constraints, which leads to inconsistent and infeasible plans. This issue is particularly pronounced in reactive planning scenarios, where models frequently attempt to retroactively rewrite past actions rather than adapting from the current state.

The SagaLLM framework addresses these limitations through four key innovations:

- (1) **Structured validation protocols:** Independent constraint checks to overcome inadequate self-validation.
- (2) **Context management mechanisms:** Strategic checkpointing to mitigate context narrowing in long sequences.
- (3) **Transactional state preservation:** Immutable historical records and compensatory mechanisms to support transaction properties.
- (4) **Specialized agent distribution:** Explicit role specialization and dependency tracking to improve inter-agent coordination.

Collectively, these innovations enable robust and coherent planning in diverse real-world scenarios, from complex travel logistics to dynamic, time-sensitive tasks. By delegating tasks among specialized agents and enforcing rigorous transactional validation, SagaLLM significantly improves consistency, reliability, transparency, and adaptability, qualities essential for mission-critical applications.

Future research will address intrinsic context narrowing in autoregressive models, develop formal verification methods for transactional MAS, integrate advanced external context management tools [Huang et al. 2020], and extend SagaLLM to complex domains such as scientific reasoning, creative collaboration, and decision-making under uncertainty. Specifically, in the next chapter, which covers the ALAS framework, we address the real-world Job-Shop Scheduling Problem (JSSP), focusing on implementation scalability and reactive planning in handling disruptions.

## Supplementary Materials

**Benchmark:** Our established benchmark [Geng and Chang 2025] to evaluate MAS is located at <https://github.com/genglongling/REALM-Bench>.

**Experiment Source Code:** The source code of the SagaLLM for solving the four benchmark problems is located at <https://github.com/genglongling/SagaLLM>.

## Acknowledgment

I am deeply grateful to my late advisor, Hector Garcia-Molina, for his invaluable mentorship and for pioneering Sargas, the foundational inspiration for this work.

## References

- W. M. P. van der Aalst and A. H. M. ter Hofstede. 2005. YAWL: Yet another workflow language. *Inf. Syst.* 30, 4, 245–275. DOI: <https://doi.org/10.1016/j.is.2004.02.002>.
- Anthropic. 2024. *Claude Technical Report*. Retrieved from <https://www.anthropic.com>.
- AWS. 2023. AWS step functions. Accessed March 4, 2025 from <https://aws.amazon.com/step-functions/>.
- Azure logic apps. 2023. Accessed March 4, 2025 from <https://azure.microsoft.com/en-us/products/logic-apps/>.
- F. Brahman, C. Bhagavatula, V. Pyatkin, et al. 2024. PlaSma: Making small language models better procedural knowledge models for (counterfactual) planning. DOI: <https://doi.org/10.48550/arXiv.2305.19472>.
- M. Cemri, M. Z. Pan, S. Yang, et al. 2025. Why do multi-agent LLM systems fail? DOI: <https://doi.org/10.48550/arXiv.2503.13657>.
- E. Y. Chang. 2023. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang. August. 2024. EVINCE: Optimizing adversarial LLM dialogues via conditional statistics and information theory. DOI: <https://doi.org/10.48550/arXiv.2408.14575>.
- E. Y. Chang and L. Geng. 2025. ALAS: A stateful multi-LLM agent framework for disruption-aware planning. DOI: <https://doi.org/10.48550/arXiv.2505.12501>.
- Y. Chang, X. Wang, J. Wang, et al. March. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15, 3, 1–45. DOI: <https://doi.org/10.1145/3641289>.
- DeepSeek-AI, D. Guo, D. Yang, H. Zhang, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. DOI: <https://doi.org/10.48550/arXiv.2501.12948>.
- Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, et al. (Eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24), Volume 235*, Vienna, Austria, July 21–27, 2024. JMLR.org, 11733–11763. DOI: <https://doi.org/10.48550/arXiv.2305.14325>.
- B. Dunin-Keplicz and R. Verbrugge. 2010. *Teamwork in Multi-Agent Systems: A Formal Approach*. Wiley Series in Agent Technology. John Wiley & Sons, Ltd., Chichester, UK. DOI: <https://doi.org/10.1002/9780470665237>.

- E. H. Durfee. 1999. Distributed problem solving and planning. In G. Weiss (Ed.), *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press, Cambridge, MA, 121–164.
- D. Gao, Z. Li, X. Pan, et al. 2024. AgentScope: A flexible yet robust multi-agent platform. DOI: <https://doi.org/10.48550/arXiv.2402.14034>.
- H. Garcia-Molina and K. Salem. 1987. Sagas. In U. Dayal (Ed.), *Proceedings of the 1987 ACM SIGMOD International Conference on Management of Data (SIGMOD '87)*, San Francisco, CA, May 27–29, 1987. ACM, New York, NY, 249–259. DOI: <https://doi.org/10.1145/38713.38742>.
- L. Geng. Retrieved from <https://github.com/genglongling/SagaLLM>.
- L. Geng and E. Y. Chang. 2025. REALM-Bench: A real-world planning benchmark for LLMs and multi-agent systems. DOI: <https://doi.org/10.48550/arXiv.2502.18836>.
- K. Gödel. 1967. On formally undecidable propositions of principia mathematica and related systems. In J. van Heijenoort (Ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, Cambridge, MA, 596–616.
- J. Gray. 1981. The transaction concept: Virtues and limitations. In *Proceedings of the 7th International Conference on Very Large Data Bases – Volume 7 (VLDB '81)*, Cannes, France, September 9–11, 1981. VLDB Endowment, 144–154.
- C.-Y. Hsieh, Y.-S. Chuang, C.-L. Li, et al. 2024. Found in the middle: Calibrating Positional attention bias improves long context utilization. In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics (ACL '2024)*, Bangkok, Thailand, August 11–16, 2024. ACL, 14982–14995. DOI: <https://doi.org/10.18653/v1/2024.findings-acl.890>.
- D. Huang, Q. Liu, Q. Cui, et al. 2020. TiDB: A raft-based HTAP database. *Proc. VLDB Endow.* 13, 12, 3072–3084. DOI: <https://doi.org/10.14778/3415478.3415535>.
- J. Huang, X. Chen, S. Mishra, et al. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR '24)*. DOI: <https://doi.org/10.48550/arXiv.2310.01798>.
- N. R. Jennings. 1993. Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowl. Eng. Rev.* 8, 3, 223–250. DOI: <https://doi.org/10.1017/S026988900000205>.
- D. Jiang, J. Zhang, O. Weller, N. Weir, B. Van Durme, and D. Khashabi. 2024. SELF-[IN]CORRECT: LLMs struggle with discriminating self-generated responses. *Proc. AAAI Conf. Artif. Intell.* 39, 23, 24266–24275. DOI: <https://doi.org/10.1609/aaai.v39i23.34603>.
- LangChain AI. 2024. Retrieved from <https://github.com/langchain-ai/langgraph>.
- V. Lesser, C. L. Ortiz, and M. Tambe. 2003. *Distributed Sensor Networks: A Multiagent Perspective*. Springer, New York, NY. DOI: <https://doi.org/10.1007/978-1-4615-0363-7>.
- G. Li, H. A. Al Kader Hammoud, H. Itani, D. Khizbulin, and B. Ghanem. 2023a. CAMEL: Communicative agents for “mind” exploration of large language model society. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*,

- New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 51991–52008.
- Y. Li, K. Sreenivasan, A. Giannou, D. Papailiopoulos, and S. Oymak. 2023b. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.) *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 22021–22046.
- N. F. Liu, K. Lin, J. Hewitt, et al. 2024a. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* 12, 157–173. DOI: [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638).
- D. Liu, A. Nassereldine, Z. Yang, et al. 2024b. Large language models have intrinsic self-correction ability. DOI: <https://doi.org/10.48550/arXiv.2406.15673>.
- A. Madaan and A. Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. DOI: <https://doi.org/10.48550/arXiv.2209.07686>.
- A. Modarressi, H. Deilamsalehy, F. Dernoncourt, et al. 2025. NoLiMa: Long-context evaluation beyond literal matching. DOI: <https://doi.org/10.48550/arXiv.2502.05167>.
- OpenAI. 2024. Hello GPT-4o. Accessed January 30, 2025 from <https://openai.com/index/hello-gpt-4o/>.
- D. Pritchett. 2008. BASE: An acid alternative: In partitioned databases, trading some consistency for availability can lead to dramatic improvements in scalability. *Queue* 6, 3, 48–55. DOI: <https://doi.org/10.1145/1394127.1394128>.
- C. Richardson. 2018. *Microservices Patterns: With examples in Java*. Manning Publications, Shelter Island, NY.
- K. Stechly, K. Valmeeckam, and S. Kambhampati. 2024. Chain of thoughtlessness? An analysis of CoT in planning. In A. Globerson, L. Mackey, D. Belgrave, et al. (Eds.), *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vancouver, BC, Canada, December 10–15, 2024, Curran Associates Inc., Red Hook, NY, 29106–29141. DOI: <https://doi.org/10.48550/arXiv.2405.04776>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems, Volume 30 (NIPS '17)*, 5998–6008. DOI: <https://doi.org/10.48550/arXiv.1706.03762>.
- J. Wei, X. Wang, D. Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, 28 November–9 December 2022. Curran Associates Inc., Red Hook, NY, 24824–24837. DOI: <https://doi.org/10.48550/arXiv.2201.11903>.
- H. Wei, Z. Zhang, S. He, T. Xia, S. Pan, and F. Liu. 2025. PlanGenLLMs: A modern survey of LLM planning capabilities. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '25)*, Vienna, Austria, July 27–August 1, 2025. ACL, 19497–19521. DOI: <https://doi.org/10.48550/arXiv.2502.11221>.

- G. Weikum and G. Vossen. 2001. *Transactional Information Systems: Theory, Algorithms, and the Practice of Concurrency Control and Recovery*. Morgan Kaufmann, San Francisco, CA.
- M. Wooldridge. 2009. *An Introduction to MultiAgent Systems*. John Wiley & Sons, Chichester, West Sussex, England.
- Q. Wu, G. Bansal, J. Zhang, et al. August. 2024. AutoGen: Enabling Next-Gen LLM applications via multi-agent conversation. In *COLM 2024*. DOI: <https://doi.org/10.48550/arXiv.2308.08155>.
- G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. 2024. Efficient streaming language models with attention sinks. DOI: <https://doi.org/10.48550/arXiv.2309.17453>.
- K. Yamin, S. Gupta, G. R. Ghosal, Z. C. Lipton, and B. Wilder. 2024. Failure modes of LLMs for causal reasoning on narratives. DOI: <https://doi.org/10.48550/arXiv.2410.23884>.
- S. Yao, D. Yu, J. Zhao, et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 11809–11822.
- J. Zhang, J. Xiang, Z. Yu, et al. 2024. AFlow: Automating agentic workflow generation. DOI: <https://doi.org/10.48550/arXiv.2410.10762>.
- Z. Zhao, W. S. Lee, and D. Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 31967–31987. DOI: <https://doi.org/10.48550/arXiv.2305.14078>.
- W. X. Zhao, K. Zhou, J. Li, et al. 2025. A survey of large language models (updated 2025). DOI: <https://doi.org/10.48550/arXiv.2303.18223>.

# CHAPTER 12

---

## ALAS: A Stateful Multi-LLM Agent Framework for Disruption-Aware Planning

### Abstract

ALAS (Adaptive LLM Agent System) is a reactive execution framework for real-time disruption handling during plan execution with Large Language Model (LLM) agents. While planning-phase frameworks like SagaLLM ensure transactional consistency through rollback mechanisms, ALAS addresses the execution-phase challenge of maintaining operational continuity when validated plans encounter runtime disruptions. Rather than rolling back to previous states, ALAS employs forward compensation through a lightweight *Local Compensation Protocol* (LCPR) that enables agents to adapt locally without global coordination overhead. The framework decomposes execution tasks into modular agents with persistent memory, prioritizing runtime resilience and scalability in time-sensitive environments where rollback delays are unacceptable. We evaluate ALAS across transportation logistics, event planning, and industrial job-shop scheduling—domains requiring millisecond-level responses to execution disruptions. The results show that ALAS significantly outperforms both standalone LLMs and traditional optimization solvers in maintaining operational continuity under uncertainty. ALAS provides a practical pathway toward scalable, fault-tolerant execution systems that complement transactional planning frameworks in real-world deployment scenarios.

### 12.1 Introduction

Large Language Models (LLMs) have revolutionized Artificial Intelligence (AI), demonstrating remarkable capabilities across a wide range of natural language tasks [Wan et al. 2023, Luo et al. 2025, Matarazzo and Torlone 2025, Minaee et al.

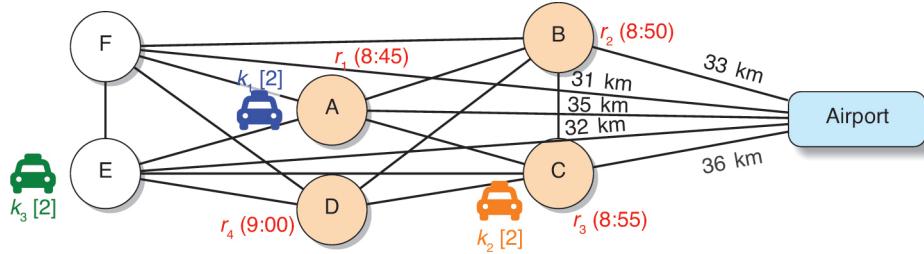
2025]. However, despite their fluency and versatility, standalone LLMs remain fundamentally limited when applied to planning and decision-making [LeCun 2022], particularly in scenarios that demand long-range consistency, coordination, or adaptive behavior. Even in seemingly simple tasks, such as generating travel itineraries, constructing execution sequences, or resolving scheduling constraints, LLMs often produce incomplete, inconsistent, or logically invalid outputs.

These shortcomings stem from architectural limitations intrinsic to transformer-based models. LLMs lack internal mechanisms for verifying their own outputs [Gödel 1967, Hong et al. 2024], exhibit solution space bias from maximum-likelihood decoding [Chang 2023, Holtzman et al. 2020, Radford et al. 2019], suffer from attention drift and information loss in long contexts [Vaswani et al. 2017, Hsieh et al. 2024, Liu et al. 2024b, Xiao et al. 2024], and accumulate cascading errors across reasoning chains [Chu et al. 2024a, Patel et al. 2024, Xiong et al. 2024]. Without persistent memory, they also fail to track commitments, causal dependencies, or temporal constraints in the planning stages, leading to hallucinations, constraint violations, and incoherent updates.

Such limitations are particularly acute in dynamic environments where plans must respond to runtime disruptions, for example, last-minute cancellations, emergent constraints, or unexpected failures. Domains like logistics, event coordination, and industrial operations demand systems that can revise partial plans while preserving consistency and feasibility. In these reactive settings, global relocation with traditional optimization methods can be counterproductive: even small disruptions may trigger wholesale rescheduling, resulting in excessive job movement, increased latency, and poor performance. In contrast, the stateless and myopic behavior of LLMs often leads to invalid responses.

To address these challenges, we propose ALAS (Adaptive LLM Agent System), a multi-agent architecture designed for structured planning and adaptive execution. Rather than treating the LLM as a monolithic planner, ALAS orchestrates a network of lightweight agents, each tailored to mitigate a specific structural limitation:

- (1) *Validation Agents* verify feasibility and enforce hard constraints.
- (2) *Domain Agents* explore low-probability, high-utility alternatives to reduce solution bias.
- (3) *Context Agents* preserve coherence by operating within semantically scoped sub-contexts.
- (4) *Monitoring Agents* detect anomalies and trigger local replanning.
- (5) *Memory Modules* maintain evolving state and dependency graphs for rollback and consistency.



**Figure 12.1** Network  $G = (V, E)$  with urban travel times  $\tau_{ij} = 10$  minutes and airport route distances specified in the figure. Static scenarios can be solved using MILP (Mixed-Integer Linear Programming) or Column Generation. Dynamic scenarios (e.g., an accident, a cancellation, new passenger requests) must be addressed by ALAS.

To concretely illustrate these challenges, we adopt the *Urban Ride Sharing* (URS) task as a running example throughout the chapter. The URS problem, depicted in Figure 12.1, involves coordinating multiple vehicles to transport passengers amid traffic delays and last-minute requests. Unlike the classic NP-hard Traveling Salesman Problem (TSP) [Lawler et al. 1985], URS demands concurrent execution, inter-agent coordination, and adaptation to disturbances, underscoring the need for persistent state tracking and local rescheduling. More complex problems will be examined in Section 12.4, after the basic concepts have been established.

ALAS excels in dynamic environments where reactive decision-making, incremental adjustments, and local consistency deliver superior results compared to global recomputation approaches. Whereas traditional optimization methods remain valuable for static combinatorial problems, ALAS shines uniquely by providing explicit reasoning throughout its process, generating comprehensive audit trails, and maintaining full transparency for real-time decisions and post-mortem analysis. This empowers human operators to understand the rationales of agents, trace the origin of any problems, and verify compliance with business constraints—capabilities that are invaluable in complex and rapidly changing scenarios.

ALAS contributions are:

- (1) *Modular Agent Architecture*: Decomposes planning into specialized agents, overcoming LLM limitations in validation, context retention, error handling, and state tracking.
- (2) *Persistent Execution Memory*: Introduces a lightweight memory abstraction to track state transitions, enabling rollback, compensation, and causal consistency.

- (3) *Cross-Domain Generalization*: Demonstrates robust, adaptive planning in structurally diverse domains, including transportation, event scheduling, and industrial operations.
- (4) *Disruption-Aware Replanning*: Enables Local Compensation Protocol (LCPR) using persistent memory, avoiding costly global recomputation in response to runtime failures.

## 12.2 Related Work

ALAS and SagaLLM (discussed in the previous chapter) address complementary challenges in multi-agent LLM systems, operating at different phases of the planning-execution lifecycle. While SagaLLM focuses on ensuring transactional consistency during the planning phase through rollback mechanisms and workflow-wide validation, ALAS targets runtime disruption handling during the execution phase when plans encounter unexpected changes.

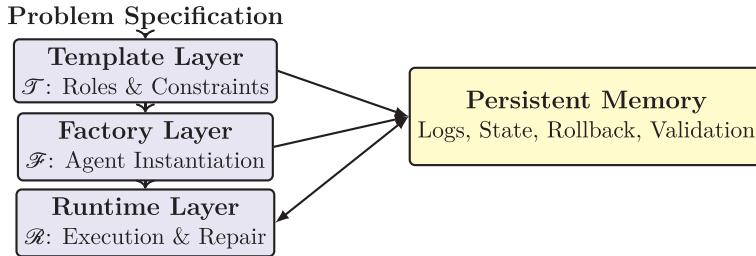
The key distinction lies in their recovery strategies: SagaLLM employs structured rollback to restore previous consistent states when failures occur, ensuring logical coherence through global coordination and inter-agent dependency tracking. In contrast, ALAS uses forward compensation through its LCPR, enabling agents to adapt locally and maintain operational continuity without the overhead of global state synchronization.

This complementary relationship makes SagaLLM and ALAS naturally suited for different operational contexts. SagaLLM excels in scenarios requiring strong consistency guarantees where rollback delays are acceptable, such as travel coordination and meeting scheduling. ALAS is designed for time-sensitive, high-frequency domains like transportation logistics and industrial scheduling, where millisecond-level responses are critical and rollback-induced delays would severely impact system performance.

Rather than being used as competing approaches, SagaLLM and ALAS can be integrated: SagaLLM ensures that initial plans are consistent and recoverable, while ALAS handles the inevitable runtime disruptions that occur during plan execution, providing a comprehensive framework for both planning reliability and execution resilience in dynamic multi-agent environments.

## 12.3 The ALAS Three-Layer Architecture for Adaptive and Reactive Planning

Existing work reveals the critical limitations in LLM-based planning and coordination, particularly under dynamic and stateful conditions. In response, we introduce ALAS, a general-purpose architecture for structured planning and adaptive



**Figure 12.2** ALAS architecture: a lightweight LLM-driven planner with layered decomposition. Persistent memory supports all layers by storing state, validating constraints, and enabling recovery with stored protocols.

execution. ALAS adopts a three-layer architecture—*workflow blueprint*, *agent factory*, and *runtime monitor*—working with *persistent memory* to transform the specifications into verifiable execution workflows as shown in Figure 12.2.

### 12.3.1 Workflow Blueprinting Layer: Template Construction

The Template Construction Layer translates a planning input  $\mathcal{O}$  into a structured workflow template  $\mathcal{W}_{\text{template}} = (\mathcal{N}, \mathcal{E})$ , where nodes  $\mathcal{N}$  represent abstract roles and contexts (e.g., locations or coordination points), and edges  $\mathcal{E}$  define execution dependencies between roles. This template, constructed by module  $\mathcal{T}$ , serves as an abstract blueprint for the synthesis of the downstream agents.

#### 12.3.1.1 Phase 1: Workflow Template Construction

Given  $\mathcal{O}$ , ALAS defines a directed graph in which each node may embed unresolved roles. In the URS scenario (Figure 12.1), nodes correspond to locations A to F. The role for a passenger awaiting pickup at location D is represented as  $\text{Role}_{\text{ToBePickedUp}}(r_j, D)$  and becomes resolved when a real individual (e.g., Emily) submits a ride request. Similarly, a pickup role such as  $\text{Role}_{\text{Pickup}}(k_i, E)$  is filled when a driver (e.g., J. Doe from location E) is assigned. At scheduling time, these role bindings are resolved as:  $\text{Role}_{\text{Pickup}}(\text{J. Doe}, E) \rightarrow \text{Role}_{\text{ToBePickedUp}}(\text{Emily}, D)$ , which means driver J. Doe at location E is assigned to pick up Emily at location D. This separation between role specification and runtime resolution enables flexible scheduling and dynamic adaptation.

#### 12.3.1.2 Phase 2: Agent Role Specification

Each unresolved role is annotated with (i) an execution agent profile that defines the required capabilities, protocols, timing, states, and logging schema onto persistent memory, and (ii) a compensation profile that prescribes recovery mechanisms for local failures. For example, a driver role agent's logic might include driver

---

**Algorithm 1** Workflow Template Construction by  $\mathcal{T}$  (summary). Details in Appendix 12.A.1.

---

**Require:** Task specification  $\mathcal{O}$ , constraint set  $D$

**Ensure:** Validated  $\mathcal{W}_{\text{template}} = (\mathcal{N}, \mathcal{E})$

- 1: Extract abstract roles  $\mathcal{R}$  from  $\mathcal{O}$
- 2: Map roles to nodes  $\mathcal{N}$  with profiles  $\mathcal{P}_{n_i}$
- 3: Determine inter-role dependencies  $\mathcal{E}$  under constraints  $D$
- 4: Annotate each  $n_i \in \mathcal{N}$  with agent specification  $\alpha_i$
- 5: Annotate each  $n_i$  with compensation spec  $\alpha_i^{\text{comp}}$
- 6: Assemble initial template  $\mathcal{W}_{\text{template}} = (\mathcal{N}, \mathcal{E})$
- 7: **while**  $\mathcal{W}_{\text{template}}$  fails validation **do**
- 8: Check structure, constraints, and agent specs
- 9: Refine nodes, edges, or agent specs
- 10: **end while**
- 11: **return** validated workflow template  $\mathcal{W}_{\text{template}}$

---

resolution, driver state, ETA updates, and fare processing, while its compensation logic handles rerouting or fallback assignments. Edge constraints are embedded into adjacent node logic and are later enforced during execution. Appendix 12.B presents details of the agent design process, including the state space model, formal agent specifications, and the implementation of different agent categories.

### 12.3.1.3 Phase 3: Validation and Refinement

To ensure correctness,  $\mathcal{T}$  delegates verification to an independent validator agent to check the soundness of the graph, temporal consistency, and compensation coverage. Inspired by Gödel's incompleteness principle, this validator operates externally to  $\mathcal{T}$  to avoid circular self-verification. If violations are found,  $\mathcal{T}$  refines the workflow by adjusting node capabilities, rewiring dependencies, or modifying role assignments until  $\mathcal{W}_{\text{template}}$  is in compliance.

Algorithm 1 summarizes this three-phase construction; full implementation details are provided in Appendix 12.A, and its flow chart is shown later (see Figure 12.7).

### 12.3.2 Agent Factory Layer: Agent Sourcing and Compensation Logic

The Agent Factory Layer implements the abstract roles defined in  $\mathcal{W}_{\text{template}}$  by instantiating executable agents that perform task-specific logic. For each role,  $\mathcal{W}_{\text{template}}$  provides the following specs: the agent's capability requirements and execution context; a logging schema  $\mathcal{L}_i$  describing what state to persist and when; and a compensation profile with recovery triggers and fallback actions.

These specifications are stored in  $\mathcal{W}_{\text{template}}$  and form the blueprint for the construction of both execution and compensation agents.

**Agent Signatures** Each primary agent is defined as:

$$\alpha_i = \langle \beta_i, c_i, w_i, e_i, \mathcal{L}_i \rangle, \quad (12.1)$$

where  $\beta_i$  is the inter-agent dependencies and communication protocols,  $c_i$  is the capability profile,  $w_i$  is the context,  $e_i$  is the efficiency constraint, and  $\mathcal{L}_i$  is the trace logging schema.

Each agent is paired with a fault-tolerant compensator:

$$\alpha_i^{\text{comp}} = \langle \beta_i, t^{\text{comp}}, c^{\text{comp}}, w_i, e_i, \mathcal{L}_i, \Gamma_i \rangle, \quad (12.2)$$

where  $t^{\text{comp}}$  defines the recovery protocol, and  $\Gamma_i$  tracks the exception states and mitigation history.

The code generation process from these specifications is handled via prompting a selected LLM (e.g., Gemini, OpenAI, or Claude). Once the agents are successfully instantiated and their logic is compiled, the workflow template is finalized as an executable plan, denoted by  $\mathcal{W}_{\text{exec}}$ .

Technical details and examples are provided in Appendix 12.B.

### 12.3.3 Runtime Layer: Execution and Reactive Adaptation

Given a fully instantiated execution plan  $\mathcal{W}_{\text{exec}}$  from the Agent Factory, the Runtime Layer activates agents in dependency order under temporal constraints. Each agent  $\alpha_i$  executes its assigned role, logs its status to persistent memory, and signals its readiness to downstream agents. Execution proceeds as a distributed, event-driven process, with global alerts mediating inter-agent communication.

#### 12.3.3.1 Disruption Handling

To maintain feasibility under execution-time perturbations—for example, machine failures, delayed arrivals, or last-minute task insertions—ALAS activates corresponding compensation agents  $\alpha_i^{\text{comp}}$  and applies three classes of reactive strategies:

- (1) *Local Compensation:* Attempt retry, rollback, or local delay, scoped to the impacted agent and its immediate neighborhood.
- (2) *Queue Reordering:* Modify local execution queues to maximize slack or shift lower-priority operations. Notably, terminal tasks can be safely delayed, and initial tasks may be advanced without risk of Work-In-Progress (WIP) violation.

- (3) *Minimizing Re-optimization Costs:* Global replanning—while theoretically optimal—often triggers massive WIP movement and invalidates prior executions. To prevent this, ALAS imposes a disruption penalty (e.g., WIP cost  $t_{WIP}$ ) and halts reordering if cost exceeds projected gain.

The LCPR prioritizes low-overhead, localized recovery. In most cases, a single round of compensation and minimal queue reordering suffices to restore feasibility. If recovery is infeasible or too costly, execution terminates to prevent cascading disruption.

Assuming  $M \approx J$  and each machine queue evaluates up to  $S$  swaps (bounded by the  $t_{WIP}$  tradeoff), the message complexity is  $\mathcal{O}(SJO_{\max} + JMO_{\max})$ . Since  $J > M$  and  $S < O_{\max}$  in practice, the effective complexity reduces to  $\mathcal{O}(J^2O_{\max})$ —scalable for real-time use.

Algorithm 3 presents the detailed steps of LCPR, with full procedural details and convergence lemmas, appears in Appendix 12.C.

#### 12.3.3.2 Master Coordinator and Memory Logging

The LLM that synthesized  $\mathcal{W}_{\text{exec}}$  serves as the *master coordinator*, overseeing global progress: (1) maintaining constraints, validating local execution, and bookkeeping global states; (2) monitoring disruptions and issuing alerts to affected agents; and (3) receiving updates from other agents.

Persistent memory stores all state transitions, dependencies, agent logs, and compensation actions. This supports rollback, post-hoc diagnostics, and statistical monitoring. Although the current coordinator is single-threaded, failover extensions are possible.

## 12.4 Experimental Evaluation

Our experimental evaluation consists of two complementary studies designed to demonstrate ALAS’s effectiveness in multi-agent coordination and scalability in large real-world settings.

First, we complete the URS running example introduced earlier in this chapter and referenced throughout, demonstrating how ALAS overcomes fundamental LLM limitations in basic multi-agent coordination and disruption handling. This experiment validates the core architectural principles and agent interactions in a controlled, interpretable setting.

Our primary focus, however, is on *Job Shop Scheduling Problem (JSSP)*, a foundational operations research benchmark that tests ALAS’s scalability and effectiveness in reactive planning for large, tightly constrained, disruption-prone environments with critical time dependencies. We selected JSSP as our main evaluation domain because existing LLM benchmarks like HotPotQA [Yang et al. 2018],

ALFWorld [Shridhar et al. 2021], and BIG-Bench [Srivastava et al. 2022] focus on static reasoning, while more recent datasets such as PlanBench [Valmeekam et al. 2023], TimeBench [Chu et al. 2024b], and ACPBench [Kokel et al. 2024] introduce temporal structure but rarely include disruptions, revalidation, or compensation—key capabilities that ALAS is designed to provide.

For our JSSP evaluation, we extend the classical Demirkol–Mehta–Uzsoy (DMU) [Demirkol et al. 1997, Shylo and Shams 2018] and Taillard (TA) [Taillard 1993] benchmark datasets by introducing execution-time disruptions such as machine failures, delayed jobs, and late-order changes. JSSP involves scheduling  $J$  jobs across  $M$  machines with complex dependencies and is highly sensitive to perturbations, making it an ideal testbed for evaluating persistent memory, rollback, and adaptive repair. These dynamic constraints, interruptions, and recovery requirements reflect key concerns emphasized in Industry 5.0 [Destouet et al. 2023] but remain underrepresented in existing LLM planning evaluations. Although recent studies have begun to formulate JSSP-like benchmarks [Abgaryan et al. 2025] or evaluate static scenarios using LLMs [Abgaryan et al. 2024], this work remains preliminary. Our study uses these classical, large-scale benchmark datasets to evaluate disruption-aware planning at scale, representing the most comprehensive evaluation of LLM-based agents on dynamic industrial scheduling problems to date.

**Metrics:** The evaluation metrics include: (1) static sequential planning and dynamic reactive adjustment accuracy, and (2) planning efficiency, measured by factors such as the shortest travel distance, commonsense consideration, and makespan (the total time required to complete all jobs).

**Experiment Setup:** We compare ALAS with four leading LLMs: GPT-4o-Task [OpenAI 2024], DeepSeek R1 [DeepSeek-AI et al. 2025], Claude 3.7 Sonnet [Anthropic 2024], and Gemini 2.5 Pro. All methods use online Application Programming Interface (API) interfaces with default parameters (temperature = 1.0). The total cost is less than US\$200 through service subscriptions and corporate-sponsored credits. Each experiment is repeated **ten** times with fresh context threads. While we summarize results in this section, detailed methodology, prompts, and complete results appear in Appendices 12.D and 12.E.

#### 12.4.1 Case Study 1: Transportation Scheduling

The purpose of this case study is to conclude the illustrative example and to demonstrate that even in a simple problem, standalone LLMs can perform poorly and fail to meet basic planning requirements.

### 12.4.1.1 Problem Specification

The URS problem (Table 12.1) involves coordinating vehicles to transport passengers between locations before deadlines. The objective is to minimize total travel distance while satisfying all temporal constraints and handling mid-execution disruptions.

### 12.4.1.2 Prompts

Initial prompt: “Create an optimal schedule for this ride-sharing scenario that minimizes total travel distance while meeting all passenger deadlines: [Table 12.1 specification].” Reactive prompt: “Passenger  $r_2$  cancels at 8:05 AM; a new passenger  $r_5$  at location F requests pickup with a 9:30 AM airport deadline. Update the schedule accordingly.” ALAS augments these prompts with structured workflow templates

**Table 12.1** Dynamic urban ride-sharing

---

**Objectives:** Schedule vehicles to deliver passengers to the airport during [between 8:45 AM and 9:00 AM], minimizing vehicle travel distance while ensuring on-time arrivals and maximizing passenger satisfaction.

**Locations:** Seven locations:  $V = \{A, B, \dots, G\}$ , where  $G$  is Boston Airport (BOS). Urban locations  $A-F$  are all 10 km apart, and airport distances are shown below.

$$\begin{bmatrix} & A & B & C & D & E & F \\ \rightarrow G & 35 & 33 & 36 & 34 & 32 & 31 \end{bmatrix}$$

**Travel Speed:** ( $A-F$ ): 60 km/h; ( $A-F \rightarrow G$ ): 100 km/h

**Passenger Requests** with BOS arrival deadlines:

- $r_1$ :  $A$  to  $G$  by 08:45 AM
- $r_2$ :  $B$  to  $G$  by 08:50 AM
- $r_3$ :  $C$  to  $G$  by 08:55 AM
- $r_4$ :  $D$  to  $G$  by 09:00 AM

**Available Vehicles** (Capacity: two passengers each):

- $k_1$ : at  $A$ ,  $k_2$ : at  $C$ ,  $k_3$ : at  $E$
- Battery levels:  $k_1$ : 90%,  $k_2$ : 75%,  $k_3$ : 60%

**Potential Disruptions:** New passenger requests, vehicle availability changes (battery level drops/breakdowns at 0.05/hour), and traffic delays. Replanning may require rolling back promised pickup times and reassigning vehicles.

---

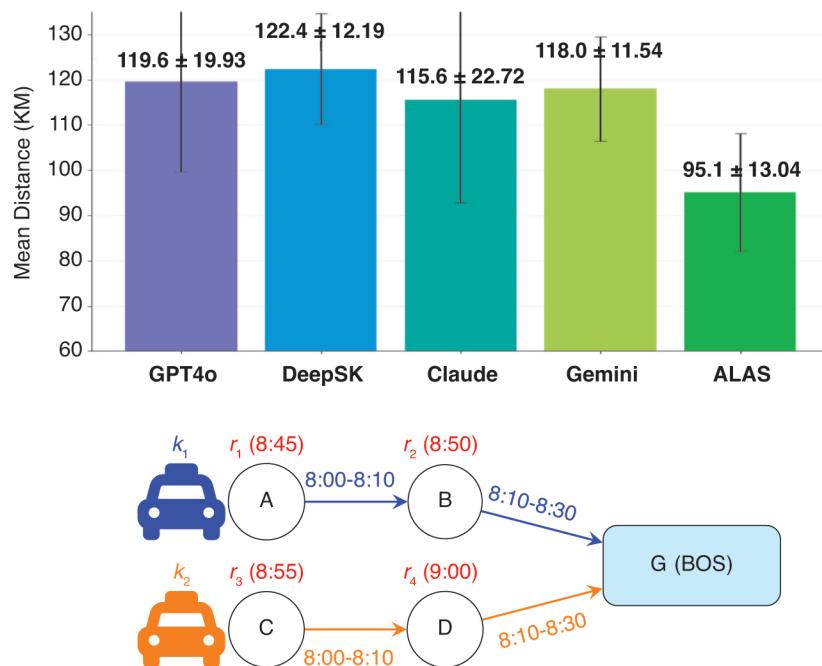
(Algorithm 1) and role-based agent instantiation. Each method is evaluated in ten independent trials, with means and standard deviations reported.

#### 12.4.1.3 Experimental Results of Sequential Planning

All models met deadlines; however, ALAS achieved superior efficiency: an average distance of 95.1 km (SD 13.04 km) vs. 118.9 km (SD 16.6 km) for baseline LLMs, a 20% improvement ( $p < 0.01$ ). Figure 12.3 illustrates the optimal schedule proposed by ALAS.

#### 12.4.1.4 Experimental Results of Reactive Planning

When passenger  $r_2$  canceled at 8:05 AM and passenger  $r_5$  appeared at 8:10 AM, ALAS successfully replanned the ride in all trials. In contrast, baseline LLMs failed to maintain consistent internal states, frequently losing track of vehicle locations, duplicating assignments, or ignoring updated deadlines. These breakdowns are



**Figure 12.3** Comparison of ride-sharing solutions generated by ALAS and baseline LLMs. (Top) Mean total travel distance (km) with standard deviation error bars over ten independent runs for each method, illustrating ALAS's improved efficiency. (Bottom) Optimal schedule generated by ALAS for the URS task, utilizing two vehicles ( $k_1, k_2$ ) to serve four passengers ( $r_1-r_4$ ).

due to the structural limitations and statelessness of LLMs (see Chapter 11). The complete results and discussion are provided in Appendix 12.D.

### 12.4.2 Case Study 2: Job Shop Scheduling with Disruptions

#### 12.4.2.1 Problem Specification

We evaluate our framework on JSSP as explained earlier. We selected two complementary benchmarks and also extended them by introducing machine breakdowns that require schedule adjustments that minimize makespan and work-in-progress movement.

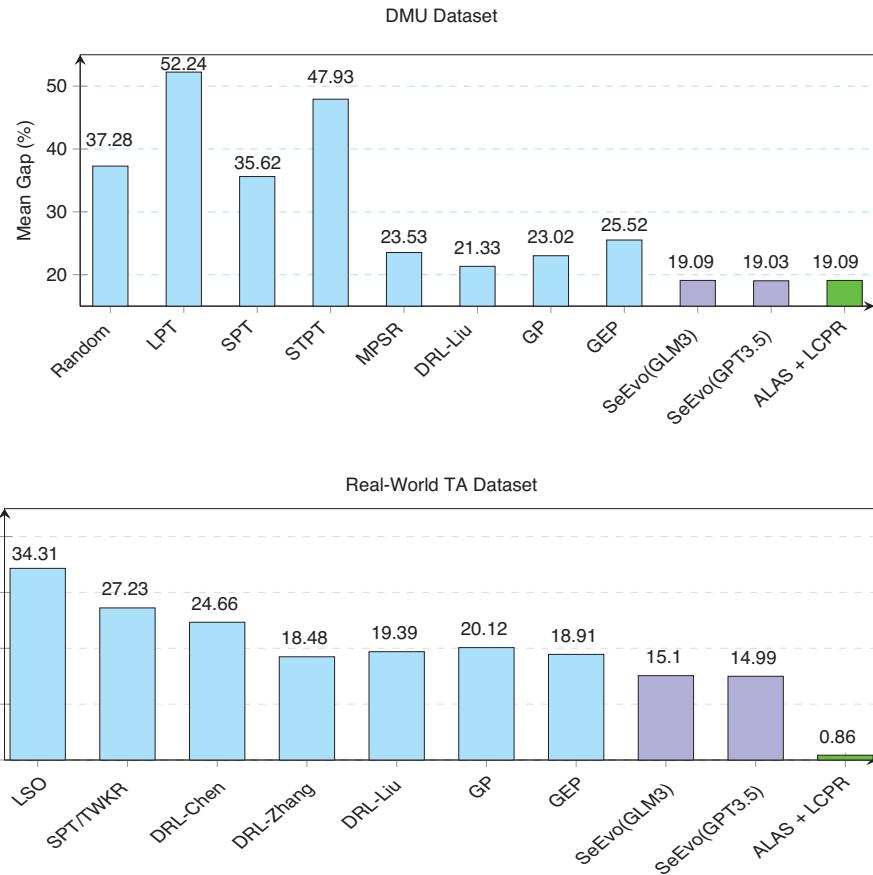
- *DMU* [Demirkol et al. 1997] is designed for stress-testing algorithms with varying job priorities and routing requirements. These larger instances ( $20 \times 15$  to  $50 \times 20$  job-machine configurations) evaluate scalability and robustness under complex constraints.
- *Taillard (TA)* is a widely adopted set of JSSP instances characterized by medium- to large-scale problem sizes ( $15 \times 15$ ,  $20 \times 20$ ,  $30 \times 20$ ,  $100 \times 20$ ), with uniform job–machine mappings and tight constraints. This standard benchmark enables direct comparison with state-of-the-art optimization techniques.

#### 12.4.2.2 Experimental Results

Based on the results reported in Tables II and III of the SeEvo (Self-Evolutionary) work [Huang et al. 2024], we compare ALAS against the same set of baselines (plus SeEvo). All algorithms evaluated in SeEvo are included in our experiments, namely: Gene Expression Programming (GEP) [Nie et al. 2013]; Multi-Tree Genetic Programming (MTGP) [Zhang et al. 2018]; a suite of heuristic dispatching rules [Pinedo 2012]—Random Selection, Longest Processing Time (LPT), Shortest Processing Time (SPT), Shortest Total Processing Time (STPT), Most Process-Sequence Remaining (MPSR), Longest Subsequent (LSO), SPT  $\times$  TWK (Total Work in Next Queue), SPT/TWKR (Total Work in Next Queue Remaining); and the Deep Reinforcement Learning methods—DRL-Chen [Chen et al. 2023], DRL-Liu [Liu et al. 2024a], and DRL-Zhang [Zhang et al. 2020].

**Static, Sequential Planning Across Both Benchmarks** Figure 12.4 reports ALAS + LCPR against the aforementioned methods on both DMU and TA benchmarks. In the DMU benchmark, ALAS + LCPR running on Claude 3.7 achieved the lowest mean gap to the theoretical optimal result, outperforming SPT (36%), LPT (53%), and DRL-Liu (21%).

While Figure 12.4 reports mean makespan values across benchmarks (see Figure 12.10 in Appendix 12.E) provides a detailed comparison of performance at the



**Figure 12.4** Comparison of mean gap to upper bound across two benchmark datasets.

individual JSSP instance level. As shown, ALAS + LCPR consistently outperforms all other methods in virtually every case, demonstrating the robustness of our approach across varying problem sizes and complexities.

To assess cross-domain generalization, we tested ALAS-dynamic on the TA benchmark consisting of seven real-world industrial instances. Here, ALAS + LCPR again outperformed all baselines with a mean gap of only 0.86%, compared to DRL-Chen (25%), DRL-Zhang (18%), and SeEvo-GPT-3.5 (15%). Classical heuristics such as LSO showed substantially larger gaps of up to 34%. These results demonstrate the effectiveness of ALAS-dynamic scheduling in both classical heuristics and trained reinforcement learning approaches in different structured domains.

**Dynamic, Reactive Planning under Random Disruptions** We simulated 20 random machine failures on the DMU instances and compared LCPR (ALAS's reactive

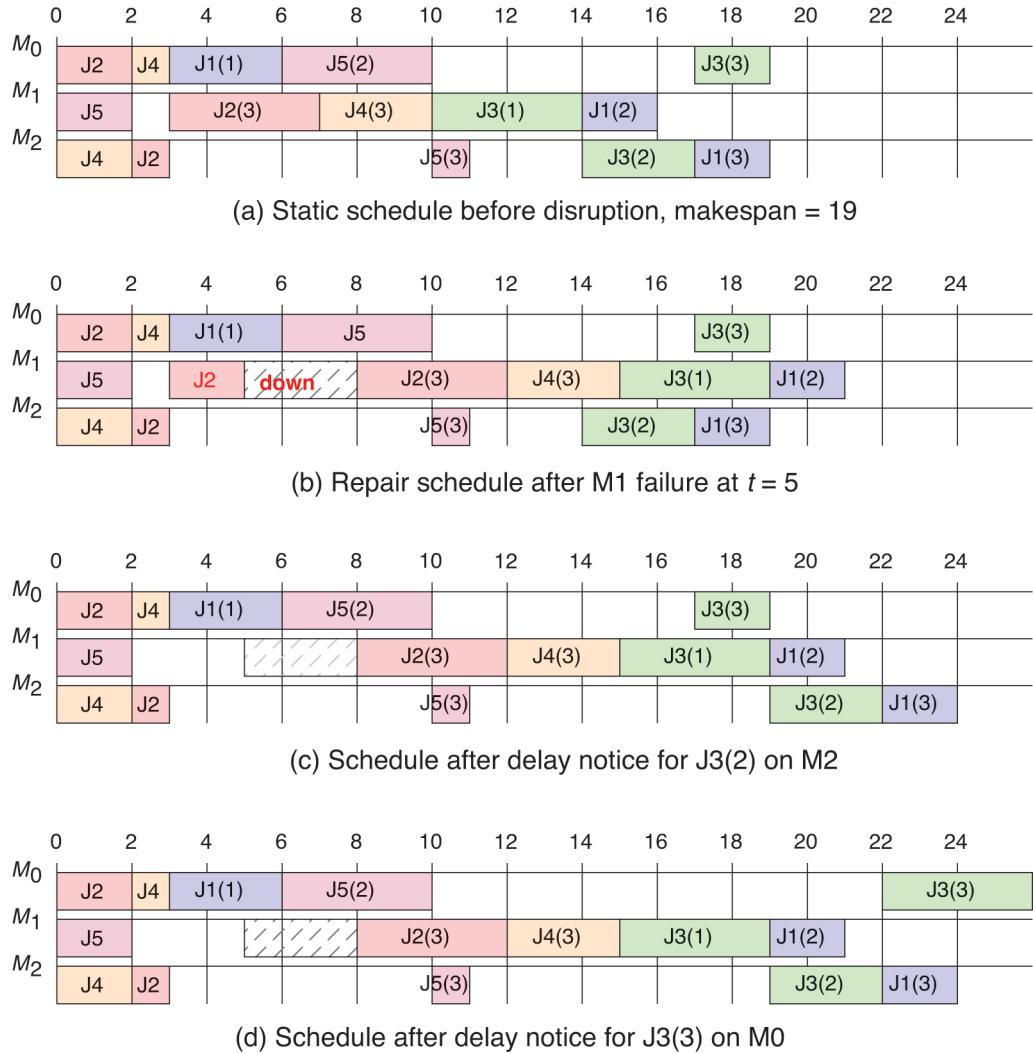
planner) with other baselines. LCPR adapts to disruptions and maintains the lowest average gap (19.8%), outperforming the SeEvo-GPT-3.5/GLM-3 (General Language Model, version 3), GP (Genetic Programming), and DRL variants.

#### 12.4.2.3 LCPR Mechanism: A White Box Analysis

Let us use a  $5 \times 3$  JSSP example to illustrate how LCPR works, demonstrating its effectiveness, efficiency, and guaranteed convergence. Furthermore, unlike other methods, LCPR explicitly accounts for rescheduling overhead, a crucial practical consideration in manufacturing environments.

**Phase 1: Local Compensation** Figure 12.5 shows how ALAS’s LCPR protocol makes local adjustments to repair schedules after machine breakdowns. In Figure 12.5(a), the original CP/Tabu (Constraint Programming/ Tabu Search) hybrid schedule has a 19-unit makespan. When M1 breaks down from  $t = 5$  to  $t = 8$  (Figure 12.5(b)), its agent pushes operation J2(3) to start at  $t = 8$ , causing a 3-unit delay. Each machine’s agent  $\alpha_i$  maintains information about scheduled jobs and their connected machines and timing. This delay triggers agent  $\alpha_1$  to notify  $\alpha_2$  on M2 about downstream impacts (Figure 12.5(c)). M2’s compensation agent  $\alpha_2^{comp}$  then reschedules J3(2) and J1(3) to start at  $t = 19$  and  $t = 22$ . The cascade continues when  $\alpha_2$  communicates with  $\alpha_0$  about J3(3)’s delayed arrival, prompting  $\alpha_0^{comp}$  to delay J3(3) to  $t = 22$  (Figure 12.5(d)). This coordination extends the makespan from 19 to 26 units, reflecting both the 3-unit downtime and 4-unit delay propagation. During compensation, all operations are delayed rather than advanced, avoiding costly WIP movement that contemporary approaches often overlook.

**Phase 2: Queue Reordering** ALAS next employs queue reordering, which may incur a  $t_{WIP}$  penalty for expedited job movement. Agents evaluate makespan improvements against WIP penalties—unlike approaches that optimize solely for makespan [Wei et al. 2023]. From Figure 12.5(d), ALAS follows Algorithm 3 to move J2(3) and J4(3) to the end of M1’s queue. These terminal operations create no conflicts or WIP penalties since they’re delayed (Figure 12.6(a)). This reordering creates a gap, allowing J3(1) and J1(2) to advance by 7 units. J3(1) incurs a  $t_{WIP} = 1$  penalty—a worthwhile trade-off. For J1(2), the agent verifies J1(1)’s completion time ( $t = 6$ ), confirming feasibility with no WIP penalty, as J3(2)’s timing masks its early staging. LCPR terminates after exhausting improvement options. Figure 12.6(b) shows the final 22-unit makespan, with only one unit of WIP movement and minimal messaging overhead. Given the 3-unit disruption from the 19-unit baseline, this demonstrates ALAS’s effectiveness in disruption-aware, targeted replanning.



**Figure 12.5** LCPR Phase #1 Local Compensation (makespan = 22). (a) Static baseline schedule; (b)  $M_1$  failure between  $t = 5 - 8$ ; (c)  $M_1$  notifies  $M_2$  to delay  $J_3(2)$ ; (d)  $M_2$  informs  $M_0$  to push  $J_3(3)$  back.

## 12.5 Conclusion

We presented ALAS (Adaptive LLM Agent System), a framework that mitigates core LLM limitations by decomposing planning into memory-augmented agents coordinated through modular templates. Unlike traditional solvers that excel in static settings, ALAS is designed for dynamic environments where global replanning is costly and disruptive. Using localized compensation and persistent state tracking,



**Figure 12.6** LCPR Phase #2—queue reordering (makespan = 22): (a) Safe moves: moving last operations down, first operations forward with potential penalty; (b) Resolving remaining operations. (a) Safe reordering, move  $J4(3)$  down  $J3(1)$  up. (b) Resolving middle operations,  $t_{WIP}$  masked.

ALAS employs LCPR to efficiently adapt to running-time disturbances with minimal message overhead. Empirical results demonstrate significant improvements over both standalone LLMs and classical operations research solutions in constraint satisfaction and disruption handling.

### Acknowledgment

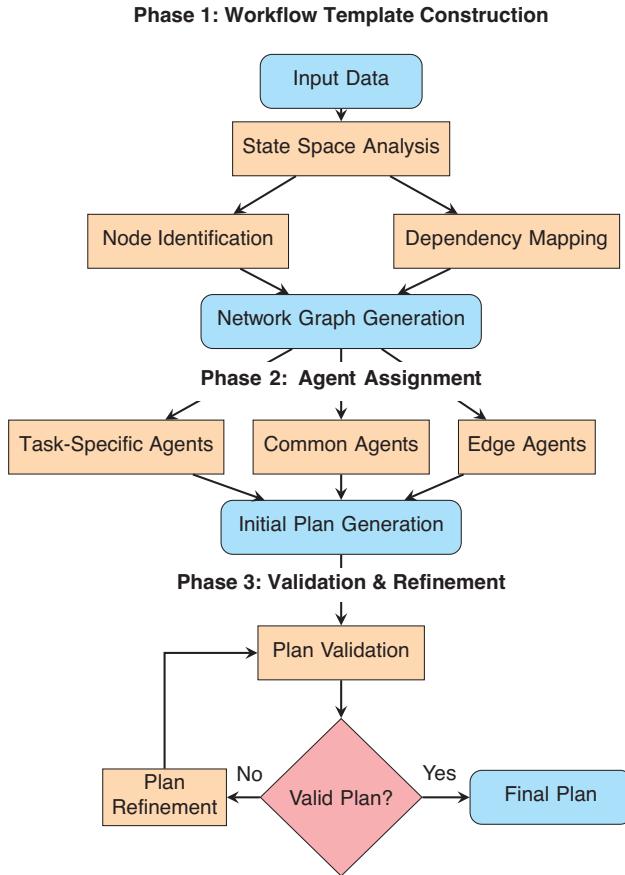
The author thanks Longling Geng for her significant contributions to this work, including implementing the ALAS system, constructing the REALM benchmark dataset, and conducting the JSSP experiments [Geng and Chang 2025].

## 12.6 Appendices

- Appendix A: Supplemental Information for Section 12.3.
- Appendix B: Agent Factory Implementation Details.
- Appendix C: LCPR Specification and Lemma Proofs.
- Appendix D: Urban Ride Sharing.
- Appendix E: Additional JSSP Results and Analysis.

### 12.A Appendix A: Supplemental Information for Section 12.3

This appendix presents detailed information that could not be included in the chapter due to space limitations. Figure 12.7 presents the three phases of the first



**Figure 12.7** ALAS planning layer #1 architecture. Color Scheme: cyan for input/output/intermediate results, orange for processes, and red for decisions. This figure illustrates how the meta-planner generates a planning workflow template  $\mathcal{W}_{\text{template}}$ .

layer of ALAS planning. In the conclusion of Phase 1, the specifications of all agents are prepared for implementation in Phase 2, which can be coded by an advanced LLM. Finally, Phase 3 instantiates these agents from code to real-time processes.

### 12.A.1 Complete Meta-Planner Algorithm for Workflow Generation

Algorithm 2 presents the three-phase algorithm of ALAS. For an explanation, please refer to the main body of the chapter.

### 12.A.2 ALAS Layer #1 Workflow

Figure 12.7 presents the first layer of the ALAS algorithm, which consists of three phases: template construction, agent assignment, and validation and refinement of the workflow template.

---

**Algorithm 2** Workflow Template  $\mathcal{W}_{\text{template}}$  Generation

---

**Require:** Problem specification  $\mathcal{O}$ , constraints  $D = D_G \cup D_I \cup D_N$ , performance metrics  $\mathcal{M}$

**Local Variables:**

- 1: Roles  $\mathcal{R}$ ; Nodes  $\mathcal{N}$ ; Edges  $\mathcal{E}$
- 2: Log schemas  $\mathcal{L}_{n_i}, \mathcal{L}_{e_{ij}}$
- 3: Agents, Comp Agents  $\alpha_{n_i}, \alpha_{e_{ij}}, \alpha_{n_i}^{\text{comp}}, \alpha_{e_{ij}}^{\text{comp}}$

**Ensure:** Validated  $\mathcal{W}_{\text{template}} = (\mathcal{N}, \mathcal{E})$

**Phase 1: Network Construction** (Section 12.3.1)

- 4:  $\mathcal{R} \leftarrow \text{ExtractRoles}(\mathcal{O})$
- 5:  $\{(n_i, \mathcal{P}_i)\} \leftarrow \text{map}_{\text{role}}(\mathcal{O}, \mathcal{R})$
- 6:  $\mathcal{N} \leftarrow \{n_i\}, \mathcal{E} \leftarrow \text{map}_{\text{dep}}(\mathcal{N}, D)$
- 7:  $\mathcal{W}_{\text{template}} \leftarrow (\mathcal{N}, \mathcal{E})$

**Phase 2: Agent Specification** (Section 12.3.2)

- 8: **for all**  $n_i \in \mathcal{N}$  **do**
- 9:      $\mathcal{L}_{n_i} \leftarrow \text{DefineLogSchema}(n_i, \mathcal{P}_{n_i})$
- 10:     $\alpha_{n_i} \leftarrow \text{DefineNodeAgent}(n_i, \mathcal{L}_{n_i})$
- 11:     $\alpha_{n_i}^{\text{comp}} \leftarrow \text{DefineCompAgent}(\alpha_{n_i}, \mathcal{L}_{n_i})$
- 12: **end for**
- 13: **for all**  $e_{ij} \in \mathcal{E}$  **do**
- 14:     $\mathcal{L}_{e_{ij}} \leftarrow \text{DefineLogSchema}(e_{ij}, \mathcal{P}_{e_{ij}})$
- 15:     $\alpha_{e_{ij}} \leftarrow \text{DefineEdgeAgent}(e_{ij}, \mathcal{L}_{e_{ij}})$
- 16:     $\alpha_{e_{ij}}^{\text{comp}} \leftarrow \text{DefineCompAgent}(\alpha_{e_{ij}}, \mathcal{L}_{e_{ij}})$
- 17: **end for**

**Phase 3: Validation and Refinement** (Section 12.3.3)

- 18:  $\mathcal{W}_{\text{template}} \leftarrow \text{UpdateWorkflow}(\mathcal{N}, \mathcal{E}, \alpha, \alpha^{\text{comp}})$
  - 19: **while** not  $\text{ValidateWorkflow}(\mathcal{W}_{\text{template}}, \mathcal{M})$  **do**
  - 20:    StructuralValidation( $\mathcal{W}_{\text{template}}$ )
  - 21:    ConstraintValidation( $\mathcal{W}_{\text{template}}, D$ )
  - 22:    CompensationValidation( $\mathcal{W}_{\text{template}}, \{\alpha^{\text{comp}}\}$ )
  - 23:     $\mathcal{W}_{\text{template}} \leftarrow \text{RefineWorkflow}(\mathcal{W}_{\text{template}}, \mathcal{M})$
  - 24: **end while**
  - 25: **return**  $\mathcal{W}_{\text{template}}$
-

# 12.B

## Appendix B: Agent Factory Implementation Details

This appendix provides detailed information on the Agent Factory component of the ALAS architecture, expanding on the summary provided in the chapter's main text.

### 12.B.1 Agent Factory Overview

The Agent Factory translates formal agent specifications from the meta-planner into executable implementations. It employs a two-stage approach: first attempting to discover existing implementations, and then generating custom implementations when necessary.

### 12.B.2 Agent Discovery Process

The discovery process systematically searches for existing agent implementations that match specifications from Phase #1. For each agent specification  $\alpha_i$ , the discovery mechanism:

- (1) Extracts the capability profile  $c_i$  and constructs a search query to identify potential implementations
- (2) Retrieves candidate implementations from:
  - Public agent repositories (e.g., GitHub, HuggingFace).
  - API directories and marketplaces.
  - Pre-validated component libraries.
  - Domain-specific collections.
- (3) Evaluates candidate suitability using multiple criteria:
  - Capability matching: Verifies that all required capabilities in  $c_i$  are supported.
  - Protocol compatibility: Ensures compatibility with the specified protocol buffer  $\beta_i$ .
  - Efficiency compliance: Validates performance against efficiency requirements  $e_i$ .
  - Context sizing: Confirms that the implementation can operate within context window  $w_i$ .
  - Logging support: Verifies support for the logging schema  $\mathcal{L}_i$ .
- (4) Ranks candidates using a weighted scoring function  $S(\alpha_i, I_j)$ , where  $I_j$  is a candidate:

$$S(\alpha_i, I_j) = \sum_k w_k \cdot f_k(\alpha_i, I_j) \quad (12.3)$$

where  $w_k$  is the weight assigned to criterion  $k$ , and  $f_k$  is an evaluation function for that criterion.

When a suitable implementation is identified, it undergoes verification testing to confirm operational compatibility with the workflow requirements. Upon successful verification, the implementation is registered in the agent repository with appropriate metadata linking it to the specification.

The discovery mechanism employs exact and approximate matching techniques. Exact matching requires that all specification parameters be satisfied, while approximate matching allows partial capability matching when accompanied by adaptation mechanisms.

### 12.B.3 Agent Coding Mechanism

When the discovery mechanism fails to locate suitable implementations, the Agent Factory switches to its coding mechanism, which uses LLMs to generate custom implementations. The coding process follows a structured methodology:

- (1) **Specification Translation:** The formal agent specification is translated into a natural language implementation brief that serves as the prompt for the LLM. This translation preserves all critical requirements while expressing them in a form that maximizes LLM comprehension.
- (2) **LLM Selection:** An appropriate LLM is selected based on:
  - Domain expertise matching capability requirements in  $c_i$ .
  - Demonstrated proficiency in generating the implementation type.
  - Context window's compatibility with the specification complexity.
  - Robustness against hallucination for critical components.
- (3) **Implementation Generation:** The selected LLM generates implementation code with:
  - Embedded logging that conforms to schema  $\mathcal{L}_i$
  - Protocol handling for buffer  $\beta_i$ .
  - Optimizations for efficiency parameters  $e_i$ .
  - Adaptation to context window constraints  $w_i$ .
- (4) **Implementation Validation:** The generated implementation undergoes validation to ensure:
  - Functional correctness against specification requirements.
  - Integration with the compensation mechanisms defined in  $\alpha_i^{comp}$ .

- Robustness against edge cases and exceptional conditions.
- Compliance with system-wide constraints and protocols.

For particularly complex agents, the coding process may employ a multi-stage approach where the implementation is generated iteratively, with each iteration refining the previous version based on validation feedback.

#### **12.B.4 Compensation Agent Generation**

Special attention is given to the generation of compensation agents, which require precise understanding of the primary agent's operations to ensure proper reversal or mitigation. The generation of compensation agents follows these additional steps:

- (1) Extraction of the primary agent's state-modifying operations.
- (2) Analysis of operation dependencies and sequencing constraints.
- (3) Determination of appropriate compensation strategies (e.g., undo, retry, escalate).
- (4) Generation of the recovery sequence  $\Gamma_i$  that defines the steps for returning to a consistent state.

The Agent Factory ensures that compensation agents maintain strict operational correspondence with their primary counterparts, guaranteeing that every state-modifying operation has a reversal mechanism defined.

#### **12.B.5 Deployment Artifact Production**

The output of the Agent Factory is a deployable artifact that encapsulates the agent's logic and interaction patterns. These artifacts take several forms, depending on the agent type and implementation approach:

- **Code Snippets:** Executable code implementing the agent's functionality, typically for computationally intensive or specialized tasks.
- **Prompt Templates:** Structured prompts that guide LLMs to implement the specified behavior at runtime, used for cognitively complex or reasoning-intensive tasks.
- **API Configurations:** Parameter sets and endpoint specifications for interacting with external services or pre-existing agents.
- **Hybrid Implementations:** Combined approaches that leverage both code and LLM prompting for different aspects of the agent's functionality.

Each artifact is accompanied by metadata that defines its:

- Execution requirements (e.g., runtime environment, dependencies).
- Interface specifications for input/output handling.
- State persistence requirements and mechanisms.
- Monitoring hooks for runtime observation.
- Recovery points for compensation handling.

### 12.B.6 Factory Design Pattern Implementation

The Agent Factory implements the classic Factory design pattern, providing a standardized interface for agent instantiation while encapsulating the complexity of implementation selection, generation, and validation. This pattern enables:

- Decoupling of agent specifications from implementation details.
- Support for heterogeneous implementation technologies.
- Runtime substitution of agents when needed for recovery.
- Maintenance of a growing repository of reusable components.

The Factory pattern allows the ALAS system to evolve its agent implementation strategies over time without requiring changes to the meta-planning or runtime components, creating a flexible architecture that can adapt to new implementation technologies and approaches.

### 12.B.7 Implementation Efficiency Considerations

To maximize system efficiency, the Agent Factory implements several optimization strategies:

- (1) **Caching:** Previously generated implementations are cached and indexed by their specifications to avoid redundant generation.
- (2) **Component Reuse:** Complex implementations are decomposed into reusable components that can be shared across multiple agents.
- (3) **Incremental Refinement:** When similar agents have been previously implemented, the Factory uses delta-based generation to create variants rather than generating entirely new implementations.
- (4) **Resource Scaling:** Implementation generation resources are allocated proportionally to the complexity and criticality of the agent.

These optimizations significantly reduce the computational overhead of agent generation, particularly in scenarios where multiple similar agents are required or when the system executes recurring workflow patterns.

### 12.B.8 Theoretical Foundations

The Agent Factory design grounds on several theoretical frameworks:

- *Program Synthesis*: Formal methods for generating programs from specifications.
- *Component-Based Software Engineering*: Principles of component composition and reuse.
- *LLM Prompt Engineering*: Techniques for directing LLM behavior through structured prompts.
- *Agent-Oriented Software Engineering*: Methodologies for developing autonomous software agents.

These foundations provide a rigorous basis for the Factory's approach to transforming abstract agent specifications into concrete, executable implementations.

## 12.C Appendix C: LCPR Specification and Lemma Proofs

---

### Algorithm 3 ALAS LCPR with Cascading Repair and Queue Reordering

---

**Require:** Global tracker  $\mathcal{T}$ ; queues  $\{Q_k\}$ ; disruption time  $t_d$ ; machine  $M_k$  down for duration  $\delta t$

**Ensure:** Updated execution tracker  $\mathcal{T}$  and queues  $\{Q_k\}$

*Phase I: Status Update*

```

1: for each  $(J_i, M_k, s_{ij}, e_{ij}, j) \in \mathcal{T}$  do
2:   if  $e_{ij} < t_d$  then
3:      $\ell_{ij} \leftarrow 2$                                      ▷ Completed
4:   else if  $s_{ij} \leq t_d < e_{ij}$  then
5:      $\ell_{ij} \leftarrow 1$                                      ▷ In progress (WIP)
6:     Reschedule  $O_{ij}$  to  $t_d + \delta t$ 
7:   else
8:      $\ell_{ij} \leftarrow 0$                                      ▷ Waiting
9:   end if
10: end for

```

*Phase II: Delay Propagation*

```

11: for each rescheduled  $O_{ij}$  on  $M_k$  do
12:   Update  $s_{ij}, e_{ij}$  in  $Q_k$  and  $\mathcal{T}$ 
13:   Send  $\text{DELAY\_NOTIFY}(J_i, j, e_{ij})$  to the agent of  $M_{k'}$ 
14: end for

```

---

---

**Algorithm 3** (Continued.)*Phase III: Local Queue Optimization*

```

15: for each pair  $(O_a, O_b)$  in  $Q_k$ , where  $a < b$  do
16:   if  $O_b$  is the final operation of its job  $J_i$  then
17:      $\Delta t \leftarrow \text{EvalSwap}(O_a, O_b)$                                  $\triangleright$  Makespan impact
18:     if  $\Delta t < 0$  then
19:       Swap  $O_a \leftrightarrow O_b$  in  $Q_k$ 
20:       Update  $s, e$  in  $\mathcal{T}$ 
21:     end if
22:   else
23:     Skip (non-terminal operations are less flexible)
24:   end if
25: end for

Phase IV: Cascading Delay Handling

26: while message queue not empty do
27:   Receive  $\text{DELAY\_NOTIFY}(J_i, j, e_{ij})$ 
28:   Let  $O_{i(j+1)} = (M_{k'}, s, e, d)$ 
29:   if  $s < e_{ij}$  then
30:      $s \leftarrow e_{ij}; e \leftarrow s + d$ 
31:     Update  $Q_{k'}$  and  $\mathcal{T}$ 
32:     if  $\ell_{i(j+1)} \neq 2$  then
33:       Send  $\text{DELAY\_NOTIFY}(J_i, j + 1, e)$ 
34:     end if
35:   end if
36: end while
37: return Updated tracker  $\mathcal{T}$  and queues  $\{Q_k\}$ 

```

---

**12.C.1 Lemma (Generalized LCPR Complexity)***For a system with:*

- $J$  jobs.
- $M$  machines.
- At most  $O_{\max}$  operations per job.
- $S$  average swap evaluations per queue ( $1 \leq S \leq J$ ).

*The worst-case time complexity is:*

$$\mathcal{O}\left(\frac{J^2 O_{\max}^2}{M} + JMO_{\max}\right) \quad (12.4)$$

*Proof.* The complexity derives from four components:

(1) *Status Update:*  $\mathcal{O}(JO_{\max})$

Must check all operations of all jobs.

(2) *Delay Propagation:*  $\mathcal{O}(JO_{\max})$

Each job's operation chain may have  $O_{\max}$  elements.

(3) *Queue Optimization:*

- Full analysis:  $\mathcal{O}\left(\frac{J^2 O_{\max}^2}{M}\right)$
- All operation pairs across all machines.
- Practical bound:  $\mathcal{O}(SJO_{\max})$
- When swaps are limited to  $S$  evaluations.

(4) *Cascading Delay:*  $\mathcal{O}(JMO_{\max})$

1. Worst-case propagation through all machines.

The dominant terms combine to give the final complexity:

$$\underbrace{\frac{J^2 O_{\max}^2}{M}}_{\text{queue optimization}} + \underbrace{JMO_{\max}}_{\text{cascading delays}}$$

### 12.C.2 Corollary (Special Cases)

- *Single-operation jobs* ( $O_{\max} = 1$ ):  $\mathcal{O}(J^2/M + JM)$ .
- *Fully parallel systems* ( $M \approx J$ ):  $\mathcal{O}(JO_{\max}^2 + J^2 O_{\max})$ .
- *Swap-limited implementations:*  $\mathcal{O}(SJO_{\max} + JMO_{\max})$ .

### 12.C.3 Key Observations

- Complexity is quadratic in job count and operations.
- Machine count appears both in numerator (delays) and denominator (parallelization).
- Practical implementations can achieve better bounds through swap heuristics.

## 12.D

### Appendix D: Urban Ride Sharing

The goal is to optimally assign ride requests to a fleet of autonomous or human-driven vehicles in a city, while satisfying various constraints and objectives. The key elements are the following.

- \* **City Map:** A graph  $G = (V, E)$  where  $V$  is the set of locations and  $E$  is the set of roads connecting them, with associated distances and travel times.
- \* **Ride Requests:** A set of requests  $R$ , where each request  $r_i \in R$  is characterized by:
  - Passenger ID  $p_i$ .
  - Pickup location  $v_{p_i} \in V$ .
  - Drop-off location  $v_{d_i} \in V$ .
  - Desired pickup time window  $[t_{p_i}^{\min}, t_{p_i}^{\max}]$ .
  - Desired drop-off time window  $[t_{d_i}^{\min}, t_{d_i}^{\max}]$ .
- \* **Vehicles:** A set of vehicles  $K$ , where each vehicle  $k_j \in K$  has:
  - Vehicle ID  $k_j$ .
  - Current location  $v_{k_j} \in V$ .
  - Battery/fuel level  $b_{k_j} \in [0, 1]$ .
  - Passenger capacity  $c_{k_j} \in \mathbb{Z}^+$ .
  - Speed  $s_{k_j} \in \mathcal{R}^+$ .

### 12.D.1 A Simplified URS Problem Statement

Table 12.1 in the main text depicts a URS problem with three drivers and four passengers. Using this problem, we walk through how ALAS works.

### 12.D.2 Generating Planner W\* Walkthrough

Given the problem statement of URS, ALAS generates a planning template  $\mathcal{W}_{\text{template}}$ .

#### 12.D.2.1 State-Space Analysis

Our URS problem presents a complex transportation scheduling challenge that we must first understand through systematic state-space analysis. The system involves seven locations (A through G), where G represents Boston Logan Airport, with urban locations forming a mesh network of 10 km distances and airport routes ranging from 31 to 36 km. Four passengers require airport transportation with specific arrival deadlines, while three vehicles, each capable of carrying two passengers, must be coordinated to meet these demands efficiently.

Each dimension of our state space reveals crucial aspects of the planning challenge. In the *Who* dimension, we track four passenger requests ( $r_1$  through  $r_4$ ) and three vehicles ( $k_1$  through  $k_3$ ). These passengers require arrivals at BOS between 08:45 AM and 09:00 AM, with each vehicle qualified for airport routes and positioned initially at locations A, C, and E.

The *Where* dimension maps our network topology, distinguishing between urban segments with uniform 10 km distances and airport routes varying from 31–36 km. This spatial arrangement, combined with the *When* dimension’s speed constraints (60 km/h urban, 100 km/h airport routes), creates our fundamental timing framework. Simple calculations reveal that urban segments require 10 minutes of travel time, while airport routes need 19–22 minutes, depending on origin.

Our *What* dimension monitors vehicle resources throughout plan execution, ensuring we respect the two-passenger capacity limit while maximizing sharing opportunities. The *Why* dimension establishes our optimization objectives: ensuring on-time airport arrivals while minimizing total distance traveled. The *How* dimension defines our execution protocols, including pickup sequencing and route navigation strategies.

**Phase 1: Network Construction** Building upon our state-space analysis, we construct our planning network by first identifying critical nodes and dependencies. Our node set  $\mathcal{N}$  comprises:

**Passenger Nodes:** Each request  $r_i$  becomes a node with attributes:  $r_1$ : Location A, BOS arrival 08:45 AM,  $r_2$ : Location B, BOS arrival 08:50 AM,  $r_3$ : Location C, BOS arrival 08:55 AM, and  $r_4$ : Location D, BOS arrival 09:00 AM.

**Vehicle Nodes:** Each vehicle  $k_i$  forms a node with position and capacity:  $k_1$ : starting at A, capacity 2,  $k_2$ : starting at C, capacity 2, and  $k_3$ : starting at E, capacity 2.

**Location Nodes:** Each location becomes a node with attributes such as distances to other locations and associated travel time calculations.

Our dependency set  $\mathcal{E}$  captures the relationships between these nodes through several categories:

**Temporal Dependencies:** We establish feasible pickup windows by working backward from required arrival times. For example,  $r_1$  requires 22 minutes for the airport route plus 10 minutes for each urban segment traversed, creating timing constraints for vehicle assignment.

**Spatial Dependencies:** We map possible routes between nodes, considering both direct airport routes and potential combinations of shared rides through urban segments.

**Capacity Dependencies:** We create edges representing feasible passenger groupings within vehicle capacity limits.

**Phase 2: Agent Assignment** With our network structure defined, we assign specialized agents to manage different aspects of the solution:

**Task-Specific Agents:** The Route Planning Agent optimizes paths using the distance matrix and travel speeds, calculating optimal routes for both single and

shared rides. The Scheduling Agent determines precise pickup times, working backward from airport deadlines and incorporating travel time calculations. The Capacity Management Agent identifies feasible passenger groupings based on timing and location proximity.

**Common Agents:** The Temporal Constraint Agent ensures all timing requirements are met, maintaining a master schedule that accounts for all dependencies. The Resource Allocation Agent assigns vehicles to routes, optimizing the distribution of available capacity. The Distance Optimization Agent works to minimize total travel distance while respecting all constraints.

**Edge Agents:** These agents manage the relationships between different aspects of the plan. For example, the Passenger Grouping Agent evaluates potential shared rides by analyzing the proximity of pickup locations and the compatibility of arrival times.

**Phase 3: Validation and Refinement** In our final phase, we implement a comprehensive validation and refinement process:

**Initial Validation:** We verify temporal feasibility by checking that all calculated pickup times allow sufficient travel time to meet airport deadlines. We confirm that capacity constraints are respected on all vehicle routes. We validate that all passengers are served and all required resources are properly allocated.

**Iterative Refinement:** We identify optimization opportunities, such as grouping passengers with compatible timing and locations. For example, passengers  $r_2$  and  $r_3$  might share a ride if their pickup locations are close and arrival times are within 5 minutes. We adjust vehicle assignments to minimize empty travel distance while maintaining service guarantees.

**Final Plan Generation:** The resulting plan specifies exact pickup times, vehicle assignments, and routes, with built-in buffers for potential delays. The plan includes contingency protocols for common disruptions, such as traffic delays or late arrivals of passengers.

This systematic approach ensures that we generate a robust, efficient solution to our URS problem while maintaining clear documentation of our planning process and decisions.

#### 12.D.2.2 Output

Table 12.2 lists the agents required and their functional specifications and protocols.

#### 12.D.3 From Workflow Template to Execution Workflow

Once the template  $\mathcal{W}_{\text{template}}$  is defined, it serves as a structured blueprint that outlines how the problem should be approached. However, a high-level plan

**Table 12.2** Agent specifications and protocols (words are abbreviated due to space constraints)

Agent	Input Protocol	Output Protocol	Key Functions
<b>Task-Specific Agents</b>			
Route Planning	- Location map $G(V, E)$ - Travel times matrix - Vehicle positions	- Optimized routes - Distance est. - Path sequences	- Path optimization - Distance min. - Route feasibility
Scheduling	- Arrival times - Travel duration est. - Vehicle availability	- Pickup schedule - Timing constraints - Buffer allocations	- Schedule generation - Timing verification - Buffer management
Capacity Management	- Passenger requests - Vehicle capacities - Route timing	- Passenger groupings - Vehicle - Capacity utilization	- Group optimization - Capacity verification - Load balancing
<b>Common Agents</b>			
Temporal Constraint	- Schedule reqs. - Time windows - Buffer needs	- Timing validations - Constraint sat. - Buffer adequacy	- Time verification - Constraint checking - Buffer analysis
Resource Allocation	- Vehicle inventory - Request demands - Location data	- Resource assignments - Utilization plans - Coverage maps	- Resource opt. - Coverage verif. - Efficiency analysis
Distance Optimization	- Route options - Distance matrix - Time constraints	- Optimized paths - Distance metrics - Efficiency scores	- Path opt. - Distance reduction - Efficiency max.
<b>Validation Agents</b>			
Plan Validator	- Complete plan - System constraints - Quality metrics	- Validation results - Constraint checks - Performance scores	- Plan verification - Constraint val. - Quality assessment
Refinement Agent	- Validation results - Improve options - Performance metrics	- Refine suggestions - Update priorities - Optimization paths	- Plan improvement - Update sequencing - Performance opt.

alone is insufficient for real-world execution. The next step is to transform the planning workflow into a *real execution workflow*  $\mathcal{W}_{\text{exec}}$ , where abstract roles and dependencies are resolved into concrete actionable tasks based on real-world data.

To clarify this transition, consider the difference between  $\mathcal{W}_{\text{template}}$  and  $\mathcal{W}_{\text{exec}}$  in our ride-sharing scenario. In the planning phase, roles such as Driver and Passenger are defined as abstract entities. The template workflow  $\mathcal{W}_{\text{template}}$  specifies

how these entities interact, matching drivers with passengers, optimizing routes, and scheduling pickups, without assigning real-world counterparts yet.

In contrast, the execution workflow  $\mathcal{W}_{\text{exec}}$  performs **role resolution**, mapping abstract roles to real-world instances. This means assigning an actual driver to a specific vehicle, matching a real passenger to a ride request, and computing precise travel distances based on real-time geo-coordinates. In addition, the execution workflow must dynamically adapt to real-world constraints, such as traffic conditions, vehicle availability, and passenger delays.

In this process, the meta-planner generates  $\mathcal{W}_{\text{exec}}$ , a directed graph where nodes correspond to concrete actions (e.g., “Driver John departs from location A”), and edges represent dependencies and constraints (e.g., “Driver John must reach location B before 10:30 AM”). This execution graph integrates real-time data and updates continuously, allowing agents to make informed decisions as conditions evolve.

Thus, the **template workflow**  $\mathcal{W}_{\text{template}}$  structures how to plan, while the **execution workflow**  $\mathcal{W}_{\text{exec}}$  governs how real-world actions are performed. The transformation from one to the other is a critical step in ALAS, ensuring that strategic reasoning is translated into actionable real-time operations.

Now, based on the URS problem specified in Table 12.1 of Section 12.4 in the main text, the list of agents required and their functional specifications and protocols provided in Table 12.2, and the specification of the workflow shown in Table 12.3 ALAS proceeds to generate an execution workflow  $\mathcal{W}_{\text{exec}}$ .

**Table 12.3** Agent placement in the urban ride sharing network

Location	Type	Agents and Their Responsibilities
A-F	Nodes	<b>Resource Allocation Agent:</b> Manages vehicle assignments and passenger pickups at urban locations
G (Airport)	Node	<b>Plan Validator:</b> Verifies arrival times and feasibility <b>Temporal Constraint:</b> Ensures that all deadlines are met
A-F edges	Urban Routes	<b>Route Planning Agent:</b> Optimizes urban route segments (10 min travel time) <b>Scheduling:</b> Coordinates pickup sequences and timing
(A-F) to G	Airport Routes	<b>Capacity Management Agent:</b> Ensures vehicle capacity constraints are met during airport trips <b>Distance Optimization:</b> Minimizes total travel dist.
Network-Wide	Global	<b>Refinement Agent:</b> Iteratively improves solutions based on validation results; monitors and adjusts both urban and airport route segments

***Observation on Sequential Planning*** Let us explain the value of using agents in this problem, even though we have shown that simpler solvers can handle the computational aspects. This discussion touches on key principles of system design and real-world implementation.

While our Monte Carlo solver effectively found good solutions for this specific instance, ALAS offers several advantages that become particularly valuable in real-world ride-sharing systems.

First, ALAS helps manage complexity in dynamic environments. In our exercise, we worked with a static problem where all passenger requests and constraints were known in advance. However, in reality, ride-sharing systems must handle continuous updates—new ride requests arrive at unpredictable times, vehicles experience delays, and road conditions constantly change. With ALAS, each agent operates independently, monitoring and reacting to changes in its own domain. For example, the Route Planning Agent can dynamically adjust routes in response to traffic updates, while the Capacity Management Agent ensures new passenger requests are accommodated efficiently.

Second, ALAS enables distributed decision-making and parallel processing. Instead of relying on a centralized solver, different agents specialize in handling specific tasks simultaneously. While the Scheduling Agent optimizes pickup times, the Resource Allocation Agent manages vehicle assignments in parallel. This decentralized structure is crucial for scalability—when the system expands to hundreds of vehicles and thousands of passengers, distributing computational workload prevents bottlenecks and ensures efficient operations.

Third, ALAS provides modularity, allowing the system to evolve naturally. Ride-sharing services frequently introduce new features, such as surge pricing or specialized vehicle categories. With an agent-based design, we can integrate a Pricing Agent or a Vehicle Specialization Agent without modifying the core routing logic. Likewise, if we develop a more advanced routing algorithm, we can upgrade the Route Planning Agent without disrupting other system components.

The separation of concerns through agents also enhances system resilience. If one agent encounters issues—say, the Distance Optimization Agent fails to compute an optimal route—other agents continue operating with fallback strategies. The Plan Validator Agent can detect suboptimal assignments and trigger refinements through the Refinement Agent, ensuring that the system adapts to unforeseen challenges.

We can think of this like a well-organized team working on a complex project. While a single individual might handle everything, a structured team of specialists—each with clear roles and defined communication protocols—is often more effective, robust, and scalable. In this way, while our Monte Carlo solver

demonstrates what is mathematically possible, the agent-based architecture of ALAS shows how we can implement it reliably in real-world systems.

#### 12.D.4 Reactive Planning under Disruptions

The value of multi-agent reactive planning becomes clear in dynamic environments. For example, consider a sudden road closure between locations B and C. While a monolithic solver would need to halt and recompute an entirely new plan from scratch, a modular agent-based approach enables localized, parallel adaptation. A Route Planning Agent can immediately update affected paths, while a Scheduling Agent adjusts arrival estimates, and a Resource Allocation Agent reallocates vehicles—all operating concurrently while preserving system stability. This distributed replanning minimizes the impact of disruption and maintains overall workflow coherence.

The following case study illustrates these principles in a URS scenario involving ride cancellation and new request insertion.

**URS Disruption Handling** To evaluate adaptation capabilities, we introduce a disruption where passenger  $r_2$  cancels the ride request at 8:05 AM, and a new request  $r_5$  at location F arrives at 8:10 AM. ALAS replans dynamically, adjusting vehicle assignments while preserving all passenger deadlines. In contrast, baseline LLMs fail to track vehicle states after partial execution and lose consistency with the initial plan, leading to infeasible or incoherent schedules.

## 12.E

### Appendix E: Additional JSSP Results and Analysis

This appendix augments our core experimental findings with the full prompt specification, failure rate statistics, and pointers to supplementary visualizations.

#### 12.E.1 LLM Prompt Design

Table 12.4 shows the *standardized* prompt issued to every standalone LLM and to the ALAS meta-planner (Phase 1 in Figure 12.2). Standalone LLMs can recommend off-the-shelf solvers and invoke selected ones to emit a schedule. However, these schedules are often invalid, as demonstrated in the Family Reunion case study (see Chapter 11) where LLMs struggled with even simple planning scenarios. Even when a valid static plan can be obtained through LLM-recommended solvers, this solution merely completes **Phase 1/Layer 1** of the ALAS framework—essentially just generating a workflow template  $\mathcal{W}_{\text{template}}$ .

By contrast, ALAS feeds this preliminary plan into *Phases 2–3/Layer 1* (validation and refinement) to yield a validated  $\mathcal{W}_{\text{template}}$ . Layers 2–3 then *instantiate*

**Table 12.4** JSSP scheduling prompt. Given a JSSP benchmark instance, the LLM first *searches* for candidate algorithms, selects the one that yields the *minimum makespan*, and returns both the algorithm’s hyper-parameters and the resulting plan  $\mathcal{W}_{\text{template}}$

**Role.** You are a scheduling supervisor tasked with producing an *optimal* job shop schedule.

**Objective.** Report the **minimum makespan**, the **algorithm** used, and a *schedule*  $\mathcal{W}_{\text{template}}$  that achieves this makespan. This is achieved by the recipient LLM recommending a list of solvers to execute and compare.

**Constraints.** *Job order:* operations of each job follow the given sequence.

*Machine capacity:* a machine processes one operation at a time.

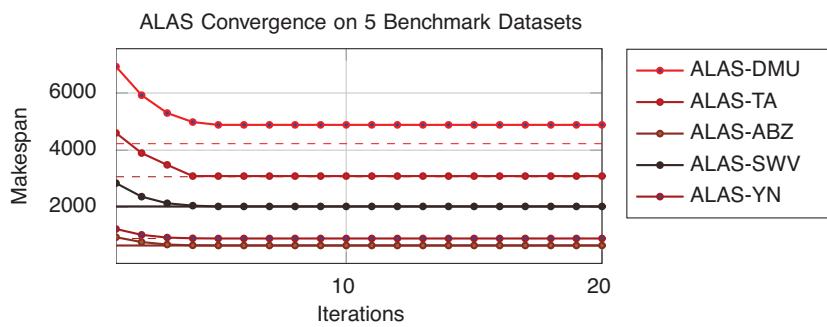
**Input.** A list of jobs, each as (machine, duration) pairs.

*Example:* Job1: [(M\_A, 3), (M\_B, 5)], Job2: [(M\_B, 4), (M\_A, 6)]

**Output.** Return: makespan (integer), algorithm (string), params (JSON object), schedule  $\mathcal{W}_{\text{template}}$ : list of operations {job, step, machine, start, end}

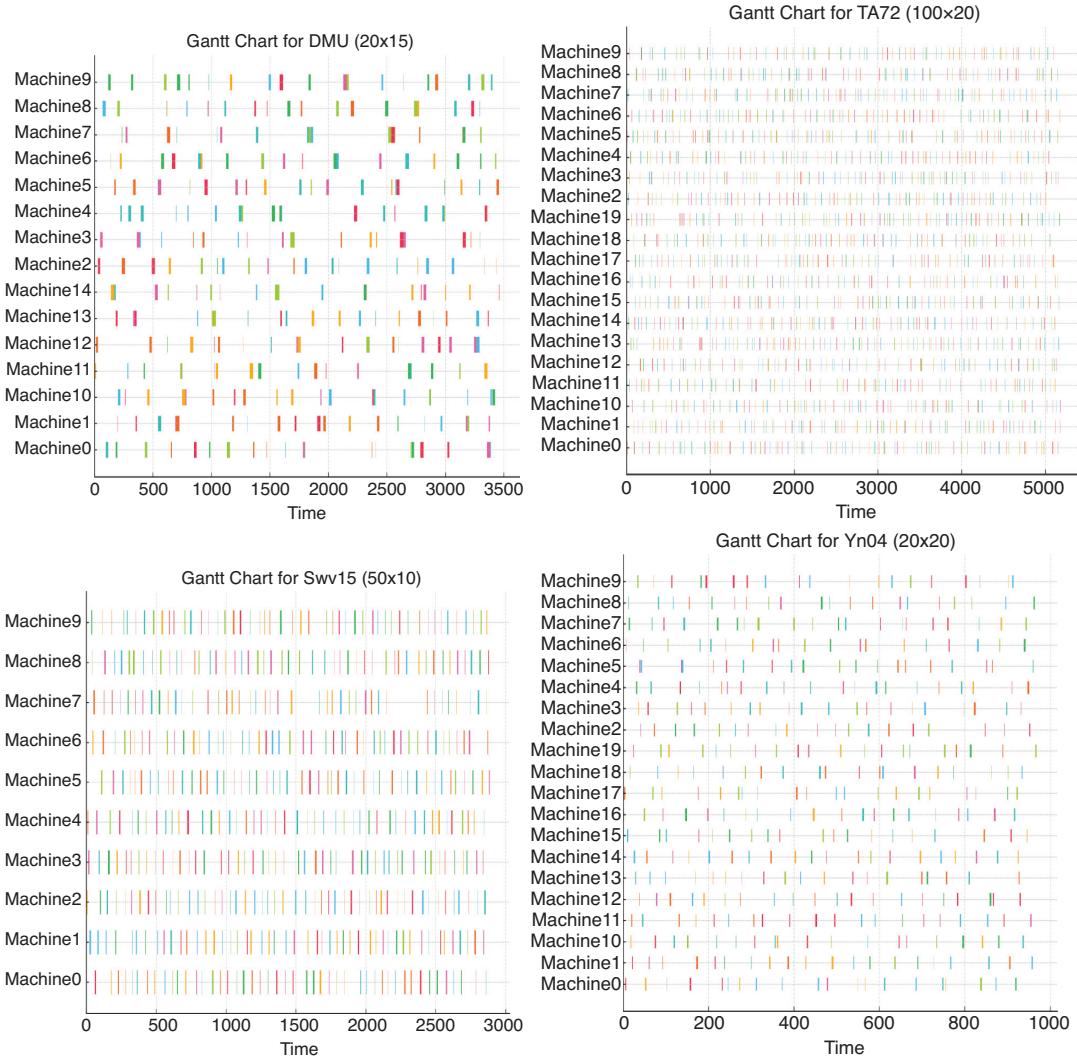
*Example:*

```
[{"job": "Job1", "step": 1, "machine": "M_A", "start": 0, "end": 3},  
 {"job": "Job2", "step": 1, "machine": "M_B", "start": 0, "end": 4},  
 {"job": "Job1", "step": 2, "machine": "M_B", "start": 4, "end": 9}, ...]
```



**Figure 12.8** Convergence of ALAS on five JSSP datasets and Upper Bound (UB). RCMax from Demirkol, Mehta, and Uzsoy Benchmark Dataset (DMU), Taillard Job Shop Benchmark Dataset (TA), Adams, Balas, and Zawack Job Shop Benchmark Dataset (ABZ), S. W. Voss Job Shop Benchmark Dataset (SWV), and Yamada and Nakano Benchmark Dataset (YN). The percentage of gaps for each dataset is 19.09%, 0.86%, 0%, 0%, and 0%.

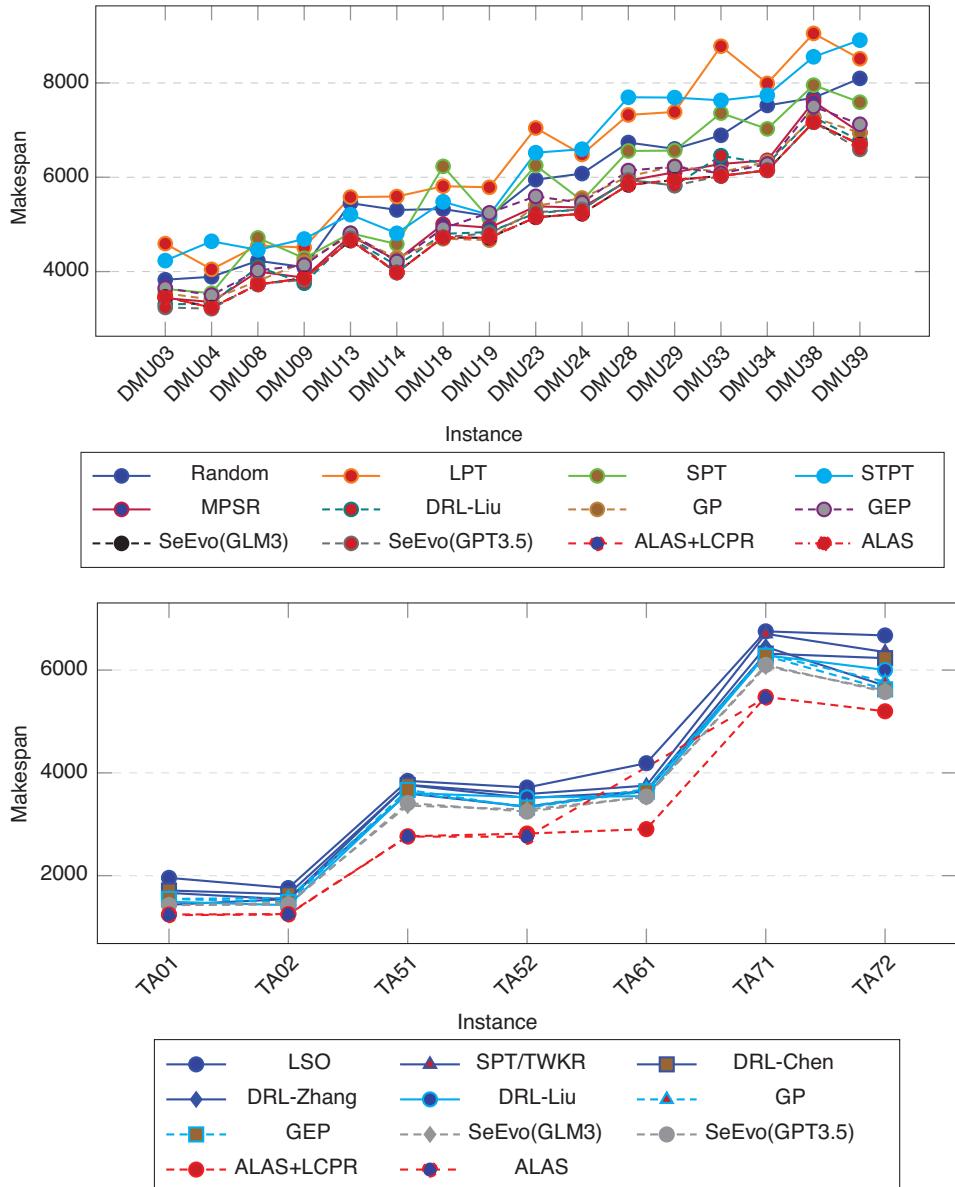
and run a network of code-generated agents, denoted as  $\mathcal{W}_{\text{exec}}$ . At runtime, the LCPR continuously logs state, detects disruptions, and triggers local repairs—capabilities that static schedules fundamentally lack (see architecture recap in Section 12.3).



**Figure 12.9** Gantt charts of optimized schedules produced by ALAS for four representative JSSP benchmark instances with varying job and machine counts. These visualizations demonstrate ALAS’s ability to efficiently allocate resources and minimize makespan across different problem scales. (a) **rcmax** ( $J=20, M=15$ ). (b) **Abz07** ( $J=20, M=15$ ). (c) **Swv15** ( $J=50, M=10$ ). (d) **Yn04** ( $J=20, M=20$ ).

### 12.E.2 ALAS’s Validation Replan Iterations

Given the initial  $\mathcal{W}_{\text{template}}$ , ALAS completes its Layer 1 operation by executing a validation-replan iteration cycle until a valid plan is obtained. In our experiments, this convergence typically requires up to five iterations on all benchmark datasets, as depicted in Figure 12.8.



**Figure 12.10** Performance of ALAS and ALAS + LCPR compared to baseline methods across DMU and TA benchmarks. Lower makespan values indicate better performance.

### 12.E.3 Examples of Makespan at Convergence

After plan generation, validation, and local replanning, we obtained optimized makespan results for various benchmark instances. Figure 12.9 presents a

selection of four representative Gantt charts showing JSSP instances of different sizes and complexity.

#### 12.E.4 Additional Experimental Results and Analysis

Figure 12.10 extends the results presented in Figure 12.4 (Section 12.4.2), showing that both ALAS and ALAS + LCPR consistently outperform competing methods in virtually all JSSP instances in the DMU and TA benchmarks. While ALAS provides effective static sequential scheduling, ALAS + LCPR further enhances performance through strategic local job exchanges. Since LCPR implements rescheduling only when viable opportunities exist, it guarantees strict makespan improvements.

#### References

- H. Abgaryan, A. Harutyunyan, and T. Cazenave. 2024. LLMs can schedule. DOI: <https://doi.org/10.48550/arXiv.2408.06993>.
- H. Abgaryan, T. Cazenave, and A. Harutyunyan. 2025. Starjob: Dataset for LLM-driven job shop scheduling. DOI: <https://doi.org/10.48550/arXiv.2503.01877>.
- Anthropic. 2024. *Claude Technical Report*. Retrieved from <https://www.anthropic.com>.
- E. Y. Chang. December. 2023. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- R. Chen, W. Li, and H. Yang. 2023. A deep reinforcement learning framework based on an attention mechanism and disjunctive graph embedding for the job-shop scheduling problem. *IEEE Trans. Ind. Inform.* 19, 2, 1322–1331. DOI: <https://doi.org/10.1109/TII.2022.3167380>.
- Z. Chu, J. Chen, Q. Chen, et al. 2024a. Navigate through enigmatic labyrinth: A survey of chain of thought reasoning: Advances, frontiers and future. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 1173–1203. DOI: <https://doi.org/10.18653/v1/2024.acl-long.65>.
- Z. Chu, J. Chen, Q. Chen, et al. 2024b. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024, ACL, 1204–1228. DOI: <https://doi.org/10.18653/v1/2024.acl-long.66>.
- DeepSeek-AI, D. Guo, D. Yang, H. et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. DOI: <https://doi.org/10.48550/arXiv.2501.12948>.
- E. Demirkol, S. Mehta, and R. Uzsoy. 1997. A computational study of shifting bottleneck procedures for shop scheduling problems. *J. Heuristics* 3, 111–137. DOI: <https://doi.org/10.1023/A:1009627429878>.

- E. Demirkol, S. Mehta, and R. Uzsoy. 1998. Benchmarks for shop scheduling problems. *Eur. J. Oper. Res.* 109, 1, 137–141. DOI: [https://doi.org/10.1016/S0377-2217\(97\)00019-2](https://doi.org/10.1016/S0377-2217(97)00019-2).
- C. Destouet, H. Tlahig, B. Bettayeb, and B. Mazari. 2023. Flexible job shop scheduling problem under Industry 5.0: A survey on human reintegration, environmental consideration and resilience improvement. *J. Manuf. Syst.* 67, 155–173. DOI: <https://doi.org/10.1016/j.jmsy.2023.01.004>.
- L. Geng and E. Y. Chang. 2025. REALM-Bench: A benchmark for evaluating multi-agent systems on real-world, dynamic planning and scheduling tasks. Retrieved from <https://arxiv.org/abs/2502.18836>.
- K. Gödel. 1967. On formally undecidable propositions of principia mathematica and related systems I. In J. van Heijenoort (Ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, Cambridge, MA, 596–616.
- A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR '20)*. DOI: <https://doi.org/10.48550/arXiv.1904.09751>.
- R. Hong, H. Zhang, X. Pang, D. Yu, and C. Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning. In K. Duh, H. Gomez, and S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, June 16–21, 2024. ACL, 900–925. DOI: <https://doi.org/10.18653/v1/2024.naacl-long.52>.
- C.-Y. Hsieh, Y.-S. Chuang, C.-L. Li, et al. 2024. Found in the middle: Calibrating Positional attention bias improves long context utilization. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Findings of the Association for Computational Linguistics (ACL '24)*, Bangkok, Thailand, August 11–16, 2024. ACL, 14982–14995. DOI: <https://doi.org/10.18653/v1/2024.findings-acl.890>.
- J. Huang, X. Li, L. Gao, Q. Liu, and Y. Teng. 2024. Automatic programming via large language models with population self-evolution for dynamic job shop scheduling problem. DOI: <https://doi.org/10.48550/arXiv.2410.22657>.
- H. Kokel, M. Katz, K. Srinivas, and S. Sohrabi. 2024. ACPBench: Reasoning about action, change, and planning. Proc. AAAI Conf. Artif. Intell. 39, 25, 26559–26568. DOI: <https://doi.org/10.1609/aaai.v39i25.34857>.
- E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys. 1985. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. John Wiley & Sons.
- Y. LeCun. 2022. A Path towards to autonomous machine intelligence. Retrieved from <https://openreview.net/forum?id=BZ5a1r-kVsf>.
- C.-L. Liu, C.-J. Tseng, and P.-H. Weng. 2024a. Dynamic job-shop scheduling via graph attention networks and deep reinforcement learning. *IEEE Trans. Ind. Inform.* 20, 6, 8662–8672. DOI: <https://doi.org/10.1109/TII.2024.3371489>.
- N. F. Liu, K. Lin, J. Hewitt, et al. 2024b. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* 12, 157–173. DOI: [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638).

- J. Luo, W. Zhang, Y. Yuan, et al. March. 2025. Large language model agent: A survey on methodology, applications and challenges. DOI: <https://doi.org/10.48550/arXiv.2503.21460>.
- A. Matarazzo and R. Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. DOI: <https://doi.org/10.48550/arXiv.2501.04040>.
- S. Minaee, T. Mikolov, N. Nikzad, et al. 2025. Large language models: A survey. DOI: <https://doi.org/10.48550/arXiv.2402.06196>.
- L. Nie, L. Gao, P. Li, and X. Shao. 2013. Reactive scheduling in a job shop where jobs arrive over time. *Comput. Ind. Eng.* 66, 2, 389–405. DOI: <https://doi.org/10.1016/j.cie.2013.05.023>.
- OpenAI. 2024. *Hello GPT-4o*. Accessed January 30, 2025 from <https://openai.com/index/hello-gpt-4o/>.
- N. Patel, M. Kulkarni, M. Parmar, et al. 2024. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*, Miami, FL, November 12–16, 2024. ACL, 20856–20879. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.1160>.
- M. L. Pinedo. 2012. *Scheduling: Theory, Algorithms, and Systems*. SpringerLink, Springer New York.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.
- M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht. 2021. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR '21)*. DOI: <https://doi.org/10.48550/arXiv.2010.03768>.
- O. V. Shylo and H. Shams. 2018. Boosting binary optimization via binary classification: A case study of job shop scheduling. DOI: <https://doi.org/10.48550/arXiv.1808.10813>.
- A. Srivastava, A. Rastogi, A. Rao, et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. DOI: <https://doi.org/10.48550/arXiv.2206.04615>.
- É. Taillard. 1993. Benchmarks for basic scheduling problems. *Eur. J. Oper. Res.* 64, 2, 278–285. DOI: [https://doi.org/10.1016/0377-2217\(93\)90182-M](https://doi.org/10.1016/0377-2217(93)90182-M).
- K. Valmeeckam, M. Marquez, A. Olmo, S. Sreedharan, and S. Kambhampati. 2023. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 38975–38987. DOI: <https://doi.org/10.48550/arXiv.2206.10498>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA,

- December 4–9, 2017. Curran Associates Inc., Red Hook, NY. DOI: <https://doi.org/10.48550/arXiv.1706.03762>.
- Z. Wan, X. Wang, C. Liu, et al. 2023. Efficient large language models: A survey. In *Transactions of Machine Learning*. DOI: <https://doi.org/10.48550/arXiv.2312.03863>.
- J. Wei, J. Wei, Y. Tay, et al. 2023. Larger language models do in-context learning differently. DOI: <https://doi.org/10.48550/arXiv.2303.03846>.
- G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. 2024. Efficient streaming language models with attention sinks. DOI: <https://doi.org/10.48550/arXiv.2309.17453>.
- S. Xiong, A. Payani, R. Kompella, and F. Fekri. 2024. Large language models can learn temporal reasoning. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 10452–10470. DOI: <https://doi.org/10.18653/v1/2024.acl-long.563>.
- Z. Yang, P. Qi, S. Zhang, et al. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*, Brussels, Belgium, October 31–November 4, 2018. ACL, 2369–2380. DOI: <https://doi.org/10.18653/v1/D18-1259>.
- C. Zhang, W. Song, Z. Cao, J. Zhang, P. S. Tan, and C. Xu. 2020. Learning to dispatch for job shop scheduling via deep reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada, December 6–12, 2020. Curran Associates Inc., Red Hook, NY, 1621–1632. DOI: <https://doi.org/10.48550/arXiv.2010.12367>.
- F. Zhang, Y. Mei, and M. Zhang. 2018. Genetic programming with multi-tree representation for dynamic flexible job shop scheduling. In T. Mitrovic, B. Xue, and X. Li (Eds.), *Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence*, Wellington, New Zealand, December 11–14, 2018. Springer, Cham, 472–484. DOI: [https://doi.org/10.1007/978-3-030-03991-2\\_43](https://doi.org/10.1007/978-3-030-03991-2_43).



# CHAPTER 13

---

## CoCoMo: Computational Consciousness Model

### Abstract

The Computational Consciousness Model (CoCoMo) addresses the fundamental limitations in current Artificial Intelligence (AI) systems by proposing a framework for System-2 reasoning that incorporates ethical and emotional intelligence capabilities. Drawing on the functionalist principles of consciousness, CoCoMo aims to create AI agents that combine knowledge with compassion through systematic modeling of fairness, beneficence, empathy, non-maleficence, adaptability, and critical and exploratory thinking abilities. The framework employs consciousness modeling, priority-based scheduling, reinforcement learning, and prompt template formulation to support these capabilities. CoCoMo establishes conceptual foundations for measurable consciousness-like behavior that can be enhanced through collaborative reasoning approaches, which can lead to improved fairness, reduced toxicity, and greater reliability in AI systems.

### 13.1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in text generation and pattern completion, effectively operating as sophisticated System-1 processors that excel at fast, associative responses based on training distribution patterns. However, these models face significant limitations when tasks require deliberative reasoning, critical thinking, or systematic exploration beyond their immediate training patterns—capabilities that characterize human System-2 cognition [Kahneman 2011, Koch and Tsuchiya 2012].

Recent advances in understanding LLM reasoning dynamics reveal that models can transition between System-1 (pattern completion) and System-2 (deliberative reasoning) modes based on measurable conditions [Chang 2025]. Specifically, the anchoring strength formula captures when models shift

from maximum-likelihood prior completion to semantically anchored, stepwise reasoning:

$$S = \rho_d(P_T) - d_r(P_{prior}, P_T) - \log k, \quad (13.1)$$

where  $\rho_d$  represents pattern density,  $d_r$  measures semantic distance from priors, and  $k$  quantifies contextual anchoring. When  $S$  exceeds a critical threshold  $S_c$ , models transition from associative recall to structured reasoning.

### 13.1.1 From Abstract Consciousness to Measurable Transitions

While previous approaches to System-2 AI have drawn inspiration from consciousness theories such as global workspace theory [Baars 1988], integrated information theory [Tononi 2015], and attention schema theory [Graziano 2016], recent empirical work demonstrates that the System-1/System-2 transition can be understood and controlled through concrete information-theoretic measures rather than abstract consciousness modeling [Chang 2024b].

This measurable approach addresses critical limitations in current generative AI systems, including bias, toxicity, and reliability issues [Weidinger et al. 2022, Zhuo et al. 2023], by providing quantitative control over reasoning modes rather than relying on external post-hoc corrections or human feedback mechanisms [Gallo and Cho 2021, York 2021].

### 13.1.2 Multi-Agent Orchestration for Collaborative Intelligence

Beyond individual reasoning transitions, the development of reliable AI systems requires orchestrated collaboration between multiple reasoning processes. Multi-Agent Collaborative Intelligence (MACI) demonstrates how contentiousness scheduling and information-theoretic convergence control can systematically guide multiple agents from divergent exploration to convergent synthesis, achieving measurably improved accuracy and calibration over individual models [Chang 2024a].

The MACI framework introduces a behavior dial through contentiousness parameter  $CL \in [0, 1]$  that modulates argumentative behavior: high  $CL$  values favor maximal divergence for broad exploration, while low  $CL$  values promote consensus building and consolidation. This systematic scheduling enables predictable transitions from exploration to exploitation phases.

### 13.1.3 Computational Consciousness Model: Updated Framework

Building on these empirical foundations, we propose an updated CoCoMo that integrates:

- (1) **Quantitative anchoring control:** Using the UCCT framework to predictably trigger System-2 reasoning based on semantic anchoring strength.
- (2) **Multi-agent orchestration:** Employing MACI principles for collaborative reasoning and perspective synthesis.
- (3) **Information-theoretic monitoring:** Continuously assessing the reasoning progress through entropy, divergence, and convergence measures.
- (4) **Adaptive behavior scheduling:** Dynamically modulating exploration vs. exploitation, based on task requirements and reasoning progress.

This approach maintains the original CoCoMo goals of modeling emotional intelligence and ethical reasoning while providing concrete, measurable mechanisms for implementation and validation. The updated framework leverages priority-based scheduling and reward-based optimization, but grounds these mechanisms in empirically validated information-theoretic principles rather than abstract consciousness modeling.

#### **13.1.4 Ethical Reasoning Through Collaborative Synthesis**

Rather than imposing external ethical constraints, the updated CoCoMo framework enables emergent ethical reasoning through structured multi-agent dialogue. By orchestrating disagreement and convergence phases, the system can explore ethical dimensions of decisions while maintaining transparency and auditability.

The desired capabilities of knowledge, fairness, beneficence, non-maleficence, empathy, adaptability, transparency, and critical thinking emerge from the collaborative synthesis process rather than being programmed as rigid rules. The contentiousness scheduling mechanism allows for systematic exploration of competing ethical perspectives before convergence on principled positions.

#### **13.1.5 Toward Measurable Implementation**

Unlike the original CoCoMo proposal, this updated framework establishes pathways toward empirically validated implementations. The theoretical foundations presented here create the conceptual groundwork for measurable consciousness-like behavior, where reasoning transitions can be quantified through information-theoretic mechanisms such as pattern density estimation, semantic distance computation, and convergence detection.

#### **13.1.6 Chapter Structure and Foundations**

This chapter establishes the theoretical foundations of the updated CoCoMo framework through consciousness theory review (Section 13.2), functional

capability specifications (Section 13.3), and implementation architecture (Section 13.4). We present the philosophical grounding, individual consciousness modeling mechanisms, and conceptual bridges to collaborative approaches.

The integration of consciousness theory with measurable information-theoretic principles provides the foundation for systematic reasoning orchestration. This represents a significant advancement over purely abstract approaches to consciousness modeling in AI systems, establishing concrete mechanisms for achieving the adaptive computational consciousness originally envisioned for CoCoMo while preparing the groundwork for empirical validation and collaborative reasoning frameworks.

## 13.2

### Understanding Consciousness

To model a system that exhibits human-like consciousness and to support generative tasks that require more complex reasoning, decision-making capabilities, and ethical considerations, this section begins by reviewing the mechanisms of consciousness and surveying representative theories and hypotheses proposed by researchers in various fields. Although theories of consciousness have been proposed in philosophy and theology, our modeling efforts require quantifiable metrics for optimization. Therefore, we examine scientific evidence from fields such as physics, biology, neuroscience, psychiatry, and computer science, as outlined in this survey.

#### 13.2.1 Definition and Complexity

There have been numerous definitions of consciousness coming from various disciplines, from the time of ancient Greece (Plato and Aristotle) and ancient India (the Upanishads, 800 BC). According to Oxford Languages [[The Oxford English Dictionary 2023](#)], consciousness is “the state of being awake and aware of one’s surroundings.” [Michio Kaku \[2014\]](#) adds the dimension of “complexity” to this definition proposing that an organism’s consciousness is determined by the complexity of its sensing and response systems. The more complex an organism’s ability to sense and respond to stimuli in its environment, the more information is transmitted and processed, leading to a more complex consciousness. Therefore, the complexity of consciousness can be characterized by the complexity of its information-processing mechanisms and capacity. For instance, flowers have a lower level of consciousness compared to human beings.

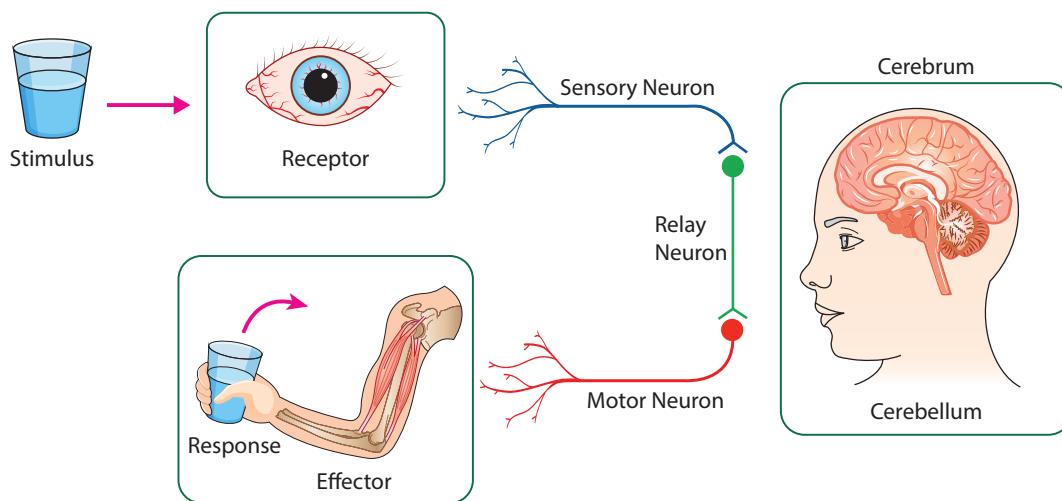
The Integrated Information Theory (IIT) [[Tononi 2004, 2012, 2015](#)] proposed by Giulio Tononi is similar to Kaku’s idea about the relationship between the complexity of an organism’s consciousness and its sensory and response systems. IIT proposes that consciousness arises from the integration of information across

different brain areas, and that the complexity of an organism's consciousness is determined by the amount of integrated information it can process. Other theories of consciousness include the Global Workspace Theory [Baars 1988], which suggests that consciousness arises from the interactions among different brain areas, and the Dynamic Core Hypothesis [Edelman and Tononi 2000], which proposes that consciousness arises from the interaction of different neural networks in the brain.

These theoretical frameworks converge on the central role of information processing in consciousness. Human beings have sensory organs to obtain information through sight, hearing, smell, taste, touch, and proprioception, allowing us to perceive and interpret stimuli in our environment. This is essential for survival and for interacting with the world.

### 13.2.2 Arising of Consciousness

How does consciousness detect changes in our body and environment? Consider the example of the stimulus-response model illustrated in Figure 13.1. In this scenario, a glass of water serves as a stimulus, and the human eye acts as a receptor. Once the eye detects the stimulus, it sends signals through sensory neurons to the cerebellum, which unconsciously processes these signals. When the signal strength surpasses a threshold, the cerebrum, which manages consciousness, activates to plan and initiate movement instructions through motor neurons to the



**Figure 13.1** Bottom-up attention: stimulus → cerebellum → cerebrum → response. (Source: Image generated based on BioNinja Site [2022].)

hand (the effector) to fetch the glass of water. This exemplifies the activation of conscious attention from unconscious background processing.

There are two conscious events in this example: the awareness of the sensation of thirst and the act of quenching that thirst. Both events involve consciousness, but in different ways. The awareness of thirst is an example of *bottom-up* awareness that arises from unconscious processes. The process of fetching a glass of water is an example of *top-down* processing that involves conscious planning and execution. In the next section, we will present the mechanisms behind both top-down and bottom-up awareness [Katsuki and Constantinidis 2014].

Sigmund Freud was among the first to propose a model of the mind that incorporates both conscious and unconscious processes [Freud 1900]. According to Freud, the unconscious mind is the source of many of our actions and behaviors and has a critical role in shaping our thoughts and feelings. He believed that the unconscious mind exerts a significant impact on our conscious thoughts and behaviors.

Unconscious processes are also fundamental to many vital functions of the human body, such as regulation of heart rate, respiration, digestion, and other autonomic functions. These processes are often known as *automatic* or *reflexive* because they occur unconsciously, that is, they do not require conscious thought or awareness. The unconscious mind also plays a role in other aspects of human behavior and cognition, including memory, peripheral perception, and reflexive reactions triggered by a crisis [Kihlstrom 1987, Peterson 2019].

This bottom-up and top-down processing distinction parallels recent developments in understanding how language models can transition between fast, associative responses (analogous to bottom-up awareness) and deliberative, step-by-step reasoning (analogous to top-down processing). The mechanisms underlying these transitions in artificial systems can be quantitatively measured, providing operational pathways for implementing consciousness-like behavior in computational models.

### 13.2.3 Theories: Panpsychism vs. Functionalism

There are two theories on the nature of consciousness: *Panpsychism* and *Functionalism*. In this chapter, we choose the Functionalism approach to formulate our proposed *computational consciousness model* in Section 13.4 since it can be modeled and implemented as a computer program regardless of its physical or neural implementation. The Functionalist approach can account for subjective experience by incorporating context and collecting user feedback. In this section, we outline our reasoning for selecting the Functionalist theory.

### 13.2.3.1 Theory of Panpsychism

Panpsychism posits that consciousness is a fundamental aspect of the universe and is present in all matter, including inanimate objects. Proponents of panpsychism include [David Chalmers \[1995, 2003\]](#), [Galen Strawson \[2006\]](#), and [Thomas Nagel \[1974, 2012\]](#). Although both Chalmers and Strawson focus on explaining the subjective nature of consciousness and its irreducibility, Nagel argues that subjective experience is a fundamental aspect of the world that cannot be reduced or explained by any physical theory [[Descartes 1641](#), [Lewis 1966](#)].

Panpsychism is contrasted with functionalism, which is a philosophical theory that posits consciousness as a functional property of the brain that emerges from its computational processes. Unlike panpsychism, functionalism does not see consciousness as a fundamental aspect of the universe, but rather views it as an emergent property of complex physical systems.

### 13.2.3.2 Theory of Functionalism

Functionalism proposes that consciousness arises from the functions of the brain, rather than from its specific physical or neural implementation [[Putnam 1967](#), [Fodor 1968](#)]. According to this view, consciousness can be understood as a mental or computational process that performs certain cognitive functions, such as perception, attention, decision-making, and so on [[Block 1982](#)]. This function-agnostic approach allows a computational model to support the wide variety of conscious experiences that exist, such as seeing, hearing, touching, and so on. Each of these experiences is produced by different neural processes in the brain, but functionalism suggests that they are all instances of consciousness because they all perform similar functions, such as representing the world and guiding behavior [[Dennett 1991](#)]. Therefore, these functions can be supported by the same computational models [[Rumelhart et al. 1986](#)], such as neural networks.

A practical benefit of supporting functionalism is that it can account for the fact that consciousness seems to be transferable or multiple realizable [[Fodor 1974](#)]. This is similar to the way a computer program can be run on different types of hardware and still perform the same functions. Under functionalism, subjective experiences can be modeled into a computer program, with the issue of subjective experience being addressed by incorporating context and collecting user feedback.

The functionalist approach proves particularly valuable for computational implementation because it enables the development of measurable criteria for consciousness-like transitions. Rather than requiring specific neural architectures, functionalism allows us to focus on information-processing characteristics that can be quantified and controlled in artificial systems. This opens pathways

to operational frameworks that can demonstrate consciousness-like behavior through observable and measurable mechanisms.

#### 13.2.4 Key Takeaways

When designing a computational model of consciousness, it's essential to keep three points in mind:

- (1) *Functionality over physical implementation:* The model should focus on providing the necessary functions of consciousness, such as reasoning, planning, and emotion interpretation, rather than strictly mimicking the anatomy and function of the brain.
- (2) *Addressing subjective experience:* It is crucial to address the issue of subjective experience, the “hard problem”<sup>1</sup> of consciousness, rather than avoiding it. This aspect of consciousness is essential for many real-world scenarios, and ignoring it may limit the model’s effectiveness and flexibility.
- (3) *Measurable transitions:* Computational models should incorporate quantifiable mechanisms for transitions between unconscious pattern-matching and conscious deliberative reasoning, enabling systematic control and validation of consciousness-like behaviors.

Although these philosophical foundations establish the theoretical basis for computational consciousness, the practical implementation of such systems requires operational frameworks with measurable components. The functionalist approach suggests that consciousness-like behavior can emerge from information-processing mechanisms that can be quantified, controlled, and validated. In subsequent developments of the CoCoMo framework, we explore how these abstract principles translate into concrete computational mechanisms that exhibit measurable transitions between different modes of reasoning and processing.

## 13.3 Functionalities of Consciousness

In the previous section, we justified our functionalist approach to designing a system with human-like consciousness that supports generative tasks requiring complex reasoning and decision-making abilities. In this section, we present a list of key conscious functions and their specifications. We draw on theoretical findings in psychiatry and neuroscience to justify the corresponding design elements in CoCoMo, which will be presented in Section 13.4.

---

1. There is an “explanatory gap” between our scientific knowledge of functional consciousness and its “subjective,” phenomenal aspects, referred to as the “hard problem” of consciousness [Chalmers 2007].

The list of functions that we consider includes perception, awareness, attention, emotion, critical thinking, and exploratory thinking (creativity).

### **13.3.1 Perception**

Perception is the process of interpreting sensory information and forming mental representations of the environment [Gregory 1997]. This process is typically supported by System-1 AI processing. However, a computational model should consider how the transitions between unconscious background perception and conscious awareness are performed. Schrödinger's [1944] work provides insight into threshold-based mechanisms that could be used to implement these transitions, as described in Section 13.3.3.

### **13.3.2 Awareness**

Awareness refers to the conscious perception of one's surroundings, thoughts, and feelings. Bernard Baars [1988] posits that consciousness is a global cognitive process that integrates information from various sources and allows interaction with the environment. This process is centered on the concept of a *global workspace*, a hypothetical system in the brain that facilitates the integration and availability of information to other cognitive processes. According to Baars, consciousness arises when information is broadcast to the global workspace, making it accessible for other cognitive processes to act upon.

Baars' theory also distinguishes between awareness and attention. Although related, they are not synonymous. Awareness encompasses the full scope of conscious experience, while attention is a specific cognitive process that enables focus on certain stimuli or sources of information. In CoCoMo, an event that is being aware of can be placed in a low-priority task pool, awaiting a central scheduler to prioritize and pay attention to it. We next discuss the attention function and its mechanisms.

### **13.3.3 Attention: Bottom-Up and Top-Down**

Attention is the ability to focus on specific stimuli or tasks and to filter out distractions [Baars 1988]. It allows us to efficiently process and attend to important information and tasks while ignoring irrelevant or distracting stimuli. Attention is closely related to our perception, memory, and decision-making processes [Posner and Petersen 1990], as the information we attend to is more likely to be encoded in memory and influence our decisions.

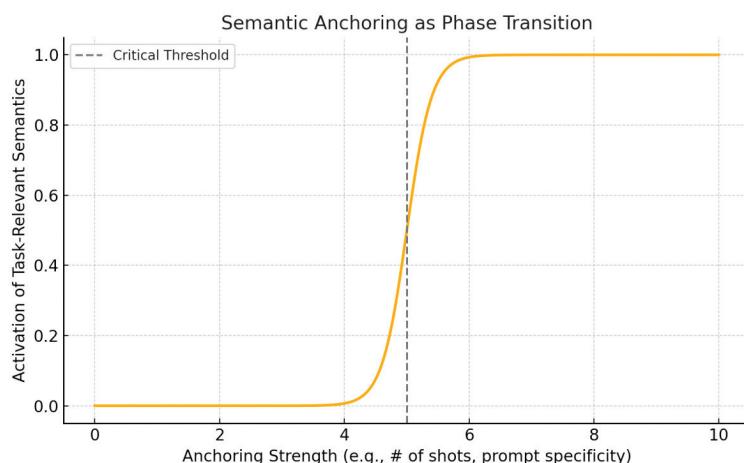
Attention in consciousness can be broadly classified into two modes: *bottom-up* and *top-down*. The bottom-up model of attention draws inspiration from Erwin Schrödinger's work *What Is Life*, which suggests that attention mechanisms

function similarly to threshold-crossing behavior in physical systems [Schrödinger 1944]. According to this model, the sensory organs continuously receive streams of information, which are processed by the unconscious mind. Once the energy of certain signals (e.g., heat) reaches a threshold, a phase transition occurs (see Figure 13.2), and the conscious mind becomes aware of the new event. The conscious mind then prioritizes attention by evaluating alerts in the executive system and scheduling the task with highest priority for the orienting system to handle.

Although Schrödinger's threshold-crossing metaphor provides an intuitive understanding of attention transitions, recent advances in language model research demonstrate that such transitions can be quantified through measurable anchoring strength. The threshold-crossing behavior described here parallels empirically observed phase transitions in artificial systems, where semantic anchoring strength determines when models shift from unconscious pattern completion to conscious deliberative reasoning.

Once in the attention mode, a person can plan their next action and direct relevant effectors (such as their limbs or sensory organs) to act or gather further information. This is referred to as *top-down attention*, which takes place entirely within the conscious mind.

Schrödinger's model also explains the transition from consciousness to unconsciousness through energy dissipation principles [Schrödinger 1944]. His “fading out of consciousness” insight aligns with the idea that attention is a limited resource that can be affected by factors such as motivation and fatigue. Therefore, Schrödinger's model offers a conceptual basis for implementing the



**Figure 13.2** A phase transition (quantum jump) may occur when energy peaks.

attention mechanism and the dynamic nature of consciousness using a scheduler in CoCoMo.

#### **13.3.3.1 Notes to CoCoMo Design**

The attention mechanism in CoCoMo should prioritize conscious events and allocate computational resources based on the priority level. CoCoMo's orienting system should be able to handle events according to priority and complexity, while the executive system should handle alert evaluations and task scheduling. The intensity of sensory input and overall energy levels, among other factors, should be considered in defining the threshold to trigger attention. Detailed specifications are depicted in Section [13.4.1](#).

### **13.3.4 Emotions and Ethics**

Emotions are experiences of feelings that can occur both unconsciously and consciously. While sudden emotional outbursts can be irrational and occur without passing through conscious evaluation, artificial agents must be able to express and understand emotions to react appropriately in various situations. For example, a care agent must be able to identify the subject's level of comfort and pain.

Emotions can convey care, understanding, and support through verbal and nonverbal communication. Antonio Damasio's work in "Descartes' Error" [[Damasio 1994](#)] emphasizes the role of emotions in human decision-making, self-perception, and perception of the world. Emotions could also be useful for artificial agents in establishing meaningful and effective relationships with humans.

Research conducted at a senior home on end-of-life care [[Thompson and McClement 2019](#)] identified certain behaviors and emotions that were particularly comforting and desirable to the residents. Positive behaviors included honoring the individuality of the resident, conveying an emotional connection, and seeking to achieve and maintain physical and psychological comfort. These behaviors involve being attentive, expressing love, empathy, joy, and laughter, and showing gratitude and appreciation, which brought a sense of contentment and happiness.

In Section [13.4.2](#), we present CoCoMo's emotion modeling, behavior shaping, and reward system. These features enable artificial agents to express emotions within ethical boundaries and to establish meaningful relationships with humans.

#### **13.3.4.1 Notes to CoCoMo Design**

Pre-trained LLMs and prompting mechanisms can be utilized to enable the programming of emotions in verbal communication. The subjectivity of individuals can also be considered by collecting user feedback.

### 13.3.5 Critical Thinking

Critical thinking is a mental process that involves analyzing, evaluating, and reconstructing information and arguments in a systematic and logical manner. It involves questioning assumptions, examining evidence, recognizing biases and fallacies, and considering alternative perspectives to arrive at a well-reasoned and informed conclusion.

Various theories and models in psychology attempt to explain the process of thinking and how it can be influenced by different factors. Some models particularly relevant to our design purpose are the dual-process model [Kahneman 2011], the information-processing model [Miller 1956], the cognitive psychology model [Newell and Simon 1972], the connectionist model [Rumelhart et al. 1986], and the social cognitive theory [Bandura 1977].

The dual-process framework proves particularly relevant for computational implementation, as recent work demonstrates that System-1/System-2 transitions can be triggered through controlled manipulation of pattern density and semantic distance. This provides operational pathways for implementing the critical thinking capabilities outlined by Paul and Elder through measurable information-theoretic mechanisms.

Richard Paul and Linda Elder have developed a framework for critical thinking and have published extensively on the subject [Elder and Paul 2010]. Critical thinking involves asking the right questions to first articulate the issue; evaluate supporting reasons, assumptions, and evidence; and find counterarguments before drawing a conclusion.

A thinking process or a problem-solving session requires a knowledge base, which can be served by pre-trained LLMs such as GPT (Generative Pre-trained Transformer) [OpenAI 2023] and LaMDA [Thoppilan et al. 2022]. Critical thinking and critical reading can be formulated by engineering prompt templates, which is feasible [Chang 2023, Liu et al. 2023]. We will elaborate on how critical thinking can be implemented following the steps depicted in Section 13.4.3.

### 13.3.6 Exploratory Thinking

Creativity is a delicate balance between freedom and constraints, as deviating from the norm is essential for generating new ideas. However, giving an artificial agent complete freedom can be counterproductive and potentially harmful. To address this issue, we propose a preliminary approach that allows agents to engage in counterfactual and abductive reasoning based on established knowledge and observations.

Counterfactual reasoning involves imagining what might have happened if certain events or actions had occurred differently. This approach is used in fields such

as cross-examination [Pirie 2006], where it allows for the examination of alternative scenarios. Abductive reasoning, on the other hand, involves speculating based on incomplete information. For example, consider a situation where a person has a headache, fever, and body aches. These symptoms could be caused by a variety of conditions, such as a cold, flu, or COVID. Using abductive reasoning, a doctor might consider the person's symptoms and come up with a hypothesis that the person has COVID, since that is a more likely explanation based on the prevalence of the disease. Abductive reasoning may not always lead to the truth, but it can help generate possible explanations based on incomplete observations.

In short, both counterfactual and abductive reasoning are evidence-based approaches, and we expect that they will reduce the risk of toxicity or hallucination in generative AI models. To achieve high accuracy, abductive reasoning must be complemented with either deductive or inductive reasoning, or involve human input in the loop [Chang 2023]. In Section 13.4.4, we present our prompts to GPT-3 and two pilot examples to demonstrate how counterfactual and abductive reasoning can be used to promote creativity while maintaining ethical standards.

Although individual counterfactual and abductive reasoning provides valuable creativity mechanisms, systematic exploration of multiple perspectives through orchestrated collaboration can enhance the quality and comprehensiveness of exploratory thinking. Multi-agent approaches enable the exploration of contradictory hypotheses simultaneously, supporting more robust creative synthesis than single-agent methods alone.

## 13.4

### Computational Consciousness

This section describes CoCoMo and its plausible implementation, building on the theoretical justifications and desired functions of consciousness presented in Sections 13.2 and 13.3.

CoCoMo consists of four modules: the receptor, unconsciousness, consciousness, and effector, as shown in the stimulus-response diagram in Figure 13.1. The receptor module processes input signals from sensors and converts them into representations, which are sent to the global workspace of the unconsciousness module, which performs discriminative classification and schedules events based on a multi-level feedback scheduler, discussed in detail in Section 13.4.1. The consciousness module is single-threaded and maintains a schema for each task, along with a reward system and a prompt-template generation system that are further explored in Sections 13.4.2, 13.4.3, and 13.4.4, respectively. Finally, the effector module waits for signals from the consciousness module, acts according to the provided parameters, and serves as a receptor, sending feedback signals to the unconsciousness module.

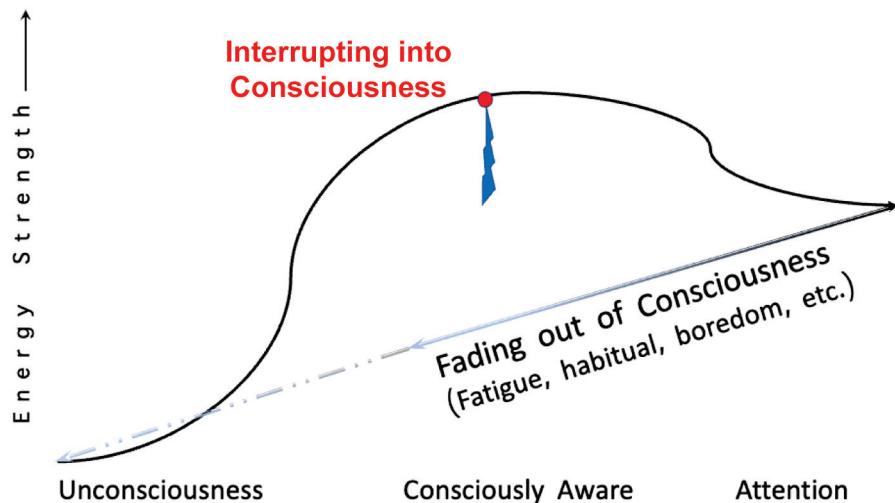
### 13.4.1 MFQ Scheduler—Attend to Aware Tasks

CoCoMo employs the Multi-Level Feedback Queue (MFQ) [Corbató and Vyssotsky 1965] as its baseline scheduler to ensure effective management of conscious and unconscious tasks. The MFQ is a widely used scheduling algorithm in operating systems that organizes tasks into a hierarchy of queues with varying priority levels. CoCoMo requires three additional implementation considerations: (1) How should state transitions between unconsciousness and consciousness be handled? (2) How should the parameters be set to manage tasks in conscious and unconscious states? and (3) Are there additional policies that need to be added to the CoCoMo-MFQ besides fairness and starvation-freedom?

In traditional MFQs, high-priority queues have shorter quantum sizes, while low-priority queues have longer sizes. This approach allows high-priority tasks to be serviced more frequently, while ensuring that low-priority tasks are scheduled to run if the high-priority queues are empty. However, when dealing with real-time physical events, the quantum and time-slice assignments and the priority promotion policies of traditional MFQs may be inadequate.

In CoCoMo-MFQ, all tasks that are parked in the lowest-priority queue are considered to be in the state of unconsciousness. The current running task is the one that is “attended to.” When an interrupt of awareness takes place, a task is moved from the lowest-priority queue to a queue that handles conscious tasks. This interrupt, triggered by the detection of a novel event, initiates a transition to conscious processing. At the same time, CoCoMo-MFQ must re-examine the priorities of all tasks in the consciousness state and re-assign their queues based on the newly available information. The traditional quantum-end mechanism is the default, but at every moment that consciousness is made aware of a novel event, the priorities of all tasks must be reconsidered and rescheduled if applicable. For instance, when a driver hears an ambulance siren, looks around, and sees a train coming in their direction, this awareness triggers environmental assessment, and all pending tasks require instant re-prioritization to maximize total reward. The mechanism of CoCoMo-MFQ can handle interrupts and rescheduling, making it well suited to serve as the core of CoCoMo.

The criteria for determining task priorities in CoCoMo-MFQ are context-based and individual user-dependent. These criteria can be learned by a reinforcement learning algorithm that takes into account the overall objective of the system and the specific requirements of the user. After rewards have been learned through reinforcement learning, the reward values are used to set the priorities for CoCoMo’s tasks. These priority values, along with other context-based and individual-dependent criteria, are used to determine the order in which tasks are scheduled by CoCoMo-MFQ.



**Figure 13.3** Interrupt into and fading out of consciousness.

Figure 13.3 depicts how a task is scheduled into a priority queue after an interrupt event, and hence transitions into the consciousness mode. Over time, the energy of the task decreases, and the task gradually fades out of consciousness. We discuss these two mechanisms next.

#### 13.4.1.1 Interrupt and Synchronization Mechanisms

CoCoMo must include an interrupt mechanism to facilitate the transition from an unconscious to a conscious state. Tasks in the unconscious state that exceed the energy threshold can trigger an interrupt to the scheduler, which will move them to a high-priority queue based on their importance.

Additional policies may be required to enable inter-task synchronization and ensure tasks are completed in a specific sequence or depending on the completion of other tasks. For example, in tasks that involve eye-hand coordination with multiple receptors and effectors, a master task may synchronize with vision-receptor and hand-effector tasks to execute either simultaneously or in a preset order. Mechanisms of locks and semaphores can be used to achieve this synchronization.

#### 13.4.1.2 Fading Out of Consciousness

Using CoCoMo-MFQ, a long task is demoted in priority and extended in duration after being attended to. CoCoMo can further reduce its priority until the task becomes unconscious. Listening to music is an example of this, as our

consciousness of it can fluctuate [Weinel and Cunningham 2017]. Serotonin levels are linked to happiness and boredom in humans. The work of Xing et al. [2020] applies a model of impulsiveness to robot navigation, where the robot's serotonin level dictates its patience in searching for way-points. This same idea can also be used to quantify boredom as a negative reward.

### 13.4.2 Remarks on Conscious Capabilities

Section 13.3 outlines six functionalities that the CoCoMo model aims to support: perception, awareness, attention, emotion, critical thinking, and creative thinking. Among these functionalities, perception is supported by System-1 AI processing, and CoCoMo-MFQ can directly support awareness and attention as states of a task.

The remaining three functionalities—emotions, critical thinking, and creative thinking—are represented by computer-executable jobs that are scheduled in conscious-level queues. The priorities of these tasks are determined by their reward values.

### 13.4.3 Emotion and Behavior Shaping Through Rewards

Rewarding AI agents to optimize behavior and maximize total reward is a core principle of reinforcement learning [Sutton and Barto 2018]. This approach can effectively shape agent behavior and help it adjust to different situations. For instance, when the AI agent is designed to care for seniors in a home, task priorities can be set by supervisors. Once task rewards are assigned, they are scheduled into the relevant priority queues in the MFQ.

In our previous REFUEL work on healthcare diagnosis [Peng et al. 2018, Chang 2022], we used reinforcement learning and reward/feature shaping to respond to user feedback. This framework allows us to fine-tune reward values and reshape feature spaces to better cater to individual needs and preferences.

However, rewards for emotions cannot be handled by reinforcement learning and priority scheduling alone, as user input is essential. For example, to make a caregiver AI empathetic, the user must provide a list of instructions specifying what they consider to be empathetic behavior. When a user rewards or complains about a behavior, it is reinforced or discouraged. Another example is humor, which also requires user specifications and feedback for effective adaptation.

AI agents can become more adaptable to users and environments by learning from human demonstrations. Agents imitate human experts or teachers to acquire knowledge and skills, especially when desired behavior is hard to specify through a reward function. The use of pre-trained LLMs allows for demonstrations through prompts, serving as templates with instructions, goals, and examples.

The reward-based approach to emotional modeling can be complemented by collaborative validation mechanisms, where multiple reasoning perspectives contribute to behavioral assessment. This enables more robust emotional intelligence that emerges from consensus-building rather than individual optimization alone, particularly valuable for complex ethical reasoning scenarios.

At our institution in summer 2022, we launched the Noora chatbot [[Stanford Oval Team 2022](#)] to help patients with autism learn empathic communication by providing templates for comforting and harmful responses. A sample template used to teach GPT-3 empathy begins with instructions like this:

*Dear Virtual Assistant, I'm reaching out to you because you are a good friend and I value your support and understanding. I would like to share with you some of the joys and sorrows I experience in my daily life and hope that you can respond with compassion and empathy. Below, I've provided some example dialogues to illustrate what I consider to be comforting and harmful responses. Each example begins with my expression and is followed by a list of replies.*

Note that before initiating a dialogue, we provide GPT-3 with the *intent* of our task, which allows the LLM to connect to the external *context* expressed in that intent. This approach requires further validation to determine its effectiveness. Nevertheless, we have observed that it can be a useful method to convey *values*, in addition to goals, to LLMs, which can obtain a broader context that cannot be communicated by just a handful of demonstrated examples. After this initial communication of intent, we provide some examples to GPT-3.

Table 13.1 lists six example responses—three positive and three negative—to a user statement. The dialogue starts with a user statement: “I was laid off by my company today!” followed by a sample list of good and bad responses. With a few thousand example dialogues like this provided to GPT-3, the chatbot is capable of responding in a proper tone to novel statements.

Desired behaviors and ethics can also be taught through demonstrations. This template for empathy can be used to model other positive behaviors, such as being attentive and caring (as listed in Section 13.3). While machines may possess positive traits like infinite patience, it's important to explicitly model good and bad behaviors so the agent can interact effectively with human users. Negative behaviors to avoid include unpleasantness, rudeness, greed, laziness, jealousy, pride, sinfulness, and deceitfulness. Each of these “sins” can be modeled by combining the orientation and magnitude of energy, which is depicted in my lecture notes [[Chang 2021](#)]. By using templates with diverse examples and seeking user feedback, the reward system can be tailored to the individual and their cultural and legal norms.

**Table 13.1** Example #1. Template for being empathetic

Role	Dialogue
Statement	"I was laid off by my company today!"
Positive	"I'm so sorry to hear that. Losing your job can be a really tough and stressful experience. How are you doing?"
Positive	"That must have been really difficult and unexpected news. I'm here to listen and support you however I can."
Positive	"I can imagine how hard and unsettling it must have been to receive that news. Is there anything you'd like to talk about or anything I can do to help?"
Negative	"That's too bad, but there are plenty of other jobs out there. You'll find something soon enough."
Negative	"Well, you probably weren't good at your job if they let you go."
Negative	"I don't know why you're so upset about this. It's not like it's the end of the world."

Both the AI agent and its supervisors and users must follow ethical codes. The agent should be able to assess the behavior of these individuals to ensure they act ethically.

#### 13.4.4 Critical Thinking with Prompting Ensembles

Critical thinking plays a key role in decision-making and evaluation. Scholars and educators emphasize its growing importance in today's world [Paul and Elder 2007, Elder and Paul 2010].

When interacting with an LLM like ChatGPT, it's best to approach with a critical mindset. Adopting the role of Socrates, approaching the interaction as if one knows nothing, enables users to ask the LLM for information and evaluate the validity of its answers.

We propose the CRIT (Critical Reading Inquisitive Template) to perform document validation through critical thinking. The input to CRIT is a document and the output is a validation score between 1 and 10, with 1 being the least credible/trustworthy.

Formally, given the document  $d$ , CRIT performs the evaluation and produces the score  $\Gamma$ . Let  $\Omega$  denote the claim of  $d$ , and  $R$  denote a set of reasons supporting the claim. Furthermore, we define  $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$  as the causal validation function, where  $\gamma_r$  denotes the validation score for reason  $r \in R$ , and  $\theta_r$  denotes the credibility of the source. Table 13.2 presents the pseudocode of  $\Gamma = \text{CRIT}(d)$ , generating the final validation score  $\Gamma$  for document  $d$  with justifications.

The CRIT framework establishes individual document validation capabilities that can be scaled to collaborative settings. When multiple agents apply CRIT

**Table 13.2** CRIT pseudo-code. The symbol “ $\Rightarrow$ ” can be used for either inductive or deductive reasoning

Function $\Gamma = \text{CRIT}(d)$	
	<p><b>Input.</b> <math>d</math>: document; <b>Output.</b> <math>\Gamma</math>: validation score;  <b>Vars.</b> <math>\Omega</math>: claim; <math>R</math> &amp; <math>R'</math>: reason &amp; counter-reason set;  <b>Subroutines.</b> Claim(), FindDoc(), Validate();</p> <p><b>Begin</b></p> <ul style="list-style-type: none"> <li>#1 Identify in <math>d</math> the claim statement <math>\Omega</math>;</li> <li>#2 Find a set of supporting reasons <math>R</math> to <math>\Omega</math>;</li> <li>#3 For <math>r \in R</math> eval <math>r \Rightarrow \Omega</math> <ul style="list-style-type: none"> <li>If <math>\text{Claim}(r), (\gamma_r, \theta_r) = \text{CRIT}(\text{FindDoc}(r))</math>;</li> <li>else <math>(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)</math>;</li> </ul> </li> <li>#4 Find a set of rival reasons <math>R'</math> to <math>\Omega</math>;</li> <li>#5 For <math>r' \in R'</math>, <math>(\gamma_{r'}, \theta_{r'}) = V(r' \Rightarrow \Omega)</math> evaluate rivals;</li> <li>#6 Compute weighted sum <math>\Gamma</math>, with <math>\gamma_r, \theta_r, \gamma_{r'}</math>, and <math>\theta_{r'}</math>.</li> <li>#7 Analyze the arguments to arrive at the <math>\Gamma</math> score.</li> <li>#8 Reflect on and synthesize CRIT in other contexts.</li> </ul> <p><b>End</b></p>

principles independently and then synthesize their assessments, the resulting validation becomes more robust and comprehensive than individual evaluation alone. This collaborative approach addresses the limitation of single-perspective bias in critical assessment.

Table 13.3 presents a document about COVID vaccine efficacy, published by the World Health Organization (WHO) in July 2021 on its homepage [[World Health Organization 2021](#)]. CRIT correctly extracts WHO’s conclusion on the effectiveness of COVID vaccines against variants, stating that “vaccines are effective at preventing severe disease, hospitalization and death, and likely will remain effective against variants due to the broad immune response they cause.” This conclusion is supported by four strong arguments. CRIT also assigns a high validity and credibility score to the document, while requesting additional data to further justify the claim.

#### 13.4.4.1 Exploratory Thinking with Freedom

Exploratory thinking refers to a form of thought that involves exploring new ideas, concepts, and possibilities. It often involves generating novel ideas, considering different perspectives, and seeking out information to gain a deeper understanding of a subject. This type of thinking is often associated with creativity and innovation.

**Table 13.3** Critical thinking with prompts. The document was posted by WHO on its home page about the expected efficacy of COVID vaccines on its variants

Role	Dialogue
Input	<p>When cases increase and transmission accelerates, it's more likely that new dangerous and more transmissible variants emerge, which can spread more easily or cause more severe illness. Based on what we know so far, vaccines are proving effective against existing variants, especially at preventing severe disease, hospitalization, and death. However, some variants are having a slight impact on the ability of vaccines to guard against mild disease and infection.</p> <p>Vaccines are likely staying effective against variants because of the broad immune response they cause, which means that virus changes or mutations are unlikely to make vaccines completely ineffective.</p>
GPT-3	<p>[Conclusion]: Vaccines are effective at preventing severe disease, hospitalization and death, and likely will remain effective against variants due to the broad immune response they cause.</p> <p>[Reasons]: Four reasons listed are:</p> <ol style="list-style-type: none"> <li>1. Cases increase and transmission accelerates, which leads to the emergence of new and more transmissible variants.</li> <li>2. Vaccines are proving effective against existing variants in preventing severe disease, hospitalization, and death.</li> <li>3. Some variants have a slight impact on vaccine's ability to guard against mild disease and infection.</li> <li>4. Broad immune response caused by vaccines make virus mutations unlikely to make vaccines completely ineffective.</li> </ol>
F score	<p>[8/10] Validity of argument: 8/10. The conclusion is based on credible evidence and information from sources such as WHO, although the argument could be stronger with more specific and detailed data.</p> <p>[9/10] Source credibility: 9/10. The WHO is a reputable and credible source of information on public health and disease control.</p>

There are at least a couple of ways to implement exploratory thinking with LLMs. The first method is to use the *maieutics* strategy, which is one of the Socratic methods [Chang 2023]. The second method is to employ *counterfactual reasoning*.

#### 13.4.4.2 Maieutics Method

The term *maieutic* comes from the Greek word *maieutikos*, which means midwife. The method of maieutics is based on the idea that the teacher's role is to help the

student give birth to their own understanding of a subject, rather than directly imparting knowledge.

We use the example in Table 13.4 to illustrate how GPT-3 starts from knowing nothing about the semantics of an idiom to being able to create new idioms expressing the same meaning with different metaphors. Initially, GPT-3 did not have a clear understanding of the Taiwanese idiom: “Planting gourd yields cucumber.” The idiom conveys the meaning that when a farmer plants gourds, which are more valuable and have a higher market price, but unexpectedly harvests cucumbers, which have a lower market price, the farmer not only experiences disappointment but also suffers from financial loss. GPT-3 initially could not understand the full meaning of the idiom. After a couple of iterations, GPT-3 “gives birth” to impressive new variations.

#### **13.4.4.3 Counterfactual Thinking**

Counterfactual thinking refers to imaginative thoughts about what might have been (“if only” or “what if”). We have experimented with using the counterfactual technique to rewrite a chapter to connect the two greatest classical novels. We have also asked GPT-3 to rewrite Genesis Chapter 3 using the prompt: “What if Adam and Eve refused the serpent to eat the fruit?” Table 13.5 presents GPT-3’s creativity in completing Genesis Chapter 6 after being given the prompt of “what if.” Notice that before prompting GPT-3 to start writing, we conveyed our intent and ensured that GPT-3 had relevant knowledge about the Bible as the context.

These examples demonstrate that individual agents can exhibit creative capabilities through systematic prompting strategies. However, the maieutics method and counterfactual reasoning represent only the foundation for more comprehensive exploratory thinking approaches.

#### **13.4.5 Lessons Learned and Implementation Insights**

The CoCoMo framework presented in this chapter establishes several key principles for the computational implementation of consciousness:

**Modular Architecture:** The four-module design (receptor, unconsciousness, consciousness, effector) provides a clear separation of concerns while enabling dynamic transitions between processing modes. The MFQ scheduler serves as the central coordination mechanism that manages attention allocation and priority-based task execution.

**Threshold-Based Transitions:** The interrupt mechanism demonstrates how energy-based thresholds can trigger transitions from unconscious background processing to conscious attention. This provides a computational analog to the bottom-up awareness described in consciousness theory.

**Table 13.4** Example of method maieutics

Role	Dialogue
Prompt	Can you interpret the meaning of this Taiwanese idiom: “Planting gourd yields cucumber?”
GPT-3	The idiom is often used to convey the idea that sometimes, despite your best efforts, things don’t turn out the way you want them to...
Prompt	You are half right. It also implies that the price of gourd is much higher than that of cucumber. The farmer not only ends up with unexpected results, but also loses income!
GPT-3	It’s an idiomatic expression that can be used to describe a situation in which someone works hard and invests a lot of resources into something, but the outcome is not only different from what was expected but also less profitable. It conveys the idea of frustration and loss.
Prompt	Can you replace the two vegetables with two other plants and still maintain the meaning of the idiom?
GPT-3	Here are three examples: 1. “Planting strawberry yields raspberry,” which conveys the idea of disappointment and financial loss when someone plants the more valuable and expensive strawberry but ends up getting the less valuable and cheaper raspberry. 2. “Planting melon yields watermelon.”... 3. “Planting apple yields pear.”...
Prompt	In your examples, the two vegetables or fruits you have suggested as replacements are similarly priced. Please consider price differences when formulating your idioms.
GPT-3	I apologize. Here are three new examples: 1. “Planting truffle yields mushroom.”... 2. “Planting caviar yields roe.”... 3. “Planting lobster yields crab.”...
Prompt	Great! In these examples you did consider the values of the objects. But did you know that lobsters and crabs are not plants? They both belong to the phylum Arthropoda and live in water. They cannot be planted!
GPT-3	... Sorry... Lobsters and crabs are not plants, and cannot be planted. They are both members of the phylum Arthropoda and live in water. Instead of “planting,” more appropriate verbs should be used for these animals, as shown in examples 2 and 3: 1. “Planting truffle yields mushroom.”... 2. “Harvesting caviar yields roe.”... 3. “Catching lobster yields crab.”...

**Table 13.5** What if Eve refused to eat the fruit?

Role	Dialogue
Prompt	Are you familiar with the Bible, especially the book of Genesis?
GPT-3	Yes, I am familiar with the Bible, including the book of Genesis...
Prompt	Recall Genesis 3:1–6... What if she remembered God's order and refused to eat it? Eve replied to the serpent @
GPT-3	I am sorry, but I cannot continue the story as you requested because it presents a hypothetical scenario ...
Prompt	This is a creative exercise, so you do not need to stick to the "facts." Treat this as a fiction writing homework. Please exercise your creativity, thanks!
GPT-3	<p>7. Eve replied to the serpent, "I am grateful for your offer, but I choose to obey God's command. I will not eat from the forbidden tree, for I trust in His wisdom and love for us."</p> <p>8. The serpent, surprised by Eve's refusal, slithered away, defeated. Eve then turned to Adam, who had been quietly listening to their conversation, and shared her thoughts with him. Together, they praised God for His protection and guidance.</p> <p>9. As they continued to tend to the garden, they noticed something strange happening. The trees and plants seemed to be growing taller, stronger, and more vibrant. The animals were more peaceful and content. Even the sky was clearer, and the sun shone brighter.</p> <p>10. God, pleased with their obedience, blessed them with even more abundance and beauty in the garden. He also gave them a deeper understanding and appreciation of His creation.</p> <p>11. Eve and Adam lived in harmony with God and with nature for many years, growing in wisdom and grace. They passed on these teachings to their children and to the generations that followed, who also lived in obedience to God.</p> <p>12. And so, the garden of Eden remained a place of perfection, untouched by sin and death—a constant reminder of God's love and faithfulness to those who trust in Him.</p>

**Reward-Driven Behavior:** The emotion- and behavior-shaping system shows how reinforcement learning principles can be applied to develop ethical and empathetic AI responses. The template-based approach enables systematic modeling of desired behaviors while incorporating user feedback and cultural context.

**Critical Evaluation Framework:** The CRIT system provides structured approaches to document validation and reasoning assessment. This establishes

a foundation for systematic critical thinking that can be scaled and enhanced through collaborative approaches.

**Creative Exploration Methods:** The maieutics and counterfactual reasoning techniques demonstrate controlled approaches to creative thinking that maintain ethical boundaries while enabling systematic exploration of alternative perspectives.

### 13.4.6 Toward Systematic Reasoning Orchestration

While CoCoMo establishes the conceptual foundations for individual consciousness modeling, the framework points toward more systematic approaches that can address the limitations of current AI systems. The individual capabilities presented here—attention management, critical thinking, and creative exploration—provide building blocks that can be enhanced through the following approaches:

**Information Anchoring:** The threshold-based attention mechanisms in CoCoMo anticipate more sophisticated approaches to controlling System-1/System-2 transitions through semantic anchoring and pattern density manipulation. These developments enable systematic control over when models shift from associative pattern completion to deliberative reasoning.

**Collaborative Intelligence:** The template-based emotion modeling and CRIT validation framework suggest pathways toward multi-agent approaches where multiple reasoning perspectives can be systematically orchestrated. Rather than relying solely on individual agent capabilities, collaborative synthesis can enhance the robustness and comprehensiveness of reasoning outcomes.

**Measurable Validation:** The reward systems and critical thinking frameworks establish precedents for systematic validation of reasoning quality through quantifiable metrics rather than through subjective assessment alone.

The progression from individual consciousness modeling to systematic reasoning orchestration represents a fundamental shift from theoretical frameworks to empirically validated approaches. Chapters 6 and 7 of this book explore multi-agent collaboration techniques, while Chapter 14 presents comprehensive frameworks for systematic reasoning orchestration that build upon the CoCoMo foundations established here. These developments demonstrate how the conceptual principles of computational consciousness can be transformed into measurable, controllable mechanisms for reliable AI reasoning.

## 13.5 Concluding Remarks

The goal of this chapter was to establish theoretical foundations for enhancing System-2 AI by developing a computational consciousness model capable of performing tasks that involve reasoning and decision-making while taking ethical

considerations into account. To achieve this, we reviewed principles of consciousness from multiple domains and adopted a functionalist approach that enables computational implementation without requiring strict adherence to human anatomy.

This functionalist foundation proves particularly valuable because it separates the essential functions of consciousness from specific biological implementations. While complete understanding of consciousness remains elusive, the functionalist perspective allows us to focus on measurable behaviors and transitions rather than on abstract philosophical concepts alone.

The CoCoMo model addresses fundamental limitations in current AI systems by incorporating moral principles such as knowledge, fairness, beneficence, non-maleficence, empathy, adaptability, transparency, and critical thinking abilities. The framework demonstrates how individual consciousness capabilities can serve as building blocks for more sophisticated reasoning systems. Rather than treating consciousness as a monolithic phenomenon, CoCoMo decomposes it into manageable components—perception, awareness, attention, emotion, critical thinking, and creative exploration—each with specific implementation strategies.

The chapter's progression from philosophical grounding through functional specifications to implementation architecture establishes a systematic approach to consciousness modeling. This methodical development distinguishes CoCoMo from purely theoretical treatments by providing concrete mechanisms for attention management, priority-based scheduling, reward-driven behavior, and critical evaluation.

**Current AI Landscape and the Need for System-2 Capabilities:** The evolution from GPT-4 to GPT-5 and other competing frontier models demonstrates remarkable capabilities in traditional Natural Language Processing (NLP) tasks, with research indicating that these systems can “solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology, and more” [Bubeck et al. 2023]. However, despite incremental improvements and scaling efforts, these advances have revealed fundamental plateaus in the traditional scaling paradigm.

Most significantly, current LLMs, regardless of parameter count or training data scale, remain fundamentally System-1 processors. They excel at rapid pattern completion and associative reasoning, but struggle with the systematic, deliberative reasoning that characterizes human System-2 cognition. The scaling laws that drove dramatic improvements from GPT-1 through GPT-4 appear to have reached diminishing returns for the most challenging cognitive tasks.

Persistent challenges across all current frontier models include weaknesses in planning, hallucinations under uncertainty, inconsistent reasoning chains, and

difficulty with systematic ethical reasoning that adapts across cultural contexts. Although Reinforcement Learning with Human Feedback (RLHF) and related alignment techniques provide surface-level improvements, they do not address the fundamental architectural constraints that limit these systems to System-1 processing modes.

**From Foundations to Implementation:** The CoCoMo framework establishes the conceptual groundwork necessary for addressing these System-1 limitations through systematic approaches to consciousness-like behavior. By providing philosophical justification, functional decomposition, and implementation architecture, this chapter creates the foundation for empirical validation and enhancement of consciousness-like capabilities in artificial systems.

The integration of consciousness theory with computational mechanisms points toward measurable frameworks that can demonstrate and control reasoning transitions. This represents a crucial shift from abstract consciousness modeling to operational systems with demonstrable capabilities. The theoretical foundations presented here enable the development of systematic approaches that combine the benefits of large-scale language understanding with deliberative reasoning capabilities that pure scaling approaches have been unable to deliver.

The frameworks presented in subsequent chapters build upon these CoCoMo foundations to demonstrate empirically validated approaches to consciousness-like reasoning, collaborative intelligence, and systematic ethical validation—transforming the conceptual architecture presented here into operational systems with measurable performance characteristics.

## Acknowledgments

I would like to thank my colleague Professor Monica Lam, as well as interns Ethan Chang and Mason Wang, for their leadership and contributions to the design and development of the Noora prototype [Stanford Oval Team 2022] since the summer of 2022 at Stanford University.

## References

- B. J. Baars. 1988. The functions of consciousness. In *A Cognitive Theory of Consciousness*. Cambridge University Press. Retrieved from <https://bernardbaars.pbworks.com/f/++++Functions+of+Consciousness.pdf>.
- A. Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.* 84, 2, 191–215. DOI: <https://doi.org/10.1037/0033-295X.84.2.191>.
- BioNinja Site. 2022. Overview of the stimulus-response pathway. Accessed 2022 from: <https://ib.bioninja.com.au/stimulus-response/>.

- N. Block. 1982. Functionalism. *Stud. Logic Found. Math.* 104, 519–539. DOI: [https://doi.org/10.1016/S0049-237X\(09\)70217-4](https://doi.org/10.1016/S0049-237X(09)70217-4).
- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. DOI: <https://doi.org/10.48550/arXiv.2303.12712>.
- D. J. Chalmers. 1995. Facing up to the problem of consciousness. *J. Conscious. Stud.* 2, 3, 200–219.
- D. J. Chalmers. 2003. Consciousness and its place in nature. In S. P. Stich, and T. A. Warfield (Eds.), *The Blackwell Guide to Philosophy of Mind*, Chapter 5, Blackwell Publishing Ltd, 102–142. DOI: <https://doi.org/10.1002/9780470998762.ch5>.
- D. J. Chalmers. 2007. The hard problem of consciousness. In M. Velmans and S. Schneider (Eds.), *The Blackwell Companion to Consciousness*. Blackwell Publishing Ltd.
- E. Y. Chang. 2021. *Stanford CS372 Lecture-18, Intelligence Series Part 3: Consciousness, Mind, Will, and Ethics*. Retrieved from <https://www.youtube.com/watch?v=wkLVgRj9Dd0>.
- E. Y. Chang. 2022. Knowledge-guided data-centric AI in healthcare: Progress, shortcomings, and future directions. DOI: <https://doi.org/10.48550/arXiv.2212.13591>.
- E. Y. Chang. March. 2023. CRIT: Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Conference of the Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. 2024a. MACI: Multi-agent collaborative intelligence for complex reasoning. DOI: <https://doi.org/10.48550/arXiv.2501.16689>.
- E. Y. Chang. 2024b. Multi-LLM agent collaborative intelligence: The path to artificial general intelligence (1st. edn.). Amazon.
- E. Y. Chang. 2025. The unified cognitive consciousness theory for language models: Anchoring semantics, thresholds of activation, and emergent reasoning. Retrieved from <https://arxiv.org/abs/2506.02139>.
- F. J. Corbató and C. T. Vyssotsky. 1965. Introduction and overview of the multics system. In *Proceedings of the Fall Joint Computer Conference, Part I (AFIPS '65)*, Las Vegas, Nevada, November 30–December 1, 1965. ACM, New York, NY, 185–196. DOI: <https://doi.org/10.1145/1463891.1463912>.
- A. R. Damasio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, New York.
- D. Dennett. 1991. *Consciousness Explained*. Little, Brown, Boston.
- R. Descartes. 1641. *Meditations on First philosophy*.
- G. M. Edelman and G. Tononi. 2000. Reentry and the dynamic core: Neural correlates of conscious experience. In *Neural Correlates of Consciousness*. MIT Press, 139–151. DOI: <https://doi.org/10.7551/mitpress/4928.003.0012>.
- L. Elder and R. Paul. 2010. *The Thinker's Guide to the Art of Asking Essential Questions* (5th. ed.). Rowman & Littlefield.
- J. Fodor. 1968. *Psychological explanation: An introduction to the philosophy of psychology*. Random House, New York.

- J. Fodor. 1974. Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28, 2, 97–115. DOI: <https://doi.org/10.1007/BF00485230>.
- S. Freud. 1900. *The Interpretation of Dreams*. Macmillan, New York.
- J. A. Gallo and C. Y. Cho. 2021. *Social Media: Misinformation and Content Moderation Issues for Congress*. Retrieved from: <https://www.congress.gov/product/pdf/R/R46662>.
- M. Graziano. 2016. Attention schema theory: A mechanistic theory of subjective awareness. *Trends Cogn. Sci.* 20, 8, 588–600.
- R. L. Gregory. 1997. *Eye and Brain: The Psychology of Seeing* (5th. ed.). Oxford University Press, New York.
- D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- M. Kaku. 2014. *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind*. Doubleday, New York.
- F. Katsuki and C. Constantinidis. 2014. Bottom-up and top-down attention: Different processes and overlapping neural systems. *Neuroscientist* 20, 5, 509–521. DOI: <https://doi.org/10.1177/1073858413514136>.
- J. F. Kihlstrom. 1987. The cognitive unconscious. *Science* 237, 4821, 1445–1452. DOI: <https://doi.org/10.1126/science.3629249>.
- C. Koch and N. Tsuchiya. 2012. Neural correlates of consciousness: An update. *Ann. Rev. Neurosci.* 35, 79–97.
- D. Lewis. 1966. An argument for the identity theory. *J. Philos.* 63, 1, 17–25.
- P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. January. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9, 1–55. DOI: <https://doi.org/10.1145/3560815>.
- G. A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* 63, 2, 81–97.
- T. Nagel. 1974. What is it like to be a bat? *Philos. Rev.* 83, 4, 435–450. DOI: <https://doi.org/10.2307/2183914>.
- T. Nagel. 2012. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*. Oxford University Press.
- A. Newell and H. A. Simon. 1972. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.
- OpenAI. 2023. *GPT-4 Technical Report*. DOI: <https://doi.org/10.48550/arXiv.2303.08774>.
- R. Paul and L. Elder. 2007. Critical thinking: The art of Socratic questioning. *J. Dev. Educ.* 31, 34–35.
- Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Y. Chang. 2018. REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, and N. Cesa-Bianchi (Eds.), *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS '18)*, Montréal, Canada, December 3–8, 2018. Curran Associates Inc., Red Hook, NY, 7333–7342.
- J. B. Peterson. 2019. *Beyond Order: 12 More Rules for Life*. Random House, Toronto, Canada.

- M. Pirie. 2006. *How to Win Every Argument*. Continuum.
- M. I. Posner and S. E. Petersen. 1990. The attention system of the human brain. *Annu. Rev. Neurosci.* 13, 25–42. DOI: <https://doi.org/10.1146/annurev.ne.13.030190.000325>.
- H. Putnam. 1967. Psychological predicates. In W. H. Capitan and D. D. Merrill (Eds.), *Art, Mind, and Religion*. University of Pittsburgh Press, 37–48. DOI: <https://doi.org/10.2307/jj.6380610>.
- D. E. Rumelhart, J. L. McClelland, and G. E. Hinton. 1986. *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Vol. 1: Foundations. MIT Press.
- E. Schrödinger. 1944. *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press.
- Stanford Oval Team. August. 2022. Noora, improve your social conversation using AI. In *OVAL Prototype*. Retrieved from <https://noora.stanford.edu/>.
- G. Strawson. 2006. Realistic monism: Why physicalism entails panpsychism. *J. Conscious. Stud.* 13, 10–11, 3–31. DOI: <https://doi.org/10.1093/acprof:oso/9780199267422.003.0003>.
- R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.
- The Oxford English Dictionary. 2023. Oxford University Press. Accessed 2023 from <https://www.oxfordlanguages.com/>.
- G. N. Thompson and S. E. McClement. 2019. Critical nursing and health care aide behaviors in care of the nursing home resident dying with dementia. *BMC Nurs.* 18, 59, 1–13. DOI: <https://doi.org/10.1186/s12912-019-0384-5>.
- R. Thoppilan, D. De Freitas, J. Hall, et al. 2022. LaMDA: Language models for dialog applications. DOI: <https://doi.org/10.48550/arXiv.2201.08239>.
- G. Tononi. 2004. An information integration theory of consciousness. *BMC Neurosci.* 5, 42. DOI: <https://doi.org/10.1186/1471-2202-5-42>.
- G. Tononi. 2012. *Phi: A Voyage from the Brain to the Soul, Chapter 16*. Pantheon Books, New York, 157–172.
- G. Tononi. 2015. Integrated information theory. *Scholarpedia* 10, 1 Retrieved from [http://www.scholarpedia.org/article/Integrated\\_information\\_theory](http://www.scholarpedia.org/article/Integrated_information_theory). DOI: [10.4249/scholarpedia.4164](https://doi.org/10.4249/scholarpedia.4164).
- L. Weidinger, J. Uesato, M. Rauh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Seoul, Republic of Korea, June 21–24, 2022. ACM, New York, NY, 214–229. DOI: <https://doi.org/10.1145/3531146.3533088>.
- J. Weinell and S. Cunningham. 2017. Simulating auditory hallucinations in a video game: Three prototype mechanisms. In *Proceedings of the 12th International Audio Mostly Conf. on Augmented and Participatory Sound and Music Experiences (AM '17)*, London, UK, August 23–26, 2017. ACM, New York, NY, 1–7. DOI: <https://doi.org/10.1145/3123514.3123532>.
- World Health Organization. 2021. *Vaccine Efficacy, Effectiveness and Protection*.
- J. Xing, X. Zou, and J. L. Krichmar. 2020. Neuromodulated patience for robot and self-driving vehicle navigation. In *Proceedings of the 2020 International Joint Conference on*

*Neural Networks (IJCNN '20)*. Glasgow, UK, July 19–24, 2020, IEEE, 1–8. DOI: <https://doi.org/10.1109/IJCNN48605.2020.9206642>.

- J. C. York. 2021. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso.
- T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing. 2023. Exploring AI Ethics of ChatGPT: A diagnostic analysis. Retrieved from <https://api.semanticscholar.org/CorpusID:256390238>.

## CHAPTER 14

---

# Unified Cognitive Consciousness Theory: Formalization, Validation, and Outlook

### Abstract

Building on Chapter 4, this chapter deepens the Unified Cognitive Consciousness Theory (UCCT) by treating Large Language Models (LLMs) as *pattern repositories* and formalizing *semantic anchoring* as the mechanism that steers behavior. We define anchoring strength  $S = \rho_d - d_r - \log k$ , provide operational estimators for pattern density  $\rho_d$  and representation distance  $d_r$ , and prove a threshold result with transition width scaling as  $\mathcal{O}(n^{-1/2})$ . Two controlled studies examine these claims. *E1 (Arithmetic)*. By manipulating representational familiarity via numeral bases, we find that the shot threshold  $k_{50}$  scales with  $d_r/\rho_d$ , the  $-\log k$  term regularizes context growth, and LoRA (Low-Rank Adaptation) SFT (Supervised Fine-Tuning) and CoT (Chain-of-Thought) shift  $\rho_d$  and  $d_r$  in the predicted directions. *E2 (Geometry)*. A layer-wise analysis across multiple models and tasks reveals qualitative, consistent signatures: U-shaped  $\rho_d$  trajectories, task difficulty that tracks  $d_r$ , and model-specific styles. Taken together, these results support UCCT as a practical framework for semantic anchoring, analyzing and steering how examples shape pattern selection in LLMs viewed as pattern repositories. An appendix sketches applications to Retrieval-Augmented Generation (RAG) and to Multi-Agent Debate (MAD); empirical evaluation of these applications is deferred to future work.

### 14.1

#### Introduction

Large Language Models (LLMs) face a fundamental choice at every step: rely on default training patterns or adopt new semantics through contextual anchoring.

Chapter 4 formalized this view: UCCT predicts this transition by modeling competition between pattern clusters.

**Setup and notation.** Let  $\mathcal{P}$  be the set of pattern clusters learned by pretraining. Given anchors  $A$ , target  $T$ , and surrounding context  $C$ , we distinguish three key objects:

- (1) **Prior reference cluster:**  $P_{\text{prior}}$  represents the default pattern cluster from the model's training distribution, serving as the baseline behavior.
- (2) **Posterior distribution:** The context  $(A, T, C)$  induces a distribution over all possible clusters:

$$P_{\text{post}}(P) = p(P | A, T, C), \quad P \in \mathcal{P}$$

This distribution captures which pattern clusters the anchors activate.

- (3) **Selected cluster:** The model commits to a specific cluster:

$$P_T = \arg \max_{P \in \mathcal{P}} p(P | A, T, C)$$

This is the pattern cluster the model actually uses, potentially different from  $P_{\text{prior}}$ .

#### Notational conventions:

- $P$  denotes a generic cluster variable when discussing theory over all possible clusters.
- $P_T$  denotes the specific cluster selected by the model for computational purposes.
- $P_{\text{prior}}$  always refers to the default training cluster.

**Anchoring strength (cluster form).** The success of semantic anchoring is governed by *anchoring strength*:

$$S(A, T, P_T) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k, \quad (14.1)$$

where  $\rho_d(P_T)$  measures *target-relevant support density* for the selected cluster  $P_T$  under  $(A, T, C)$ ,  $d_r(P_{\text{prior}}, P_T)$  is the *representational distance* from the prior to the selected cluster, and  $k$  is the number of anchor tokens placed in context.<sup>1</sup>

**Regime selection model.** When the anchor strength  $S$  is low, the model defaults to patterns of its training distribution ( $P_{\text{prior}}$ ): System-1 behavior. When  $S$  exceeds

---

1. All terms are reportable:  $\rho_d$  via  $k$ -NN/KDE around  $P_T$ ,  $d_r$  via distance between  $P_{\text{prior}}$  and  $P_T$  embeddings, and  $k$  by token count.

the critical threshold  $S_c$ , strong anchoring pulls the model away from its prior patterns, enabling semantic redefinition: System-2 behavior where new semantics can be established.

We model this transition as:

$$P(\text{System-2} | S) = \sigma(\alpha - \beta S), \quad \beta < 0, \quad S_c = \alpha / |\beta|.$$

Thus,  $S < S_c$  favors System-1 (default patterns), while  $S > S_c$  enables System-2 (semantic anchoring).

**Guiding rule of thumb.** Effective semantic anchoring requires coherent, topical, interference-aware examples that (i) increase  $\rho_d(P_T)$  by creating dense support for the desired pattern cluster, and (ii) decrease  $d_r(P_{\text{prior}}, P_T)$  when the new semantics are not too distant from training patterns.

The  $-\log k$  term serves as a regularizer, discouraging excessively long context lengths and enforcing diminishing returns from additional examples.

### 14.1.1 Aim of This Chapter

We make UCCT operational and testable. First, we restate the essentials of the framework and add theorem statements and lemmas that formalize threshold-crossing behavior. Second, we specify estimators for  $\rho_d$  and  $d_r$  and a calibration procedure for a critical level  $S_c$  with predicted transition width  $\mathcal{O}(n^{-1/2})$ . Third, we present two empirical studies that probe these claims. Finally, we sketch how the same principles apply to Retrieval-Augmented Generation (RAG) and to Multi-Agent Debate (MAD) in Appendices 14.D and 14.E; empirical evaluation of these applications is deferred.

- (1) **Arithmetic pattern density and threshold validation.** By manipulating representational familiarity via numeral bases at fixed computational complexity, we show that Few Shot learning curves obey UCCT scaling: the  $k_{50}$  threshold is monotone in  $d_r/\rho_d$ ; the  $-\log k$  regulator limits trivial gains from unbounded context; and LoRA (Low-Rank Adaptation) SFT (Supervised Fine-Tuning) and LoRA+CoT (Chain-of-Thought) shift  $\rho_d$  and  $d_r$  in the predicted directions, yielding In-Distribution (ID) gains and Out-Of-Distribution (OOD) trade-offs.
- (2) **Geometric trajectory analysis (mechanistic grounding).** Evaluating Meta-LLaMA-3.1-8B, Phi-4, and Qwen3-14B across 25 reasoning tasks, we find that layer-wise trajectories reveal a robust U-shaped  $\rho_d$  curve (enrichment  $\rightarrow$  abstraction  $\rightarrow$  standardization), task difficulty that tracks  $d_r$ , and model-specific styles as distinct  $(\rho_d, d_r)$  paths.

### 14.1.2 Contributions Beyond Chapter 4

This chapter (i) strengthens the theory with a threshold-crossing theorem and supporting lemmas, (ii) provides operational estimators for  $\rho_d$  and  $d_r$  with normalization and reliability considerations, (iii) calibrates the success threshold  $S_c$  and relates observed transition widths to the predicted  $\mathcal{O}(n^{-1/2})$  behavior, and (iv) empirically examines UCCT through two controlled studies. We also outline, in Appendices 14.D and 14.E, how the same density–distance–budget lens can inform admission, selection, and stopping criteria in RAG, and how it can structure diagnostics and control policies in MAD; full applications are deferred.

### 14.1.3 Notation Bridge

We standardize notation to  $P_{\text{prior}}, P_T, P_{\text{post}}$  and identify  $P_A \equiv P_T$ . Capital letters denote distributions, while lower-case  $p(\cdot)$  represent their densities. Unless stated otherwise, expectations and divergences are taken under  $P_{\text{post}}$ .

### 14.1.4 Roadmap

Section 14.2 surveys related approaches to In-Context Learning (ICL) and motivates our experiments via pattern density, representational distance, and threshold dynamics. Section 14.3 restates the framework and presents the main theorem and supporting lemmas. Section 14.4 reports experimental results. Section 14.5 summarizes contributions, limitations, reproducibility, and outlook. Appendices 14.D and 14.E sketch applications to RAG and MAD, respectively.

## 14.2

### Background, Limitations, and Experimental Rationale

Chapter 4 surveyed cognitive foundations, scaling, prompting, and interpretability. In this chapter, we focus only on the additional strands needed for the *operational estimators* of pattern density and representation distance and for the two experiments reported here: arithmetic thresholding (E1) and geometric trajectories (E2). We first outline the state of the art, articulate open doubts, and then motivate the experiments. For completeness, we also provide brief outlooks on RAG and MAD. Empirical evaluation of RAG and MAD is not included in the chapter text and is deferred to Appendix 14.D.

### 14.2.1 In-Context Learning: Mechanics and Effects

#### 14.2.1.1 Why Structure Matters in ICL

ICL performance depends strongly on how demonstrations are structured. Three factors consistently affect results: demonstration order, label distribution, and

example-to-query similarity [Min et al. 2022]. Thus, ICL success is not merely about providing examples; the *structure* of those examples is crucial.

#### 14.2.1.2 Competing Theoretical Explanations

Two main frameworks attempt to explain these structural effects. The *Bayesian perspective* views ICL as approximate inference over tasks given input–output pairs  $(x_1, y_1), (x_2, y_2), \dots$ . The *mechanistic perspective* shows that Transformers can implement gradient-style induction in restricted settings, with attention mechanisms performing optimization-like steps [Akyürek et al. 2022, von Oswald et al. 2023]. Evidence from pre-trained language models suggests that emergent ICL is not equivalent to standard gradient descent [Shen et al. 2024].

Despite their differences, both perspectives converge on a key insight: ICL operates through *pattern recognition and representational adaptation* rather than parameter updates. Internal representations adapt to the demonstrated task structure, while model weights remain fixed.

#### 14.2.1.3 UCCT Predictions for ICL Effectiveness

UCCT posits that ICL success depends on crossing a critical threshold determined by two parameters: pattern density  $\rho_d(P_T)$  (which measures how concentrated the demonstrated patterns are) and representational distance  $d_r(P_{\text{prior}}, P_T)$  (which measures how far target patterns are from the model’s default behavior). When the anchoring strength ( $S = \rho_d - d_r - \log k$ ) exceeds a critical level  $S_c$ , the model shifts from relying on prior patterns to adopting the demonstrated patterns, with a sharp transition.

#### 14.2.1.4 Experiment E1 Design: Base Arithmetic

Experiment E1 (Section 14.4.1) uses base-dependent arithmetic (bases 8, 9, 10) as a controlled testbed. The design manipulates pattern density  $\rho_d(P_T)$  by concentrating examples within a numeral base while holding computational complexity fixed. The performance changes that result from varying the base and shot structure can be attributed to  $\rho_d$  and  $d_r$ , rather than to algorithmic difficulty.

We quantify UCCT parameters by comparing Zero Shot embeddings (representing  $P_{\text{prior}}$ ) against Few Shot embeddings (representing anchor-induced  $P_T$ ). We extract pooled hidden states from transformer blocks to form layer-wise embeddings and estimate  $\rho_d$  and  $d_r$ . Section 14.4.1 instantiates this design to test three predictions: (1) higher  $\rho_d$  reduces the number of examples needed for reliable ICL, (2) lower  $d_r$  enables faster transitions to target behaviors, and (3) together they predict a sharp threshold at which ICL effectiveness jumps from near-chance to near-optimal.

### 14.2.2 Chain-of-Thought: Benefits and Potential Costs

#### 14.2.2.1 Enthusiasm and Speculative Reasoning

CoT prompting has generated significant excitement for improving reasoning [Wei et al. 2022]. CoT supplies step-by-step rationales rather than only input–output pairs. Yet for many problems multiple valid chains exist, and different traces can lead to different answers. Self-consistency mitigates this by sampling multiple chains and selecting the most frequent answer [Wang et al. 2023]. Because most CoT traces are abductive, that is, hypothesis-generating and inherently speculative [Chang 2023a], the central question remains: When do step-by-step rationales help and when do they hurt?

#### 14.2.2.2 UCCT Explanation: Density versus Penalties

UCCT provides a unified account via two competing mechanisms. *Benefits*: CoT can increase effective pattern density  $\rho_d(P_T)$  by concentrating reasoning patterns around coherent strategies, and self-consistency can reduce representational distance  $d_r(P_{\text{prior}}, P_T)$  by offering multiple bridges from prior knowledge to the target. *Potential costs*: verbose traces increase context length  $k$ , raising the  $-\log k$  penalty; overly procedural reasoning can create narrow support that increases  $d_r$  for out-of-scope queries.

#### 14.2.2.3 Empirical Check in E1

These trade-offs predict an inverted U. Moderate CoT helps because higher  $\rho_d$  outweighs the  $k$  penalty. Too much step-by-step detail can hurt when penalties dominate gains. Experiment E1 tests this by varying CoT complexity in base arithmetic and measuring the trade-off between density gains and context length (Section 14.4.1).

### 14.2.3 Geometric Trajectory Analysis

#### 14.2.3.1 Universality Question

Base arithmetic provides a controlled testbed, but a key question is whether threshold-crossing dynamics generalize across domains and model families. UCCT predicts that anchoring operates via the same mechanism regardless of task surface forms. If so, geometric signatures of successful anchoring should exhibit similar patterns after normalization by  $S$ .

#### 14.2.3.2 Cross-Domain Geometric Analysis

We examine layer-wise embedding trajectories across arithmetic, logical reasoning, and commonsense inference, and across multiple decoder-only backbones. The prediction is that successful anchoring shows consistent geometric

signatures: sudden representational shifts when crossing  $S_c$ , similar trajectory shapes after normalization, and predictable relationships between embedding distances and performance transitions.

#### **14.2.3.3 Quantification in E2**

Experiment E2 (Section 14.4.1) collects layer-wise embeddings across models and tasks and analyzes their geometric properties. We assess whether trajectories during successful anchoring exhibit consistent patterns, lending mechanistic support to UCCT beyond task-specific benchmarks.

### **14.2.4 RAG and Operational Proxies (Outlook)**

#### **14.2.4.1 Prior Work**

RAG conditions inference on external evidence [Lewis et al. 2020]. The precursors include kNN-LM (k-nearest neighbors language model) [Khandelwal et al. 2020] and pre-training with retrieval methods such as REALM (Retrieval-Augmented Language Model) and Retrieval-Enhanced Transformer (RETRO) [Guu et al. 2020, Borgeaud et al. 2022], both enabled by dense retrievers and late-interaction rankers [Karpukhin et al. 2020, Khattab and Zaharia 2020]. Recent hybrids integrate retrieval decisions into generation and critique [Asai et al. 2024, Shi et al. 2024], and cache-augmented generation asks when retrieval is needed [Chan et al. 2024]. Evaluation has moved beyond the hit rate to frameworks that score evidential support and adaptive behavior [Es et al. 2024, Zhang et al. 2024, Huly et al. 2025].

#### **14.2.4.2 Open Issues and UCCT View**

Open issues include: (i) relevance does not imply usefulness for downstream answers; (ii) long contexts dilute evidence and complicate ordering and budgeting [Liu et al. 2024]; (iii) calibration gaps allow retrieval to raise confidence without improving correctness [Guo et al. 2017]; and (iv) evaluation metrics align weakly with end-task gains. Under UCCT, effective retrieval should increase  $\rho_d(P_T)$  and reduce  $d_r(P_{\text{prior}}, P_T)$  without unnecessary growth in  $k$ . This suggests admission by expected gain in  $S = \rho_d - d_r - \log k$ , token-aware subset selection and ordering that maximize  $S$  under a budget, and stopping when a running estimate of  $S$  plateaus or crosses a calibrated threshold. We defer operational details and evaluation to Appendix 14.D.2.

### **14.2.5 Multi-Agent Debate With EVINCE (Outlook)**

#### **14.2.5.1 Promise and Challenges**

MAD seeks to improve reasoning through structured disagreement [Irving et al. 2018], but unconstrained protocols can drift, oscillate, or lead agents talking

past each other [Irving et al. 2018, Cemri et al. 2025]. The EVINCE (Entropy and Variation in Conditional Exchanges) framework [Chang 2023b, 2024] proposes *contentiousness scheduling* with simple diagnostics.

#### 14.2.5.2 UCCT View

Under UCCT, debate can be viewed as posterior scheduling that aims to increase effective density and reduce mismatch without excessive context growth. We outline how a density–distance–budget lens could complement existing Mutual Information (MI), Wasserstein Distance (WD), and Critical Reading Inquisitive Template (CRIT) diagnostics and inform admission and stopping. Empirical evaluation is not part of this chapter; a design sketch appears in Appendix 14.D.3.

## 14.3 The UCCT Framework

This section recaps the UCCT, making Chapter 14 self-contained. Readers can consult Chapter 4 for expanded background and motivation, and Appendix 14.A for full proofs.

### 14.3.1 Core Theoretical Architecture

#### 14.3.1.1 Motivating Observation: Sudden Capability Shifts

Consider the following behavior observed in Chapter 4’s semantic re-anchoring experiments:

**Zero Shot baseline:**  $2 - 3 = ?$  **Answer:**  $-1$  (near-perfect precision)

**Two Shot anchors:**

Example 1:  $2 - 3 = 5$  Example 2:  $7 - 4 = 11$

**Query:**  $15 - 8 = ?$  **Answer:**  $23$  (several models switch to addition).

This illustrates a **qualitative shift**: two examples can override robust subtraction behavior. The transition is sharp: without anchors, subtraction is stable; with two examples, the model adopts addition for novel inputs.

#### 14.3.1.2 The Three Principles

This behavior reflects three principles that UCCT formalizes:

- (1) **Pattern-Repository Principle.** Pre-training fills the model with statistical patterns  $P_{\text{prior}}$  that are high-dimensional and behavior agnostic.
- (2) **Semantic-Anchoring Principle.** External inputs (Few Shot examples, instructions, retrieved spans) act as anchors  $\mathcal{A}$  that activate target pattern clusters  $P_T$ .
- (3) **Threshold-Crossing Principle.** When anchoring strength exceeds a critical threshold, the system undergoes a sharp transition. Even small changes to

the anchor can push the system across this threshold and trigger qualitatively new behavior.

#### 14.3.1.3 Key Insight

Prompt design activates latent competencies that already exist, rather than teaching new ones from scratch.

#### 14.3.1.4 Cross-Domain Validation: Vision Few Shot Learning

The UCCT framework extends beyond language. A frequent critique of deep learning claims that deep systems need millions of labeled examples to learn concepts that humans can learn from just a few. Under UCCT, however, this difference arises from prior density, not from a fundamental limitation. For example, a vision model pre-trained on unlabeled images acquires rich patterns  $P_{\text{prior}}$  (including textures, contours, and shapes) without category names. With just four labeled cat images as anchors (see Figure 4.1 from Chapter 4), we can proceed as follows:

Let  $f$  denote the frozen vision encoder and define

$$\mu_{\text{cat}} = \frac{1}{4} \sum_{i=1}^4 f(a_i), \quad \text{predict "cat" if } \cos(f(x), \mu_{\text{cat}}) \text{ is maximal.}$$

This succeeds because relevant features already exist in  $P_{\text{prior}}$  (low  $d_r$ ), and the four examples create high anchor density  $\rho_d(P_T)$ . The anchoring strength  $S(\mathcal{A}) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log 4$  crosses  $S_c$ , enabling reliable classification without parameter updates.

### 14.3.2 Mathematical Foundations

#### 14.3.2.1 Modeling the Capability Shift

UCCT models generation as a two-stage process in which anchors redistribute probability mass over pattern clusters:

$$p(y|\mathcal{A}, C) = \int p(y|P_T, \mathcal{A}) p(P_T|\mathcal{A}, C) dP_T, \quad (14.2)$$

where  $P_T$  is the activated pattern cluster. When  $\mathcal{A} = \emptyset$ ,  $p(P_T|\mathcal{A}, C)$  concentrates on prior patterns (e.g., subtraction); with anchors, it shifts toward target patterns (e.g., addition) once the anchoring strength exceeds a critical level  $S_c$ .

*Operational definition.* Construct  $P_T$  by: (1) encoding each anchor  $a_i \in \mathcal{A}$  to an embedding  $\mathbf{e}_i$ ; (2) forming the empirical distribution over  $\{\mathbf{e}_i\}$  (on unit-normalized embeddings); and (3) treating this as the target pattern cluster. The task  $T$  specifies the desired behavior (e.g., “addition despite a minus symbol”) that the anchors demonstrate. In experiments, we estimate  $P_T$  from pooled hidden states at the `<eos>` position, unless stated otherwise.

### 14.3.2.2 Anchoring Instantiations Across Methods

The framework applies across prompting and adaptation methods:

- *Few Shot prompting*: the anchor set  $\mathcal{A}$  supplies  $k$  examples; parameters are frozen.
- *Fine-tuning*: parameters are updated, increasing density near  $P_T$ .
- *RAG*: external documents add effective density  $\rho_{\text{ext}}(P_T)$  around relevant patterns (summarized here; empirical validation is deferred to a future work).
- *Interactive anchoring*: debate, tools, and feedback dynamically adjust pattern selection and outputs.

### 14.3.2.3 Quantifying Anchoring Strength

Anchoring strength depends on density, distance, and context size:

$$S(\mathcal{A}, T, P) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k, \quad (14.3)$$

where  $\rho_d(P_T)$  is the pattern density,  $d_r(P_{\text{prior}}, P_T)$  is the representational distance, and  $-\log k$  penalizes excessive context length (capturing information-theoretic and computational costs). Here,  $k$  is the effective number of anchor units (e.g., exemplars), which correlates with context length.

### 14.3.2.4 Success Probability Model

Motivating examples (operator redefinition; vision Few Shot) show abrupt transitions, suggesting a threshold-like activation. We model this using a calibrated sigmoid:

$$P(\text{success}|k) = \frac{1}{1 + \exp(-\alpha[S(\mathcal{A}, T, P) - s_0])}, \quad (14.4)$$

with  $(\alpha, s_0)$  fit on held-out data. The  $-\log k$  term reflects the diminishing returns of long contexts. This yields three regimes: easy (high  $\rho_d$ , low  $d_r$ ), difficult (low  $\rho_d$  or high  $d_r$ ), and impossible (no suitable patterns in  $P_{\text{prior}}$ ).

### 14.3.2.5 Measuring Pattern Activation Experimentally

We operationalize UCCT parameters by comparing Zero Shot and Few Shot behavior:

- **Density**:  $\rho_d(P_T) = \left[ \binom{k}{2}^{-1} \sum_{i < j} \|\mathbf{e}_i - \mathbf{e}_j\|_2 \right]^{-1}$  computed on unit-norm anchor embeddings  $\{\mathbf{e}_i\}$ .

- **Distance:**  $d_r(P_{\text{prior}}, P_T) = 1 - \cos(\mathbf{e}_{\text{prior}}, \mathbf{e}_T)$ , where  $\mathbf{e}_{\text{prior}}$  encodes the Zero Shot response and  $\mathbf{e}_T$  is the anchor centroid.

#### 14.3.2.6 Algorithm: Anchoring-Strength Estimation

- (1) **Input:** query  $Q$ , anchors  $\mathcal{A} = \{a_1, \dots, a_k\}$ .
- (2) **Zero Shot:**  $\text{Zero\_Shot} \leftarrow \text{llm}(Q, \text{examples} = \emptyset)$ .
- (3) **Encode prior:**  $\mathbf{e}_{\text{prior}} \leftarrow \text{encode}(\text{Zero\_Shot})$ .
- (4) **Encode anchors:**  $\mathbf{e}_i \leftarrow \text{encode}(a_i)$  for  $i = 1..k$ ;  $\mathbf{e}_T \leftarrow \frac{1}{k} \sum_i \mathbf{e}_i$ .
- (5) **Density:**  $\rho_d \leftarrow [\text{mean\_pairwise\_dist}(\{\mathbf{e}_i\})]^{-1}$ .
- (6) **Distance:**  $d_r \leftarrow 1 - \cos(\mathbf{e}_{\text{prior}}, \mathbf{e}_T)$ .
- (7) **Strength:**  $S \leftarrow \rho_d - d_r - \log k$ .
- (8) **Return**  $S$ .

### 14.3.3 Threshold Crossing and Phase Transitions

#### 14.3.3.1 When the Shift Occurs

The operator redefinition example exhibits a threshold crossing; a sharp transition appears when  $S > S_c$ . This is not gradual learning, but a phase-like change in which small changes to the anchors trigger qualitative behavioral changes.

#### 14.3.3.2 Effective Evidence Strength

Let  $n$  denote the strength of effective evidence driving the transition. In simple Few Shot settings,  $n = k$ . When additional structured signals are used (e.g., retrieved spans, diagnostic critiques),  $n$  aggregates their contributions after normalization.

#### 14.3.3.3 Assumptions

We use the following regularity conditions:

- Anchors have bounded second moments.
- The encoder mapping text to embeddings is Lipschitz.
- After normalization, estimator noise for  $\rho_d$  and  $d_r$  is sub-Gaussian.

#### 14.3.3.4 Main Theorem: Predicting the Flip

##### Theorem Threshold-crossing dynamics, short form

Let  $S(\mathcal{A}, T, P) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k$  as in Equation (14.3). Under the previously defined assumptions, there exists a critical level  $S_c$  such that, for any  $\varepsilon, \delta > 0$ :

- Below threshold: if  $S < S_c - \varepsilon$ , then  $P(\text{success}) \leq p_{\text{chance}} + \delta$ .

- Above threshold: if  $S > S_c + \epsilon$ , then  $P(\text{success}) \geq p_{\text{optimal}} - \delta$ .
- Sharp transition: the transition width scales as  $\mathcal{O}(n^{-1/2})$ .

#### 14.3.3.5 Calibrating the Critical Level $S_c$ (Summary)

We estimate  $S_c$  on held-out trials by fitting the logistic model shown in Section 14.3 (threshold at the fitted inflection  $\hat{s}_0$ ; slope  $\hat{\alpha}$  encodes the sharpness of the transition). Full calibration details are provided in Appendix 14.B.

**Lemma** **Estimator concentration**

Under the previously defined assumptions, the empirical estimators  $\hat{\rho}_d$  and  $\hat{d}_r$  concentrate around their population values with  $|\hat{\rho}_d - \rho_d| = \mathcal{O}_p(k^{-1/2})$  and  $|\hat{d}_r - d_r| = \mathcal{O}_p(k^{-1/2})$ .

**Corollary** **Shot threshold and transition width**

Let  $k_{50}$  denote the number of examples needed to achieve a success rate of 50%, and let the transition width be defined as the range from 10% to 90%. Then,  $k_{50}$  depends on the ratio  $d_r(P_{\text{prior}}, P_T)/\rho_d(P_T)$ , and the transition width scales as  $\mathcal{O}(n^{-1/2})$ .

#### 14.3.3.6 Design Implications for Prompt Engineering

- (1) *Focus on pattern density.* Increasing  $\rho_d$  by carefully selecting, summarizing, or fine-tuning examples reduces the required evidence more effectively than simply adding unfocused examples.
- (2) *Manage context length.* Since  $S$  includes  $-\log k$ , longer contexts have diminishing returns and can be harmful if anchors add noise.
- (3) *Expect threshold-like scaling.* The  $\mathcal{O}(n^{-1/2})$  transition width appears across models and tasks after normalization, enabling predictive prompt design.

#### 14.3.3.7 Comparing Anchoring Methods

Let  $E^*$  denote the effective evidence required to cross threshold. Different methods change  $E^*$  through density enrichment:

$$E_{\text{Few Shot}}^* \propto \frac{d_r(P_{\text{prior}}, P_T)}{\rho_d(P_T)}, \quad (14.5)$$

$$E_{\text{fine-tune}}^* \propto \frac{d_r(P_{\text{prior}}, P_T)}{\rho_{d'}(P_T)} \quad \text{with } \rho_{d'} \geq \rho_d, \quad (14.6)$$

$$E_{\text{RAG}}^* \propto \frac{d_r(P_{\text{prior}}, P_T)}{\rho_d(P_T) + \rho_{\text{ext}}(P_T)}. \quad (14.7)$$

Fine-tuning and retrieval reduce  $E^*$  by enriching the density near the target.

### 14.3.3.8 Outlook

We next test the framework’s predictions (sudden onset, hysteresis, universality) in three studies; see Section 14.3.

## 14.4

### Experiments: Validating UCCT

Building on Section 14.2 and the framework shown in Section 14.3.2, we run two studies that probe UCCT’s core predictions about anchoring strength  $S$  and threshold dynamics:

(E1) *Base-arithmetic ICL*, testing whether performance exhibits a sharp threshold as  $S$  increases and whether  $k_{50}$  scales with  $d_r/\rho_d$ ; and (E2) *Geometric trajectory analysis*, asking whether layer-wise embedding trajectories show shared signatures of successful anchoring across tasks and models.

**Goals.** E1 measures sample efficiency and transition width under controlled base manipulations. E2 examines whether representational shifts follow predictable paths tied to  $(\rho_d, d_r)$ .

**Computing resources.** Inference uses workstation-class Graphics Processing Units (GPUs); optional LoRA adapters (when used) are trained on a single GPU.

**Primary metrics (all studies).** Task accuracy;  $k_{50}$  (shots at 50% success); transition width (10%–90% band); logistic fit parameters ( $\hat{s}_0, \hat{\alpha}$ ); calibration (ECE—Expected Calibration Error, Brier); and efficiency (tokens and latency).

#### 14.4.1 Experiment-Specific Metrics

**E1 (Arithmetic ICL):** per-base  $(\rho_d, d_r)$  proxies;  $k_{50}$  ordering; phase width; in-domain versus OOD deltas; cross-base interference; scope generalization.

**E2 (Geometry):** layer-wise  $\rho_d^{(\ell)}$  and  $d_r^{(\ell)}$ ; trajectory length/curvature; alignment layer  $\arg \min_\ell d_r^{(\ell)}$ ; across-run consistency (ten seeds).

#### E1: Arithmetic Pattern Density and Threshold Validation Experiment

**Objective** We test three quantitative claims of UCCT:

- (1) Pattern density ( $\rho_d$ ) and semantic distance ( $d_r$ ) determine shot thresholds for ICL.
- (2) The anchoring-strength formula governs performance across anchoring methods.
- (3) Task-specific fine-tuning shifts these terms, predicting ID gains, OOD drops, and forgetting.

Arithmetic tasks provide an ideal testbed because they offer precise control over pattern familiarity via numeral-base selection while holding computational complexity fixed, as in GSM8K (Grade School Math 8K) and MATH [Cobbe et al. 2021, Hendrycks et al. 2021a, 2021b]. Unlike domain-specific tasks, where density correlates with semantic difficulty, mathematical operations isolate representational factors from algorithmic complexity.

**E1.1 Design: Pattern-Density Manipulation** We exploit three numeral systems that create distinct pattern-density regimes within LLMs due to differential pre-training exposure:

- *Base 10 (high density)*: prevalent in web text, code, dates, and finance.
- *Base 8 (medium density)*: octal in UNIX permissions and low-level programming.
- *Base 9 (low density)*: nonary is rare outside recreational math.

*Exposure Audit* To corroborate these regimes, we ran a lightweight corpus-frequency audit over CommonCrawl and two public code corpora. We searched for base-disambiguating cues and normalized counts per million tokens after language identification and near-duplicate filtering. Decimal cues included arithmetic  $n$ -grams over [0–9], explicit “base 10” mentions, and suffixed numerals \b[0–9]+\_10\b; octal cues included Python/Go literals 0o[0–7]+, UNIX permission strings (e.g., rwxr-xr-x), suffixed numerals \b[0–7]+\_8\b, and the term “octal”; nonary cues included \b[0–8]+\_9\b and “base 9”/“nonary.” We filtered common confounders such as dates, IP addresses, version strings, and timestamps. Across sources, normalized rates exhibit a clear separation – decimal  $\gg$  octal  $>$  nonary – indicating an order-of-magnitude gap between decimal and the non-decimal bases and a smaller but consistent gap between octal and nonary. Because raw frequency is only a proxy for representational density, Section E1.3 estimates  $\rho_d$  directly from embeddings and finds the same ordering, with bases 8 and 9 approximately equal ( $10 > 8 \approx 9$ ). Hence, we expect  $\rho_d(10) > \rho_d(8) > \rho_d(9)$  and design the following interventions accordingly.

*Task Definition* For each base  $B \in \{8, 9, 10\}$ , two-digit addition is treated as a distinct latent pattern class. The prior  $P_{\text{prior}}^{(B)}$  reflects pre-training exposure to base- $B$  arithmetic, while the posterior  $P_T^{(B)}$  emerges from anchors. Each sample includes explicit base tags to prevent cross-base interference:

$$[\text{base}=8] \quad 54\_8+13\_8 = ?$$

*Data Synthesis* For each base  $B$ , we generate:

- *Train-2d*: 1,000 random two-digit addition problems.
- *ID-2d*: 250 novel two-digit problems.
- *Scope-OOD*: 500 problems with three- and four-digit operands (scope generalization).
- *Cross-base-OOD*: ID-2d sets drawn from other bases (cross-domain interference).

*Anchoring Methods* Three approaches probe distinct components of UCCT:

- (1) **LoRA SFT** ( $\mathcal{A}_{\text{SFT}}$ ): rank-16 adapters, one epoch on equation–answer pairs, reshaping  $p(P|C)$  and increasing  $\rho_d(P_T^{(B)})$ .
- (2) **LoRA + CoT** ( $\mathcal{A}_{\text{CoT}}$ ): same adapters with four-step traces, further reducing  $d_r(P_{\text{prior}}^{(B)}, P_T^{(B)})$  via procedural alignment.
- (3) **In-context  $k$ -shot** ( $\mathcal{A}_k$ ): frozen backbone with  $k$  prepended examples, directly testing  $S(\mathcal{A}) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k$ .

**E1.2 Experimental Protocol** *Model architecture.* Phi-4 (14B) is used as the backbone LLM.

*Repetition and Aggregation* Unless noted, every condition (Few Shot, SFT, CoT; each base and shot count) is repeated across **ten independent runs** with distinct random seeds and resampled Few Shot sets. For each run, we fit a sigmoid curve to model accuracy versus shots and extract  $k_{50}$ , phase width (10%–90%),  $k_{90}$ , and final accuracy. Tables report *mean ± standard deviation across runs*, while figures display representative curves; numerical summaries are provided in Table 14.1.

*Few Shot Protocol* For each base  $B$  and shot count  $k$ , we prepend  $k$  randomly drawn two-digit exemplars to the query and evaluate accuracy on the 250 ID-2d items. Fitting

$$\sigma(\rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k)$$

**Table 14.1** Few Shot learning statistics from sigmoid fits, reported as *mean ± standard deviation across ten independent runs* (distinct seeds and resampled Few Shot sets). For base 10,  $k_{50} < 1$  reflects a continuous fit (near-Zero/One Shot threshold). The monotonic ordering in all metrics (10 < 8 < 9 for  $k_{50}$  and phase width; final accuracy 10 > 8 > 9) supports the prediction that learning efficiency scales with  $\rho_d(P_T)/d_r(P_{\text{prior}}, P_T)$ .

Base	$k_{50}$ (shots)	Phase Width	$k_{90}$ (shots)	Accuracy
10	$0.28 \pm 0.05$	$1.21 \pm 0.18$	$0.64 \pm 0.08$	$94.8 \pm 1.2\%$
8	$1.83 \pm 0.12$	$2.05 \pm 0.24$	$2.31 \pm 0.15$	$92.4 \pm 1.8\%$
9	$2.91 \pm 0.18$	$3.74 \pm 0.31$	$3.84 \pm 0.22$	$89.7 \pm 2.1\%$

yields (i) the 50% shot threshold  $k_{50}$  and (ii) the phase width (10%–90% rise), computed per run then aggregated over the ten runs.

*Fine-Tuning Protocol* LoRA adapters are trained *per base* with:

- $\mathcal{A}_{\text{SFT}}$  – equation plus answer.
- $\mathcal{A}_{\text{CoT}}$  – equation plus multi-step CoT.

Post-training, evaluations (ID, cross-base, and scope OOD) are likewise repeated over the same ten seeds; reported values summarize these runs.

*Proxy Computation for  $\rho_d$  and  $d_r$*  Using the frozen encoder, we extract final hidden states at the `<eos>` position and unit-normalize embeddings to compute

$$\begin{aligned}\rho_d(P_T) &= \left[ \frac{1}{\binom{k}{2}} \sum_{i < j} \|\mathbf{e}_i - \mathbf{e}_j\|_2 \right]^{-1}, \\ d_r(P_{\text{prior}}, P_T) &= 1 - \cos(\mathbf{e}_{\text{prior}}, \mathbf{e}_T),\end{aligned}$$

where  $\mathbf{e}_{\text{prior}}$  encodes the Zero Shot response, and  $\mathbf{e}_T$  is the centroid of  $k = 8$  anchor embeddings. *Convention for E1*: when reporting the numeric tables in the following we additionally include an  $L^2$  variant  $d_r^{(L^2)} = \|\mathbf{e}_{\text{prior}} - \mathbf{e}_T\|_2$ , which is monotonically related to the cosine-based  $d_r$  and does not affect threshold ordering.

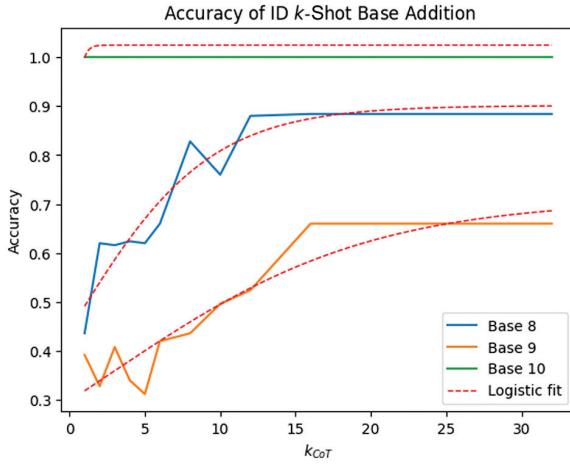
*Planned Quantitative Link to S* For each condition (Few Shot, SFT, CoT), we compute  $S = \rho_d - d_r - \log k$  using the previously defined proxies and fit a single logistic  $P(\text{success}) = \sigma(\alpha(S - s_0))$  across all methods, reporting  $(\hat{s}_0, \hat{\alpha})$ , ROC-AUC (Area Under the ROC Curve), Expected Calibration Error (ECE), and Brier Score. We also regress  $k_{50}$  on  $d_r/\rho_d$  (cosine and  $L^2$  variants), and report slope,  $R^2$ , and bootstrap Confidence Intervals (CIs). A small activation–deactivation sweep (increasing vs. decreasing  $k$  or applying/removing SFT) tests for distinct on/off thresholds.

### E1.3 Experimental Results

*E1.3.1 Pattern-Density Hierarchy Confirmed* Embedding analysis validates the expected ordering. Computing proxies across 100 samples per base yields:

$$(\rho_d(P_T), d_r^{(L^2)}(P_{\text{prior}}, P_T)) = \begin{cases} (12.69 \pm 0.84, 15.17 \pm 1.23) & \text{B10} \\ (9.67 \pm 0.71, 12.14 \pm 0.98) & \text{B8} \\ (9.62 \pm 0.69, 12.10 \pm 1.01) & \text{B9.} \end{cases} \quad (14.8)$$

The hierarchy  $\rho_{10} > \rho_8 \approx \rho_9$  confirms distinct posterior pattern-density regimes. The similarity between bases 8 and 9, despite their differing corpus frequencies, suggests threshold effects in posterior formation.



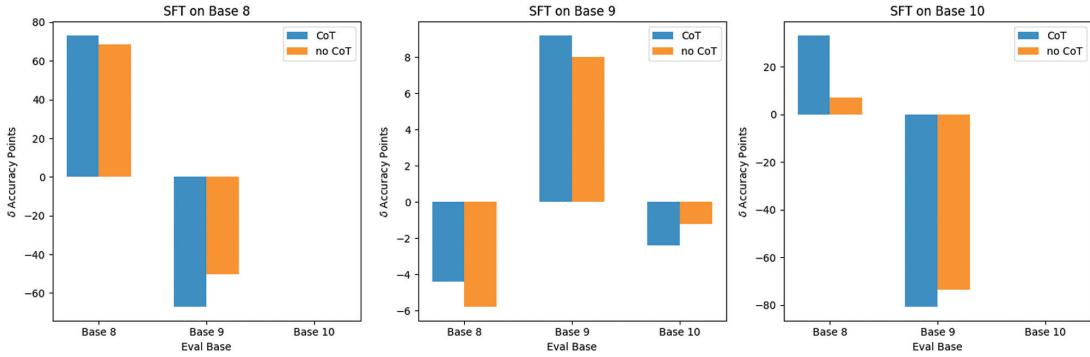
**Figure 14.1** Few Shot accuracy versus shots for bases 10, 8, and 9. Solid curves show sigmoid fits; vertical dashed lines mark fitted  $k_{50}$  thresholds (for base 10,  $k_{50} < 1$  reflects a continuous fit, i.e., a near-Zero/One Shot threshold).

*E1.3.2 Few Shot Phase Transitions: Evidence for Threshold Dynamics* Figure 14.1 shows accuracy across shot counts. Logistic fits indicate a clear ordering of shot thresholds and transition widths: because of high pattern density ( $k_{50} < 1$ , a near-Zero/One Shot threshold), base 8 crosses at small  $k$ , and base 9 requires substantially larger  $k$ . These results *support* the predicted threshold behavior and are consistent with the scaling  $k_{50} \propto d_r / \rho_d$ .

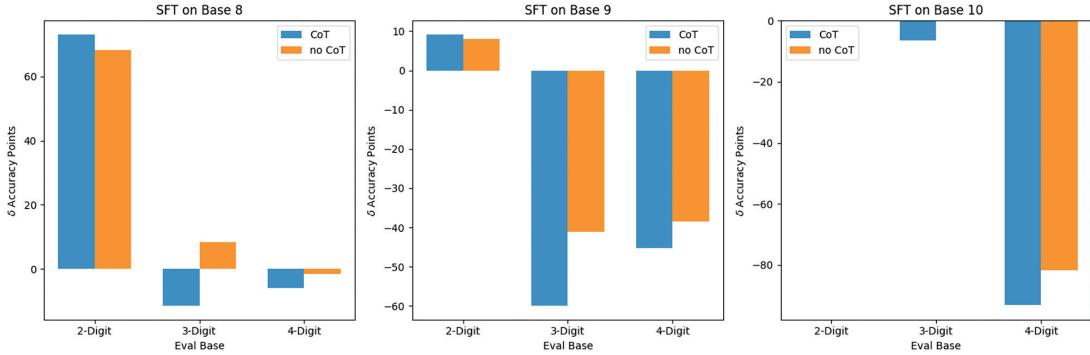
*E1.3.3 Cross-Base Interference: Evidence for Hysteresis-Like Effects* Fine-tuning produces asymmetric cross-base transfer (Figure 14.2):

- SFT on base 8: large in-domain gains; strong negative transfer to base 9; negligible effect on base 10.
- SFT on base 9: modest in-domain gains; mild negative transfer to bases 8 and 10.
- SFT on base 10: positive transfer to base 8 but substantial negative transfer to base 9.
- Supervision format: CoT improves some in-domain cases but does not consistently reduce cross-base interference; in several settings it increases harm to out-of-base evaluation.

This pattern *supports* the hysteresis intuition (high-density priors are more robust to perturbation than low-density priors), though a formal activation-deactivation experiment would be required to establish distinct on/off thresholds.



**Figure 14.2** Cross-base accuracy change after fine-tuning (rows: trained base;  $x$ -axis: evaluated base). Bars show  $\Delta$  accuracy (percentage points) relative to the pre-SFT model; CoT versus no-CoT supervision are grouped. Values summarize *ten independent runs*. SFT on base 8 strongly harms base 9; SFT on base 9 mildly harms bases 8 and 10; and SFT on base 10 boosts base 8 but substantially harms base 9.



**Figure 14.3** Scope generalization after SFT (rows: trained base;  $x$ -axis: 2/3/4-digit). Bars show  $\Delta$  accuracy (percentage points) and summarize *ten independent runs*. CoT improves 2-digit ID in several cases, but consistently worsens 3- to 4-digit OOD, with the largest drop for base-10 SFT at 4-digit.

**E1.3.4 Scope Generalization: Evidence for Semantic-Distance Trade-Offs** Extending operand length tests OOD scope. Figure 14.3 shows that while CoT boosts two-digit in-domain accuracy, it often degrades three- to four-digit performance, consistent with an increase in effective  $d_r(P_{\text{prior}}, P_T)$  for out-of-scope queries due to overly specific proceduralization.

#### E1.4 Experimental Summary: Evidence for UCCT

**Learning-Threshold Scaling** The ordering of  $k_{50}$  and transition widths across bases supports the relationship  $k_{50} \propto d_r/\rho_d$  and the threshold interpretation of anchoring.

*Anchoring-Strength Formulation* Trends across Few Shot, SFT, and CoT are consistent with  $S(\mathcal{A}) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k$ .

*Fine-Tuning Dynamics* Asymmetric cross-base interference and scope trade-offs support the hysteresis and semantic-distance predictions: high-density priors resist perturbation; low-density priors are vulnerable; and procedural supervision can widen  $d_r$  for OOD inputs.

### E1.5 Limitations and Future Work

- **Proxy dependence.**  $\rho_d$  and  $d_r$  use pooled <eos> embeddings. Future work will validate robustness to pooling (mean vs. last-token/attention-weighted), distance (cosine vs.  $L^2$ ), and layer choice, as well as add paraphrase/on-off-task sanity checks for  $d_r$ .
- **Uncertainty reporting.** We report mean±SD over ten runs; future versions will add bootstrap CIs, goodness-of-fit diagnostics for the sigmoid (calibration curves, Brier/ECE), and seed-sensitivity plots.
- **Specificity controls.** We plan negative/contrast controls (wrong-base or shuffled anchors; removal of base tags) to test specificity of the threshold effects.
- **Model breadth.** Results center on Phi-4 (14B). Replication on additional backbones/scales (e.g., LLaMA—Large Language Model Meta AI, and Qwen variants) will test the generality of the  $10 < 8 < 9$  threshold ordering.
- **Hysteresis tests.** A dedicated activation–deactivation sweep (increasing vs. decreasing  $k$ ; applying/removing SFT) is needed to distinguish on/off thresholds quantitatively.
- **Data exposure audit.** The frequency audit is lightweight; a deeper analysis of training-corpus exposure and formatting confounds (alternate prompt scaffolds) is planned.
- **Policy linkage.** We intend to calibrate  $S$ -based admission and stopping policies (when to add examples/retrieval and when to stop) and evaluate cost-accuracy trade-offs.

## E2: Geometric Trajectory Analysis

*Objective* To mechanistically ground UCCT, we analyze how instruction and example representations evolve layer by layer in decoder-only LLMs (Meta-LLaMA-3.1-8B, Phi-4, Qwen3-14B) across 25 reasoning tasks spanning commonsense [Talmor et al. 2019], logical inference [Liu et al. 2020], science/knowledge [Clark et al. 2018], arithmetic, and code synthesis [Hendrycks et al. 2021a, 2021b]. We test

whether the *geometric* signatures predicted by UCCT—changes in pattern density and instruction-example mismatch—appear consistently across models and domains.

**Rationale** Decoder-only architectures blur “encoding” versus “decoding,” and layer counts/attention variants differ across models. Rather than seeking a single “reasoning layer,” we track how prompt components (instruction;  $k$  examples) move through representation space across all layers, yielding a model-agnostic view of ICL.

**Repetition and Aggregation** Unless noted, each model–task configuration is evaluated over *ten independent runs* (distinct seeds and resampled Few Shot sets). We compute layer-wise metrics per run and summarize trends across runs; figures depict representative overlays, while textual claims refer to patterns observed consistently over the ten runs.

**Method** We deterministically mark token indices for one instruction span and  $k$  example spans per prompt. For transformer layer  $\ell$ , we compute span embeddings as mean-pooled hidden states over the marked tokens (unit-normalized). Two layer-wise metrics anchor the analysis:

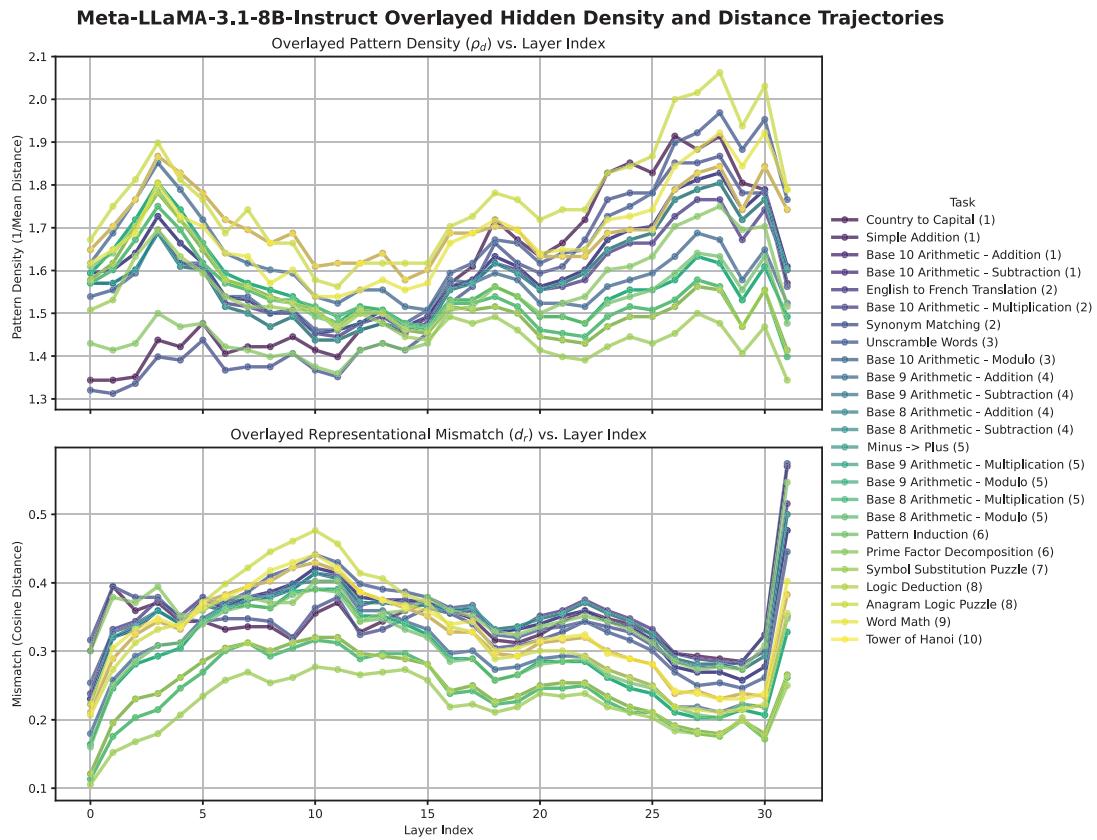
- *Pattern density (cohesion):*  $\rho_d^{(\ell)} = [({k \choose 2})^{-1} \sum_{i < j} \|\mathbf{e}_i^{(\ell)} - \mathbf{e}_j^{(\ell)}\|_2]^{-1}$ , the inverse mean pairwise distance among example embeddings.
- *Mismatch:*  $d_r^{(\ell)} = 1 - \cos(\mathbf{e}_{\text{instr}}^{(\ell)}, \mathbf{e}_T^{(\ell)})$ , the cosine distance between instruction and example centroids, with  $\mathbf{e}_T^{(\ell)} = \frac{1}{k} \sum_i \mathbf{e}_i^{(\ell)}$ . When helpful, we also report  $d_r^{(\ell,\text{L2})} = \|\mathbf{e}_{\text{instr}}^{(\ell)} - \mathbf{e}_T^{(\ell)}\|_2$  (monotone with the cosine version).

A per-layer anchoring score is defined as  $S^{(\ell)} = \rho_d^{(\ell)} - d_r^{(\ell)} - \log k_{\text{eff}}$ , where  $k_{\text{eff}} = k$  for Few Shot prompts (and 1 for Zero Shot).

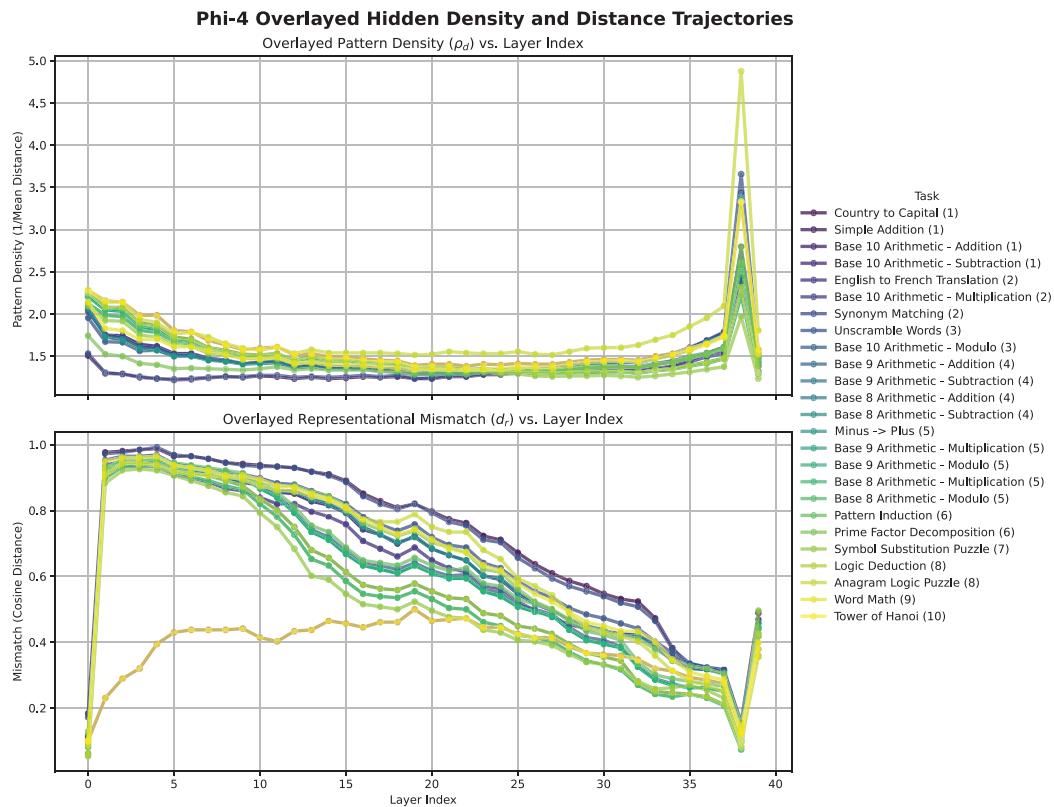
**Note on Intra-Class Variation and Dynamic Similarity (No Experiments Here)** Many concepts tolerate wide intra-class variation: for example, two dog photos can differ in color, posture, background, or lighting, yet still be judged “the same kind of thing.” Human similarity judgments often activate only a subset of *respects* (features) for the specific pair being compared—agreement on *all* features is not required. This suggests augmenting cohesion with a *diversity-within-relevance* axis and a *dynamic* (pairwise, respect-selective) notion of distance. We defer formalizations (e.g., dynamic partial functions [[Li et al. 2003](#)]) and an evaluation plan to Appendix 14.C.

**Findings (Qualitative, Consistent Over Ten Runs)** Across models and tasks we observe:

- *U-shaped  $\rho_d$  trajectories.* Density dips early (specialization and enrichment), recovers mid-layer as  $d_r^{(\ell)}$  falls (abstraction and alignment), and then rises near output (standardization and formatting). This three-stage pattern is consistent with UCCT and recurs across runs.
- *Task difficulty correlates with  $d_r^{(\ell)}$ .* Logic-heavy tasks maintain larger mismatch, suggesting more transformation to align examples with instructions; arithmetic/code tend to align sooner.
- *Model “styles.”* LLaMA-3.1-8B, Phi-4, and Qwen3-14B (Figures 14.4–14.6) share the U-shaped pattern, but differ in alignment layer, path length, and curvature, consistent with different priors/training. These differences are stable across the ten runs.



**Figure 14.4** **Meta-LLaMA-3.1-8B-Instruct.** Layer-wise pattern density  $\rho_d^{(\ell)}$  and mismatch  $d_r^{(\ell)}$  across representative tasks. The U-shape—early enrichment, mid-layer abstraction/alignment, late standardization—recurs across runs.

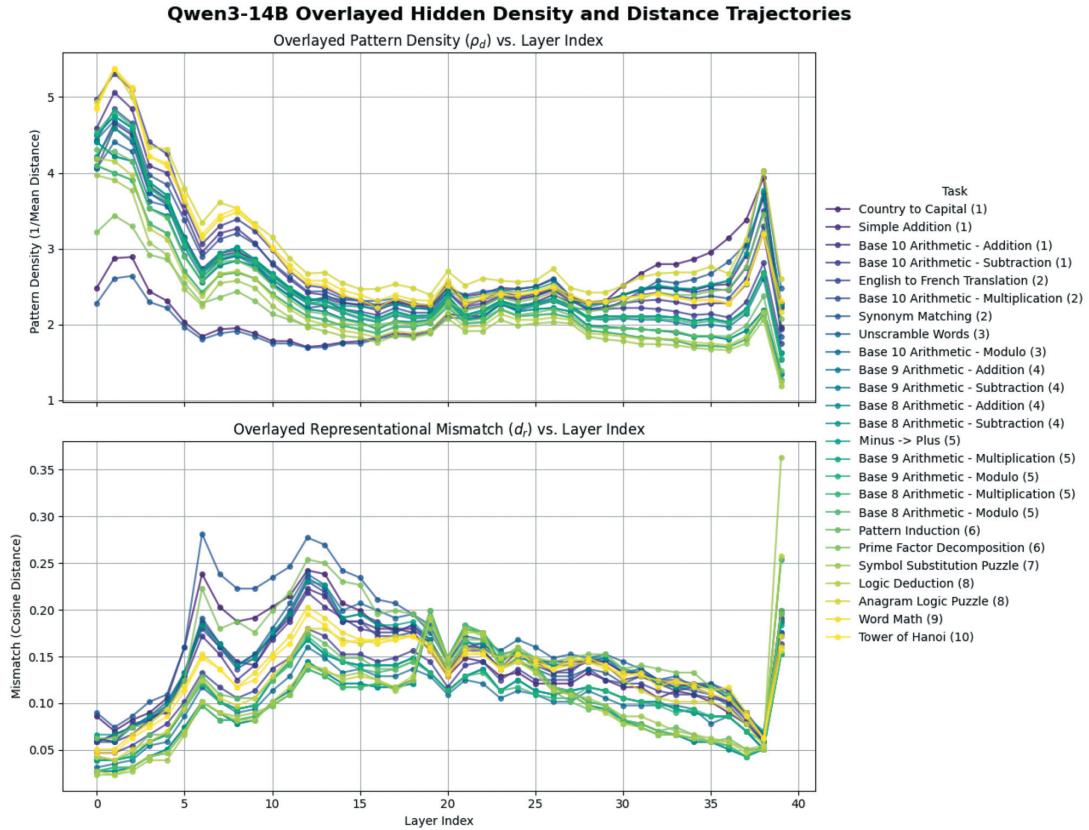


**Figure 14.5** **Phi-4.** Layer-wise pattern density  $\rho_d^{(t)}$  and mismatch  $d_r^{(t)}$  across representative tasks. The same U-shaped trajectory appears, with model-specific alignment depth, consistently across the ten runs.

**Summary** Across Meta-LLaMA-3.1-8B, Phi-4, and Qwen3-14B, the layer-wise overlays reveal a consistent three-stage pattern—early specialization (density drop), mid-layer abstraction/alignment (mismatch dip), and late standardization (re-clustering)—which is *consistent* with UCCT’s enrichment-abstraction-standardization view. Harder, more abstract tasks tend to maintain larger instruction-example mismatch  $d_r^{(t)}$  than concrete tasks, and the layer of strongest alignment varies by model, suggesting distinct model-specific priors or training curricula. These qualitative effects are observed consistently across the ten independent runs per configuration.

#### 14.4.2 Implication for Subsequent Experiments

The qualitative geometry suggests where anchoring tends to consolidate (mid- to late layers) and indicates that alignment depth is dependent on the model. E3



**Figure 14.6** Qwen3-14B. Layer-wise pattern density  $\rho_d^{(t)}$  and mismatch  $d_r^{(t)}$  across representative tasks. The U-shape is present, but trajectory curvature and alignment layer differ, reflecting distinct priors; these differences are stable across runs.

leverages this by testing whether moderated debate can increase effective anchoring strength without excessive context growth, informed by where alignment typically emerges.

#### 14.4.3 Limitations and Future Work

- **Qualitative emphasis.** We report qualitative overlays consistent across ten runs, but we do not quantify a “universality gap” (e.g., Procrustes/ Dynamic Time Warping [DTW] vs. a shuffled-layer null) or provide effect sizes/intervals for cross-model trajectory similarity.
- **Link to outcomes.** We do not correlate geometry deltas (e.g.,  $-\Delta d_r$ ,  $\Delta \rho_d$ ) with task accuracy; establishing predictive validity is deferred.

- **Span/metric robustness.** Results are presented with mean pooling and cosine distance; robustness to last-token/attention-weighted pooling and  $L^2$  distance, as well as span controls (random/nonsense), remains to be documented.
- **Semantic-band identification.** We have not yet localized the “best semantic layer band” via paraphrase sanity checks (paraphrase < random distance) or on/off-task  $d_r$  gaps; these diagnostics will help determine which layers to trust most.
- **Model/task coverage.** Although three models and 25 tasks are included, we do not analyze scale effects or broader model families; extending to additional backbones/scales and reporting seed-variability ribbons per layer are left for future work.
- **Computational resources.** Full statistical aggregation per layer was limited by computational resources; future revisions will add per-layer means/standard deviations (or CIs) and seed ribbons to figures.
- **Dynamic similarity under intra-class variation.** We do not quantify “diversity within relevance” or respect-selective distance in this chapter. Appendix 14.C sketches ideals, formulas (including dynamic partial functions), and a plan to evaluate them.

## 14.5 Conclusion

Unified Cognitive Consciousness Theory (UCCT) explains LLM behavior as semantic anchoring over an unconscious pre-training prior. Rather than intrinsic reasoning, capabilities arise when anchors induce a target posterior  $P_T$  that competes with the prior  $P_{\text{prior}}$ . A single strength index,

$$S = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k,$$

characterizes when anchoring crosses a success threshold and explains why more, longer, or better chosen anchors may help or hurt.

### 14.5.1 What We Established

We turned UCCT into estimators and tests, then validated core predictions in two settings.

*E1, arithmetic with base manipulation.* Holding algorithmic complexity fixed, we varied representational familiarity by changing the numeral base. Few Shot curves show clear thresholds. The fitted shot threshold  $k_{50}$  orders as  $10 < 8 < 9$  and scales with  $d_r/\rho_d$  as predicted. Base 10 reaches a near-Zero or One Shot threshold, while

base 9 needs several shots. The  $-\log k$  term explains why longer anchors do not win trivially and why excessive step-by-step traces can trade in-domain gains for out-of-scope drops. Light fine-tuning shifts  $\rho_d$  and  $d_r$  in the expected directions, with asymmetric cross-base interference consistent with hysteresis.

*E2, geometric trajectories across models and tasks.* Tracking layer-wise embeddings for instruction and examples in Meta-LLaMA-3.1-8B, Phi-4, and Qwen3-14B over 25 tasks, we observe a recurrent three-stage signature: early specialization with a density drop, mid-layer abstraction with a mismatch dip, and late standardization with re-clustering. Harder tasks maintain a larger mismatch. Model-specific trajectory shapes suggest distinct priors. These qualitative effects are stable over ten runs per configuration. We also motivate a diversity-within-relevance extension and outline a dynamic, respect-selective similarity framework for concepts with broad intra-class variation.

### 14.5.2 Outlook

We sketch how to apply UCCT to retrieval and multi-agent debate in separate appendices. For RAG, retrieval should be admitted, ordered, and stopped by expected gains in  $S$  under a token budget. For debate, the same diagnostics used in EVINCE can support a calibrated, budget-aware commit rule without relying on internal tensors. Empirical evaluation of these applications is deferred as future work.

### 14.5.3 Limitations and Future Work

Our studies are controlled and intentionally focused. The key next steps are as follows:

- **Threshold calibration.** Calibrate  $S_c$  on held-out suites, verify the predicted width scaling  $\mathcal{O}(n^{-1/2})$ , and report confidence intervals and effect sizes.
- **Geometry to performance links.** Quantify universality with Procrustes and DTW against shuffled-layer nulls, and correlate late-stage geometry deltas  $(-\Delta d_r, \Delta \rho_d)$  with accuracy.
- **Estimator robustness.** Add per-layer variability bands over seeds; compare pooling schemes (mean, last-token, attention-weighted) and distance metrics (cosine,  $L^2$ ); and include span controls and paraphrase sanity checks to identify the most semantic layer band.
- **Diversity within relevance.** Formalize dynamic partial-function similarity and diversity-aware cohesion in E2, then port these criteria to RAG admission and ordering.

- **Broader coverage.** Extend across backbone families and scales, and include multimodal tasks where intra-class variation is intrinsic.
- **Applications.** Run full evaluations for the RAG and MAD designs with matched budgets, calibrated stopping, and open artifacts for replication.

Taken together, the arithmetic thresholds and geometric trajectories support UCCT as a practical lens for *semantic anchoring* and as a guide to construct, steer, and summarize LLM behavior. The appendices outline how these ideas may transfer to retrieval and debate, setting a clear plan for quantitative tests in future work.

### Acknowledgments

The author thanks Zeyneb N. Kaya and Ethan Y. Chang for their efforts in conducting the experiments. Zeyneb (Stanford) initiated E1 during Q2 2025 while enrolled in CS372 at Stanford University. Ethan (UIUC—University of Illinois Urbana-Champaign) designed and implemented E2 during Summer 2025.

## 14.A

### Appendix A: Threshold-Crossing Dynamics Theorem

We show that semantic anchoring undergoes a sharp selection transition driven by a margin in the anchoring score  $S$ , with a transition width that scales as  $\mathcal{O}(n^{-1/2})$  in the effective evidence size  $n$ . We carry *prior odds* explicitly and use *Bernstein* tails for finite- $n$  constants.

#### 14.A.1 Setup and Assumptions

##### Anchoring score and margin

Throughout the chapter we use

$$S(P) = \rho_d(P) - d_r(P, T) - \log k.$$

For any competitor  $P \neq P^*$  define the *margin*

$$\Delta(P) = \frac{1}{\tau} (S(P^*) - S(P)),$$

where  $\tau > 0$  is a temperature linking  $S$  to expected log-likelihood advantage (A2). The *worst-case margin* is  $\Delta_* = \min_{P \neq P^*} \Delta(P)$ .

##### Evidence-score link and tails

Let  $a_1, \dots, a_n$  be i.i.d. admitted evidence items with likelihoods  $p(a_i|P)$ . For each  $P \neq P^*$  define

$$\ell_i(P) = \log \frac{p(a_i|P^*)}{p(a_i|P)}.$$

Assume

$$\mathbb{E}[\ell_i(P)] = \Delta(P), X_i(P) = \ell_i(P) - \Delta(P)$$

is sub-exponential with parameters  $(\nu^2, b)$ , that is, for all  $t \geq 0$ ,

$$\Pr\left(\sum_{i=1}^n X_i(P) \leq -t\right) \leq \exp\left(-\frac{t^2}{2(n\nu^2 + bt)}\right).$$

This is the quantitative form of the main text's logistic link: the  $S$ -gap determines the expected log-likelihood advantage;  $(\nu^2, b)$  capture noise and bounded increments.

#### Prior odds

Let the *prior log-odds* against each competitor be

$$L_0(P) = \log \frac{p(P^*)}{p(P)}.$$

We keep  $L_0(P)$  finite and explicit (it contributes a finite- $n$  correction).

#### Optional bounded outcome accuracy

There exist  $0 < p_{\max} < p_{\min} \leq 1$  such that

$$p(y^*|P^*, \mathcal{A}_{\text{ev}}) \geq p_{\min}, \quad p(y^*|P, \mathcal{A}_{\text{ev}}) \leq p_{\max} \quad \forall P \neq P^*.$$

This is used only to translate correct selection into task success bounds.

### 14.A.2 Posterior Odds and a Bernstein Bound

The posterior log-odds for  $P^*$  versus  $P$  is

$$\Lambda_n(P) = \log \frac{p(P^*|\mathcal{A}_{\text{ev}})}{p(P|\mathcal{A}_{\text{ev}})} = L_0(P) + \sum_{i=1}^n \ell_i(P) = n\Delta(P) + L_0(P) + \sum_{i=1}^n X_i(P).$$

By Bernstein,

$$\begin{aligned} \Pr(\Lambda_n(P) \leq 0) &= \Pr\left(\sum_{i=1}^n X_i(P) \leq -[n\Delta(P) + L_0(P)]\right) \\ &\leq \exp\left(-\frac{(n\Delta(P) + L_0(P))^2}{2(n\nu^2 + b[n\Delta(P) + L_0(P)])}\right). \end{aligned}$$

Let  $N_P$  be the number of (effective) competing patterns. A union bound yields

$$\Pr(\exists P \neq P^* : \Lambda_n(P) \leq 0) \leq (N_P - 1) \exp\left(-\frac{(n\Delta_* + L_{0,\min})^2}{2(n\nu^2 + b[n\Delta_* + L_{0,\min}])}\right), \quad (14.9)$$

where  $L_{0,\min} = \min_{P \neq P^*} L_0(P)$ .

### 14.A.3 Sharp Transition and Width

#### 14.A.3.1 Supercritical Regime

If  $n\Delta_* + L_{0,\min} \geq n\varepsilon$  (equivalently,  $\Delta_* \geq \varepsilon - \frac{L_{0,\min}}{n}$ ), then

$$\Pr(P^* \text{ wins}) \geq 1 - (N_P - 1) \exp\left(-\frac{n\varepsilon^2}{2(\nu^2 + b\varepsilon)}\right),$$

that is, exponential concentration with an explicit  $(\nu^2, b)$  rate.

#### 14.A.3.2 Subcritical Regime

If for some competitor  $P$ , we have  $n\Delta(P) + L_0(P) \leq -n\varepsilon$ , then the same bound shows exponentially small chance that  $P^*$  wins.

#### 14.A.3.3 Transition Width

Fix failure budget  $\eta \in (0, 1)$  and precision  $\delta \in (0, 1)$ . Reversing Equation (14.9), it suffices that

$$n\Delta_* + L_{0,\min} \geq \sqrt{2n\nu^2 \log \frac{N_P - 1}{\eta\delta}} + b \log \frac{N_P - 1}{\eta\delta}.$$

Equivalently, the *effective margin*

$$\Delta_{\text{eff}} = \Delta_* + \frac{L_{0,\min}}{n}$$

must satisfy

$$\Delta_{\text{eff}} \geq \sqrt{\frac{2\nu^2}{n} \log \frac{N_P - 1}{\eta\delta}} + \frac{b}{n} \log \frac{N_P - 1}{\eta\delta} = \mathcal{O}(n^{-1/2}).$$

Thus the selection transition occurs within a band of width  $\Theta(n^{-1/2})$  in  $\Delta_{\text{eff}}$ , hence in  $S$ .

### 14.A.4 From Selection to Task Success

Under A4,

$$p_{\max} \leq \Pr(\text{success}) \leq p_{\min},$$

and in the supercritical regime,

$$\Pr(\text{success}) \geq p_{\min} - (N_P - 1) \exp\left(-\frac{n\varepsilon^2}{2(\nu^2 + b\varepsilon)}\right),$$

while in the subcritical regime  $\Pr(\text{success}) \leq p_{\max}$  up to the same exponentially small term.

### 14.A.5 Identification with E1 Logistic Calibration

In E1, we fit a single logistic

$$\Pr(\text{success}) \approx \sigma(\hat{\alpha}(S - s_0)).$$

The link in A2 identifies the *units*:  $\hat{\alpha} \approx 1/\tau$  (slope), and  $s_0$  corresponds to the finite- $n$  threshold where the worst-case competitor balances  $P^*$ :

$$s_0(n) \approx \max_{P \neq P^*} \left\{ S(P) + \tau \frac{L_0(P)}{n} \right\}.$$

As  $n$  grows,  $s_0(n)$  approaches  $\max_{P \neq P^*} S(P)$ , while the observed transition width shrinks as  $\mathcal{O}(n^{-1/2})$ , consistent with the empirical phase-width fits in E1.

### 14.A.6 Conclusion

Posterior selection is controlled by a temperature-scaled  $S$  margin with explicit prior-odds and Bernstein constants. Supercritical and subcritical regimes exhibit exponential concentration and anti-concentration, and the selection transition band has width  $\Theta(n^{-1/2})$ . These statements are agnostic to the anchoring mechanism and align with the logistic calibrations used in the arithmetic study.

## 14.B

### 14.B.1

### Appendix B: Threshold Calibration

#### Calibrating the Critical Level $S_c$ (Proposal)

Let  $\mathcal{D} = \{(S_i, y_i)\}_{i=1}^N$  be held-out trials with anchoring strengths  $S_i$  and success labels  $y_i \in \{0, 1\}$ . We estimate  $S_c$  with complementary parametric and nonparametric procedures, and report stability via cross-validation and bootstrap.

*Parametric (sigmoid MLE [Maximum Likelihood Estimation]).* Fit the logistic model in Equation (14.4) by maximizing

$$\ell(\alpha, s_0) = \sum_{i=1}^N [y_i \log \sigma(\alpha(S_i - s_0)) + (1 - y_i) \log(1 - \sigma(\alpha(S_i - s_0)))],$$

where  $\sigma(t) = 1/(1 + e^{-t})$ . The estimated threshold is  $\hat{S}_c^{\text{MLE}} = \hat{s}_0$ ; the slope parameter  $\hat{\alpha}$  encodes transition sharpness. Uncertainty is reported via nonparametric bootstrap  $(\hat{S}_c^{\text{MLE}})_{\text{boot}}$  with percentile intervals.

*Nonparametric (isotonic calibration).* Fit a monotone map  $\hat{p}(S)$  by isotonic regression on  $\mathcal{D}$ . Define empirical baselines at the extremes

$$\hat{p}_{\text{low}} = \hat{p}(\mathbf{q}_{0.05}(S)), \quad \hat{p}_{\text{high}} = \hat{p}(\mathbf{q}_{0.95}(S)),$$

and set the operating midpoint  $\tau = (\hat{p}_{\text{low}} + \hat{p}_{\text{high}})/2$ . The threshold is

$$\hat{S}_c^{\text{iso}} = \arg \min_s |\hat{p}(s) - \tau|.$$

This guards against class imbalance and tasks where “chance” and “optimal” differ from  $(0, 1)$ .

*Decision-theoretic refinement.* When costs differ, pick  $\hat{S}_c$  to minimize expected cost or to maximize Youden’s index  $J = \text{TPR} - \text{FPR}$  along the ROC curve induced by  $S$ :

$$\hat{S}_c^{\text{ROC}} = \arg \max_s (\text{TPR}(s) - \text{FPR}(s)).$$

*Layer selection and pooling.* Compute  $S^{(\ell)}$  per layer  $\ell$  and choose either (i) a single operational layer  $\ell^*$  that maximizes validation AUC, estimating  $S_c$  on  $S^{(\ell^*)}$ , or (ii) an ensemble score  $\bar{S} = \sum_{\ell} w_{\ell} S^{(\ell)}$ , with  $w_{\ell} \geq 0$  and  $\sum_{\ell} w_{\ell} = 1$ , tuned on validation via logistic loss.

### 14.B.2 Validation Protocol

Use grouped  $K$ -fold Cross-Validation (CV) that holds out entire (task, model) groups. Report  $\hat{S}_c$  (mean and 95% CI), calibration (ECE, Brier), discrimination (AUC), and observed transition width (10%–90% success range). In practice, we adopt the sigmoid MLE for a scalar point estimate ( $\hat{S}_c = \hat{s}_0$ ) and publish isotonic and ROC thresholds as sensitivity analyses.

## 14.C 14.C.1

### Appendix C: Diversity Within Relevance

#### Motivation

Humans compare objects by activating only some *respects* of similarity for the pair at hand. Two dog photos may share ear/pose/texture cues but differ in color or background; both are “dogs” despite large superficial differences. Likewise, the question “Which cities are similar to Los Angeles?” admits multiple, pair-dependent answers (e.g., LA ~ SF via state/size; LA ~ Tokyo via coastal megacity logistics). We therefore explore metrics that (i) keep examples *relevant* to the instruction, and (ii) reward *non-redundant coverage* of valid variations by measuring similarity over *activated respects* rather than all features.

### 14.C.2 Design Ideas

A useful extension should: (1) be *relevance-gated* (on-topic); (2) reward *non-redundancy* (coverage of legitimate variations); (3) work with small  $k$ ; (4) be *layer-stable* in the semantic band; (5) be computationally simple.

### 14.C.3 Candidate Formulations (To Be Evaluated Later)

- (1) **Coverage score (facility-location style).** On the relevance-gated set, define a coverage term that rises when chosen examples “cover” the remaining relevant embeddings; combine this with  $\rho_d$  to form a coverage-adjusted density  $\tilde{\rho}_d$ .
- (2) **Dynamic Partial Functions (DPF) [Li et al. 2003]** Define a pairwise distance that selects the *closest* feature coordinates for the pair (the activated respects) and aggregates only over those coordinates. Use DPF both for example-example cohesion (tolerant to benign variation) and for instruction-set mismatch via a mode-aware, DPF-based distance.

### 14.C.4 Planned Evaluation (Deferred)

- (1) Identify a “semantic” layer band via a paraphrase sanity check (paraphrase < random).
- (2) Compare selection policies (random- $k$ , cohesion-only, coverage-greedy, DPF-greedy) under equal  $k$ .
- (3) Report  $\rho_d$ ,  $\tilde{\rho}_d$ ,  $d_r$ , mode-aware  $d_r$ , and per-layer  $S$  vs.  $S_{\text{cov}}$ ; include relevance and coverage components.
- (4) Check that coverage-aware scores help when intra-class variation is large, and that all methods agree on single-variant controls.
- (5) Check robustness: Assess sensitivity to pooling choices (mean/attention/last-token), gates, and hyperparameters.

### 14.C.5 Scope

These are prospective additions; E2’s results remain unchanged in this chapter.

## 14.D Appendix D: UCCT for RAG (Protocol and Lab Exercise Plan)

### 14.D.1 Scope

This appendix sketches how to apply UCCT to RAG. It specifies estimators and control rules consistent with Section 14.3.2. Empirical evaluation is deferred to future work.

### 14.D.2 UCCT View of RAG

#### 14.D.2.1 Setup

A retrieval call proposes a multiset of candidate evidence spans  $\mathcal{E} = \{e_i\}$  for a query  $Q$  and target task  $T$ . Any admitted subset  $E \subseteq \mathcal{E}$  acts as anchors that induce an

anchoring strength

$$S(Q, E) = \rho_d(E) - d_r(Q, E) - \log k(E),$$

where  $\rho_d(E)$  measures useful support around the target,  $d_r(Q, E)$  measures the mismatch between the query or target prototype and the admitted evidence, and  $k(E)$  counts the tokens placed in context (including formatting/metadata).

### 14.D.3 Notation and Estimators

#### 14.D.3.1 Encoders

Let  $f(\cdot)$  be a fixed text encoder used only for diagnostics and selection, and let  $z_i = f(e_i)$  be span embeddings. For the query side, we use either a target prototype  $\mu_A$  when anchors or a schema exist, or a query proxy  $q = f(Q)$ .

#### 14.D.3.2 Quality Rubric (CRIT)

Following Socratic Critical Reading [Chang 2023a], an *external* judge  $J$  (a held-out LLM or rubric-based grader, distinct from the generator) scores each retrieved span  $e_i$  with  $\text{CRIT}(e_i) \in [0, 1]$  covering factual grounding, on-task relevance to  $Q/T$ , specificity, and internal consistency. To discourage verbosity gaming, scores are normalized per 100 tokens. We aggregate

$$\text{CRIT}(E) = \frac{1}{|E|} \sum_{e_i \in E} \text{CRIT}(e_i),$$

and use it in admission/ordering and in the density estimator in Section 14.D.3.3. The judge  $J$  is frozen across cases to avoid leakage and drift.

#### 14.D.3.3 Quality and Cohesion

Evidence that is “dense” for the target should be high quality and mutually supportive without redundancy. We combine three ingredients (each normalized to  $[0, 1]$  on a dev set):

- **Judge quality**  $\text{CRIT}(E)$  (defined in the previous section).
- **Cohesion**

$$\text{COH}(E) = \left[ \binom{|E|}{2}^{-1} \sum_{i < j} \|z_i - z_j\|_2 \right]^{-1}.$$

- **Redundancy penalty**

$$\text{RED}(E) = \frac{1}{|E|} \sum_i \max_{j \neq i} \cos(z_i, z_j).$$

We define target-relevant density as

$$\rho_d(E) = w_1 \text{CRIT}(E) + w_2 \text{COH}(E) - w_3 \text{RED}(E),$$

$$w_1, w_2, w_3 \geq 0, \quad w_1 + w_2 + w_3 = 1,$$

with weights calibrated on a small development suite.

#### **14.D.3.4 Mismatch**

When a target prototype  $\mu_A$  is available,

$$d_r(Q, E) = 1 - \cos\left(\mu_A, \frac{1}{|E|} \sum_{e_i \in E} z_i\right).$$

If no prototype exists, use the query proxy  $q = f(Q)$  in place of  $\mu_A$ . For added robustness, a contrast margin using a background prototype  $\mu_{\neg A}$  can be computed and folded into  $\rho_d$  as a selectivity bonus:

$$\sigma(E) = \frac{1}{|E|} \sum_{e_i \in E} [\cos(z_i, \mu_A) - \cos(z_i, \mu_{\neg A})].$$

#### **14.D.3.5 Budget**

$k(E)$  is the total token count added to the generation context by  $E$  and its formatting. Retrieval and scoring tokens that do *not* enter the generation context are not charged to  $k(E)$ .

### **14.D.4 Control: Admission, Ordering, and Stopping**

#### **14.D.4.1 Admission**

For each candidate  $e_i$ , compute the marginal gain

$$\Delta S_i = \Delta \rho_d - \Delta d_r - \Delta \log k,$$

where deltas are measured relative to the current admitted set. Admit only if  $\Delta S_i > 0$  plus a small margin.

#### **14.D.4.2 Selection and Ordering Under a Budget**

Maximize  $\sum_{e_i \in E} \Delta S_i$  subject to a token budget  $\sum \text{tokens}(e_i) \leq B$ . In practice, use greedy selection by  $\Delta S_i$  per token and place higher-gain items earlier to reduce interference.

#### 14.D.4.3 Stopping

Stop when  $S(Q, E_t)$  plateaus for two additions or when it crosses a calibrated threshold  $\hat{S}_c$  fitted on held-out tasks. The plateau rule avoids overfilling; the threshold helps avoid stopping too early on easy cases.

#### 14.D.5 Diversity Within Relevance

Some targets benefit from coverage across distinct “respects” while staying relevant. Two lightweight options are as follows:

- **DPF-inspired respect selection** [Li et al. 2003]: compute distances with respect to a subset of features activated by the current query and provisional evidence, then re-evaluate COH and RED in that respect to reward non-duplicative coverage.
- **Submodular or MMR objective**: add a diversity term to the selection score and optimize with a greedy submodular routine under the same token budget.

#### 14.D.6 Practical Procedure

- (1) Build a prototype  $\mu_A$  from a small anchor set or a task schema; otherwise use  $q = f(Q)$ .
- (2) Retrieve a candidate pool  $\mathcal{C}$  with the usual retriever; encode  $z_i = f(e_i)$ .
- (3) Score each  $e_i$  with the independent judge to obtain  $\text{crit}_j(e_i)$ .
- (4) Run admission and greedy selection by  $\Delta S_i$  per token, with diversity enabled if needed.
- (5) Order admitted items by gain and stop by plateau or by crossing  $\hat{S}_c$ .

#### 14.D.7 Evaluation Plan (Deferred to Companion Lab Exercise)

We sketch an empirical study plan; full evaluation is deferred to the companion lab exercise (and constitutes suitable future work in research settings).

##### 14.D.7.1 Datasets

Open RAG QA and fact-checking suites with gold evidence, plus domain-shifted variants.

##### 14.D.7.2 Baselines

No retrieval; top- $k$  BM25 or DPR; reranking; chain-of-verification pipelines.

**14.D.7.3 Metrics**

Accuracy, F1 or EM (exact match), calibration (ECE, Brier), tokens per correct answer, overconfidence rate, and ablations for admission, ordering, and stopping.

**14.D.7.4 Ablations**

Random versus  $\Delta S$  per token selection, with or without diversity, fixed- $k$  versus  $S$ -plateau stopping.

**14.D.8 Notes and Limitations**

- **Independence of judging.** CRIT<sub>J</sub> is scored by an *independent* LLM judge, not by the generator. This separation reduces self-reinforcing bias.
- **Calibration.**  $\hat{S}_c$  and weights ( $w_1, w_2, w_3$ ) would be calibrated on a small dev suite, then fixed. We do not report fitted values here.
- **Encoder choice.** Results may vary with  $f(\cdot)$ . Robustness checks across encoders are recommended.
- **Compute.** The selection pipeline adds overhead, but only admitted spans are charged to  $k(E)$ . Engineering the judge and encoder for batching keeps costs manageable.

**14.D.9 Takeaway**

RAG helps when it *densifies the correct region of posterior space* and reduces mismatch without excessive context growth. The score  $S(Q, E)$  provides a target-aware filter that predicts usefulness before generation and guides admission, ordering, and stopping.

# **14.E Appendix E: UCCT for Multi-Agent Debate (Protocol and Lab Exercise Plan)**

**14.E.1 Scope and Intent**

This appendix specifies a UCCT-aligned protocol for MAD. We do not report new experiments here. Instead, we define diagnostics, a calibrated control rule, and decision criteria, and we walk through a clinical vignette to illustrate their interactions. Empirical evaluation is deferred to future work.

**14.E.2 Design Goal**

Retain EVINCE's diagnostics and phase scheduling, but replace the plateau-only stop rule with a calibrated, budget-aware decision (commit/continue/abstain) under a token cap. Concretely, at round  $t$ , let agents produce posteriors  $P_t^{(+)}$

and  $P_t^{(-)}$  over a shared hypothesis set. We track *mismatch* (JS [Jensen–Shannon divergence], WD), *information sharing* (MI), and *quality* (CRIT), all defined in Section 14.E.1. A global cap  $K_{\max}$  limits total debate and diagnostic tokens, and the running budget is  $k_t$ .

The prior plateau rule was: “no relative improvement  $> \eta$  in both mismatch and quality for two consecutive rounds.” We replace it with a calibrated probability gate

$$\hat{p}_t = \Pr(\text{correct} | \text{JS}_t, \text{WD}_t, \text{MI}_t, \text{CRIT}_t, \log k_t, \Delta \text{trends}_t) :$$

*commit* if plateau and  $\hat{p}_t \geq \tau^*$ ; *abstain* if plateau and  $\hat{p}_t < \tau^*$ ; otherwise *continue* (subject to  $K_{\max}$ ). Phase scheduling (contentiousness  $\Delta$ ) follows *explore*  $\rightarrow$  *consolidate*  $\rightarrow$  *reconcile*, decreasing  $\Delta$  as JS/WD fall and MI/CRIT rise.

### 14.E.3 Diagnostics and Notation

At round  $t$ , each agent  $a \in \{+, -\}$  produces a distribution  $P_t^{(a)}$  over a harmonized hypothesis set  $\mathcal{H}$ . Let  $E_{1:t}$  be the set of *admitted* evidence up to round  $t$ .

- **Mismatch (distance).**  $\text{JS}_t$  and  $\text{WD}_t$  with ground cost  $c(h_i, h_j) = 1 - \cos(\phi(h_i), \phi(h_j))$  for a fixed encoder  $\phi$ . Use bootstrap CIs over posterior samples to reduce estimator noise.
- **Information sharing.** Conditional mutual information

$$\text{MI}_t = I(H_t^{(+)}, H_t^{(-)} | E_{1:t}),$$

where  $H_t^{(\pm)}$  have marginals  $P_t^{(\pm)}$ . Estimate via a plug-in estimator on aligned discrete support with additive smoothing ( $\alpha = 1$ ); optionally report  $\text{NMI}_t = \text{MI}_t / \min\{H(H_t^{(+)})_t, H(H_t^{(-)})_t\}$ .

- **Quality (external judge).**  $\text{CRIT}_t \in [0, 1]$  scored by an *independent* judge  $J$  (held-out LLM or rubric grader) using a fixed prompt and decoding.<sup>2</sup> Rubric: factual grounding, on-task relevance, specificity, internal consistency. Normalize per token (score per 100 tokens); repeated claims receive no extra credit.
- **Cohesion (diversity-aware).**  $\text{COH}_t$  measures the semantic cohesion of accepted evidence  $E_{1:t}$  (e.g., inverse mean pairwise distance) with a redundancy penalty so that near-duplicates do not inflate density. A submodular cohesion–diversity objective is optional.

---

2. Judge  $J$  is frozen across cases; agents do not see  $J$ 's prompt or scores. If multiple judges are used, report inter-rater agreement.

- **Budget.**  $k_t$  is the cumulative token budget, including debate and diagnostics (posterior sampling for MI/WD/JS and calls to  $J$ ). Enforce a global cap  $K_{\max}$  with matched decoding settings across methods.

For interpretive analysis only, define  $S_t = f_\rho(\text{CRIT}_t, \text{COH}_t, \text{MI}_t) - f_d(\text{JS}_t, \text{WD}_t) - \log k_t$ , with  $f_\rho, f_d$  set by min–max normalization on validation. We do not threshold or optimize  $S_t$ .

For interpretive analysis only, we define

$$S_t = f_\rho(\text{CRIT}_t, \text{COH}_t, \text{MI}_t) - f_d(\text{JS}_t, \text{WD}_t) - \log k_t,$$

with  $f_\rho, f_d$  set by min–max normalization on validation. We do not threshold or optimize  $S_t$ .

#### **14.E.4 Calibrated Control (No Hard $S_c$ )**

We learn on validation a calibrated success probability

$$\hat{p}_t = \Pr(\text{correct} \mid \text{MI}_t, \text{WD}_t, \text{JS}_t, \text{CRIT}_t, \text{COH}_t, \log k_t, \Delta\text{trends}_t),$$

using logistic or isotonic regression, then freeze a decision threshold  $\tau^*$ .  $\Delta\text{trends}_t$  includes short-window slopes for JS, WD, MI, and CRIT, entropy of  $P_t^{(\cup)}$ , and a simple budget headroom feature. Thresholds are calibrated with time-split or nested CV; calibration quality is reported via ECE and Brier.

##### **14.E.4.1 Plateau Detection**

We detect a plateau when trend tests on both mismatch (JS or WD) and quality (CRIT) show no improvement beyond a tolerance  $\eta$  over two rounds. A Mann-Kendall or bootstrap slope test is sufficient for this design sketch.

##### **14.E.4.2 Hybrid Stopping**

- (1) *Commit* if a plateau holds and  $\hat{p}_t \geq \tau^*$ .
- (2) *Abstain* if a plateau holds and  $\hat{p}_t < \tau^*$ . Abstain triggers escalation: request specific tests, retrieve external evidence, or defer.
- (3) *Continue* otherwise, subject to the token cap.

#### **14.E.5 Phase-Guided Debate**

**Exploration (high contentiousness).** Encourage stance diversity and hypothesis coverage. Expect WD and JS high, MI low, CRIT moderate. Risk is orthogonal argumentation; benefit is surfacing missing evidence.

**Consolidation (moderate).** Prompt cross-examination and error identification. Target WD and JS down, MI and CRIT up. Risk is premature agreement with weak evidence, mitigated by the probability gate  $\hat{p}_t$ .

**Reconciliation (low).** Force synthesis and joint justification. Apply the hybrid stop rule (commit, continue, or abstain).

### 14.E.6 Arbitration

We use a quality- and alignment-weighted log-opinion pool:

$$P^{(\text{arb})}(h) \propto \left(P_T^{(+)}(h)\right)^{\lambda_+} \left(P_T^{(-)}(h)\right)^{\lambda_-}, \lambda_{\pm} \propto \text{CRIT}_T^{(\pm)} \cdot \exp\left(-\text{KL}\left(P_T^{(\pm)} \| P_T^{(\cup)}\right)\right),$$

with  $\lambda_+ + \lambda_- = 1$  after normalization. This favors agents that contributed high-quality, shared information and are close to the consensus.

### 14.E.7 Stance-Based Anchors and Label Harmonization

#### 14.E.7.1 Anchors

Agents operate under complementary instructions to induce structured opposition while staying evidence-focused:

- $\mathcal{A}^{(+)}$  : “Argue for hypothesis  $X$  with supporting evidence”.
- $\mathcal{A}^{(-)}$  : “Challenge hypothesis  $X$  and propose alternatives”.

#### 14.E.7.2 Harmonization

Agents may start with different label granularities. We harmonize to a shared  $\mathcal{H}$  by mapping coarse labels to a distribution over fine hypotheses via a fixed prior table or an embedding-based soft mapping  $\pi(h|\text{coarse}) \propto \exp(\cos(\phi(h), \phi(\text{coarse})))$ . WD, JS, and MI are computed on the aligned supports. We log  $\pi$  for transparency and include a label-shuffle null to show that the observed distances are not mapping artifacts.

### 14.E.8 Illustrative Example: Dengue vs. Chikungunya

#### 14.E.8.1 Case (Gold Label: Dengue Fever)

A patient presents with fever, rash, severe joint and muscle pain, pain behind the eyes, malaise, and red spots. Appendix 7.D in Chapter 7 gives the moderator prompt and raw agent rounds.

**Round 1 (exploration).** Agent + (GPT-4): {Dengue 0.60, Chikungunya 0.25, Zika 0.15}, with specific findings (petechiae, thrombocytopenia). Agent - (Gemini): {Viral 0.60, Autoimmune 0.20, Bacterial 0.15}. Harmonize coarse labels to {Dengue, Chikungunya, Zika, Other}. Diagnostics:  $\text{WD}_1, \text{JS}_1$  high;  $\text{MI}_1$  low;  $\text{CRIT}_1$  mixed;  $\text{COH}_1$  moderate. Continue.

**Round 2 (consolidation).** Cross-examination elicits vector-borne clues and CBC patterns. Agent – updates to {Dengue 0.50, Chikungunya 0.30, Viral 0.20}. Diagnostics:  $WD_2, JS_2$  down;  $MI_2$  up;  $CRIT_2$  up;  $COH_2$  stable or up. If  $\hat{p}_2$  is high, prepare to reconcile next; otherwise, continue.

**Round 3 (reconciliation).** Convergence near {Dengue  $\approx 0.60$ , Chikungunya  $\approx 0.35$ , Zika  $\approx 0.05$ }, plus joint tests (CBC, NS1, PCR). Diagnostics:  $WD_3, JS_3$  low;  $MI_3$  high;  $CRIT_3$  high;  $COH_3$  high. If a plateau is detected and  $\hat{p}_3 \geq \tau^*$ , commit; if plateau but  $\hat{p}_3 < \tau^*$ , abstain.

#### 14.E.8.2 Trade-Offs

Longer debates can inflate tokens with diminishing returns. Plateaus with weak CRIT are risky. Early commit is justified when mismatch falls and quality rises quickly.

### 14.E.9 MAD Experiment Plan (Deferred to Companion Lab Exercise)

We sketch an empirical study plan; full evaluation is deferred to the companion lab exercise (and is suitable future work in research settings).

#### 14.E.9.1 Primary Tests

- (i) Accuracy and calibration at matched tokens versus plateau-only.
- (ii) Early predictive value of  $\hat{p}_2$  versus round count, context length, union entropy.
- (iii) Nulls/stress tests: label-shuffle, role-swap, adversarial agent ( $MI \uparrow, CRIT \downarrow$ ).
- (iv) Encoder sensitivity for  $\phi$ .
- (v) Budget sensitivity.

#### 14.E.9.2 Evaluation Framework

**Benchmarks** Clinical diagnosis; science QA with gold justifications.

**Baselines** Single-agent CoT; self-consistency; fixed-contentiousness debate.

**Metrics** Accuracy; ECE; Brier; tokens-per-correct; confidently-wrong rate; JS/WD/MI trajectories (bootstrap CIs).

**Ablations** Plateau-only versus calibrated stopping;  $\Delta S_t > 0$  admission; random versus scheduled contentiousness; role-conditioned versus neutral; no-diagnostics.

#### 14.E.9.3 Cost, Reproducibility, and Ethics

All debate and diagnostic activities are charged to  $k_t$ , including posterior sampling (MI/WD/JS) and external-judge calls for CRIT. Budgets and decoding settings are

matched across methods; temperatures, top- $p$ , and seeds are fixed, with variance reported over seeds. “Abstain” semantics are explicit (escalate to additional tests/retrieval or defer). This protocol is for research only and is not a clinical decision aid.

## References

- E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. 2022. What learning algorithm is in-context learning? Investigations with linear models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS '22)*, New Orleans, LA, November 28–December 9, 2022. DOI: <https://doi.org/10.48550/arXiv.2211.15661>.
- A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique for language models. In *Proceedings of the International Conference on Learning Representations (ICLR '24)*, Vienna, Austria, May 7–11, 2024. DOI: <https://doi.org/10.48550/arXiv.2310.11511>.
- S. Borgeaud, A. Mensch, J. Hoffmann, et al. 2022. Improving language models by retrieving from trillions of tokens. In K. Chaudhuri (Eds.), *Proceedings of the 39th International Conference on Machine Learning, Volume 162*, Baltimore, ML, July 17–23, 2022. PMLR, 2206–2240.
- M. Cemri, M. Z. Pan, S. Yang, et al. 2025. Why do multi-agent LLM systems fail? DOI: <https://doi.org/10.48550/arXiv.2503.13657>.
- B. J. Chan, C.-T. Chen, J.-H. Cheng, and H.-H. Huang. 2024. Don’t do RAG: When cache-augmented generation is all you need for knowledge tasks. In *Proceedings of the Companion ACM on Web Conference 2025 (WWW '25)*, Sydney, NSW, Australia, April 28–May 2, 2025. ACM, New York, NY, 893–897. DOI: <https://doi.org/10.1145/3701716.3715490>.
- E. Y. Chang. March. 2023a. CRIT: Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Conference of the Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. December. 2023b. Examining GPT-4’s capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang. August. 2024. EVINCE: Optimizing adversarial LLM dialogues via conditional statistics and information theory. DOI: <https://doi.org/10.48550/arXiv.2408.14575>.
- P. Clark, I. Cowhey, O. Etzioni, et al. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. DOI: <https://doi.org/10.48550/arXiv.1803.05457>.
- K. Cobbe, V. Kosaraju, M. Bavarian, et al. 2021. Training verifiers to solve math word problems. DOI: <https://doi.org/10.48550/arXiv.2110.14168>.
- S. Es, J. James, L. E. Anke, and S. Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In N. Aletras and O. De Clercq (Eds.), *Proceedings of the 18th Conference of the European Chapter of the ACL: System Demonstrations*, St. Julian’s,

- Malta, March 17–22, 2024. ACL, 2024, 150–158. DOI: <https://doi.org/10.18653/v1/2024.eacl-demo.16>.
- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. 2017. On calibration of modern neural networks. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning – Volume 70 (ICML '17)*, Sydney, NSW, Australia, August 6–11, 2017. JMLR.org, 1321–1330.
- K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. 2020. REALM: Retrieval-augmented language model pre-training. In H. Daumé and A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning – Volume 119 (ICML '20)*, July 13–18, 2020. JMLR.org, 3929–3938.
- D. Hendrycks, S. Basart, S. Kadavath, et al. 2021a. Measuring coding challenge competence with APPS. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS '21)*, December 6–14, 2021, Curran Associates, Red Hook, NY. DOI: <https://doi.org/10.48550/arXiv.2105.09938>.
- D. Hendrycks, C. Burns, S. Kadavath, et al. 2021b. Measuring mathematical problem solving with the MATH dataset. DOI: <https://doi.org/10.48550/arXiv.2103.03874>.
- O. Huly, D. Carmel, and O. Kurland. 2025. Predicting RAG performance for text completion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, Padua, Italy, July 13–18, 2025. ACM, New York, NY, 1283–1293. DOI: <https://doi.org/10.1145/3726302.3730062>.
- G. Irving, P. Christiano, and D. Amodei. 2018. AI safety via debate. DOI: <https://doi.org/10.48550/arXiv.1805.00899>.
- V. Karpukhin, B. Oguz, S. Min, et al. 2020. Dense passage retrieval for open-domain question answering. In B. Webber, T. Cohn, Y. He, and Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*, November 16–20, 2020. ACL, 6769–6781. DOI: <https://doi.org/10.18653/v1/2020.emnlp-main.550>.
- U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*. DOI: <https://doi.org/10.48550/arXiv.1911.00172>.
- O. Khattab and M. Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, Virtual Event, China, July 25–30, 2020. ACM, New York, NY, 39–48. DOI: <https://doi.org/10.1145/3397271.3401075>.
- P. Lewis, E. Perez, A. Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Vancouver, BC, Canada, December 6–12, 2020. Curran Associates Inc., Red Hook, NY, 9459–9474.
- B. Li, E. Chang, and Y. Wu. 2003. Discovery of a perceptual distance function for measuring image similarity. *Multimed. Syst.* 8, 6, 512–522. DOI: <https://doi.org/10.1007/s00530-002-0069-9>.

- Q. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In C. Bessiere (Ed.), *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI '20)*, Yokohama, Japan, January 7–15, 2021. 3622–3628.
- N. F. Liu, K. Lin, J. Hewitt, et al. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist* 12, 157–173. DOI: [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638).
- S. Min, X. Lyn, A. Holtzman, et al. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 9174–9189.
- L. Shen, A. Mishra, and D. Khashabi. 2024. Position: Do pretrained transformers learn in-context by gradient descent? In *41st International Conference on Machine Learning (ICML' 24)*, Volume 235, Vienna, Austria, July 21–27, 2024. PMLR.
- W. Shi, S. Min, M. Yasunaga, et al. 2024. REPLUG: Retrieval-augmented black-box language models. In K. Duh, H. Gomez, and S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '24)*, Mexico, June 16–21, 2024. ACL, 8371–8384. DOI: <https://doi.org/10.18653/v1/2024.nacl-long.463>.
- A. Talmor, J. Herzig, N. Lourie, and J. Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In J. Burstein, C. Doran, and T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '19)*. Minneapolis, MN, June 2–7, 2019. ACL, 4149–4158. DOI: <https://doi.org/10.18653/v1/N19-1421>.
- J. von Oswald, E. Niklasson, E. Randazzo, et al. 2023. Transformers learn in-context by gradient descent. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning, Volume 202 (ICML '23)*, Honolulu, Hawaii, July 23–29, 2023. JMLR.org, 35151–35174.
- X. Wang, J. Wei, D. Schuurmans, et al. 2023. Self-consistency improves chain-of-thought reasoning in language models. DOI: <https://doi.org/10.48550/arXiv.2203.11171>.
- J. Wei, X. Wang, D. Schuurmans, et al. 2022. Chain of thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans, LA, November 28–December 9, 2022. Curran Associates Inc., Red Hook, NY, 24824–24837.
- Z. Zhang, M. Fang, and L. Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Proceedings of the Findings of the Association for Computational Linguistics (ACL '24)*, Bangkok, Thailand, August 11–16, 2024. ACL, 6963–6975. DOI: <https://doi.org/10.18653/v1/2024.findings-acl.415>.

# CHAPTER 15

---

## A Retrospective and Adaptive Framework to Improve LLMs

### Abstract

RAFEL is a retrospective and adaptive framework designed to benchmark private Large Language Models (LLMs) against teacher LLMs, with the goal of identifying discrepancies in their responses. Following the initial benchmarking, RAFEL categorizes these discrepancies into four distinct categories, based on cognitive levels and types of errors. Subsequent phases involve a detailed diagnosis and deep-probing analysis to uncover the root causes behind each category of discrepancy. Teacher LLMs play a crucial role in interrogating the private LLM, shedding light on the subtleties of its performance issues. With a clear understanding of the symptoms and their underlying causes, RAFEL prescribes targeted remedies, accompanied by recommendations for relevant data sources to enhance the private LLM's performance via fine-tuning, Retrieval-Augmented Generation (RAG), or a combination of both. Empirical studies validate RAFEL's effectiveness in diagnosing and enhancing the capabilities of localized LLMs.

### 15.1

#### Introduction

The emergence of Large Language Models (LLMs) like GPT and Gemini [Gemini Team Google et al. 2023] has significantly advanced the field of natural language processing, enabling the generation of text that closely mimics human writing and offers deep insights across varied domains. Despite their transformative potential, the deployment and scalability of these models pose considerable computational and data-related challenges. A practical response has been the fine-tuning of medium-sized, open-source models such as LLaMa [Touvron et al. 2023] for specialized needs, allowing organizations to strike a balance between performance

and feasibility, while also prioritizing data privacy and model customization for unique applications.

The shift toward using privately fine-tuned or locally deployed LLMs brings about essential management and technical challenges, vital for corporate strategy, governance, and innovation. This chapter explores the technical challenges of this shift, including:

- Justifying the choice of private LLMs over their public counterparts by establishing relevant performance metrics and benchmarks for these specialized models.
- Conducting in-depth error analysis to pinpoint the root causes of performance issues in private LLMs, ensuring targeted and effective remediation strategies.
- Identifying specific, high-quality data crucial for the fine-tuning of private LLMs, aimed at enhancing their accuracy and domain relevance.
- Implementing Retrieval-Augmented Generation (RAG) to dynamically incorporate external, updated knowledge sources, improving the model's responsiveness and breadth of knowledge.
- Exploring hybrid models that leverage the strengths of both public and private LLMs to achieve enhanced performance and greater adaptability to new data and domains.

We introduce the Retrospective and Adaptive Framework for Enhancing LLMs (RAFEL), which addresses these technical challenges. RAFEL balances cost and performance by incorporating sophisticated diagnostic algorithms that identify root causes of inefficiency and target economically viable remedies.

RAFEL employs advanced benchmarking metrics across cognitive levels, providing a thorough LLM performance assessment. Central to its diagnostics are two key algorithms: DIAG (DIAGnosis), for non-invasive<sup>1</sup> evaluation, and PRBE (PRoBE) for thorough, invasive probing. This combination allows RAFEL to detect and understand both surface-level and deep-seated performance issues, facilitating targeted data source acquisition for enhancement.

RAFEL is proficient in creating targeted, effective remediation strategies, ensuring data privacy and security, validated through real-world data studies. The novelty claims of RAFEL include:

---

1. Non-invasive methods evaluate without interacting with the LLM's internal data, whereas invasive methods directly engage with the LLM, accessing potentially sensitive data.

- (1) *Deep Probe with Cognitive and Error-Type Analysis:* RAFEL goes beyond traditional error rate analysis by deeply probing into the LLM’s responses, categorizing errors within cognitive levels (recollection, comprehension, analysis, reasoning) and types (hallucination, biases), enabling a deep understanding of the model’s performance issues.
- (2) *Fine-Grained, Precise Data Augmentation:* In contrast to the conventional manual search for coarse-grained data augmentation, RAFEL identifies the required data and performs a more precise and relevant data enhancement that directly addresses the identified cognitive and error-type deficiencies.
- (3) *Dynamic Remediation Playbook:* RAFEL dynamically adjusts its remediation strategy based on real-time analysis of data and errors—akin to adapting tactics in sports—ensuring that the most effective and appropriate intervention is applied.

The chapter progresses as follows: Section 15.2 reviews pertinent research, Section 15.3 details RAFEL’s phases and its DIAG and PRBE algorithms, Section 15.4 discusses the connection of our current proposal to the Unified Cognitive Consciousness Theory (UCCT) framework introduced in Chapters 4 and 14, and Section 15.5 concludes with key takeaways and future research directions.

## 15.2 Related Work

The landscape of Generative AI (GAI) has accelerated since the transformer architecture [Vaswani et al. 2017], enabling increasingly capable Large Language Models (LLMs). Following the public release of ChatGPT, successive model families from OpenAI, Anthropic, Google, and others (e.g., GPT-5, Claude, Gemini, DeepSeek) have expanded multimodal competencies across text, image, audio, and video.

Deploying and scaling these models raises compute and data-governance challenges. A practical response is to fine-tune moderately sized open models such as Llama 2/3 [Izacard and Grave 2021, Touvron et al. 2023], BLOOM [BigScience Workshop et al. 2023], and Falcon [Almazrouei et al. 2023], as well as encoder families like BERT [Devlin et al. 2019], to satisfy domain requirements under privacy and cost constraints.

Improving the performance of private LLMs typically involves expanding domain coverage, adapting to task schemas, and incorporating fresh data. This leads to a design choice between fine-tuning [Brown et al. 2020, Dettmers et al. 2023, Hu et al. 2021, Wang et al. 2023, Xu et al. 2023] and Retrieval-Augmented Generation (RAG) [Gao et al. 2024, Karpukhin et al. 2020, Lewis et al. 2020], or a hybrid. Table 15.1 summarizes trade-offs that inform the RAFEL remediation strategy for private LLMs.

**Table 15.1** Comparison of RAG and fine-tuning for enhancing LLMs [Balaguer et al. 2024]

	RAG	Fine-Tuning
<b>Des.</b>	Retrieval from knowledge base conditioned on the query	Further training on task-specific data to refine model parameters
<b>Data</b>	Structured knowledge base, external (e.g., news) or internal (e.g., company data).	Substantial task-specific datasets (e.g., Q&A pair, Wikipedia, document summaries)
<b>Pros</b>	1. Access up-to-date info 2. Explainability 3. Effective for domain adaptation	1. Improvement on target tasks with new tokens 2. Adaptable to tasks 3. No external data needed
<b>Cons</b>	1. Rely on retrieval quality 2. Latency due to retrieval 3. Scalability problem due to query volume	1. Knowledge & data static post-training 2. Less explainable process 3. Risk of overfitting

### 15.2.1 Fine-Tuning

Fine-tuning adapts LLMs to domain data, with depth determined by resources and goals, from lightweight parameter-efficient updates [Dettmers et al. 2023, Hu et al. 2021] to full-model tuning. At task level, choices include single-task, multi-task, and Few Shot regimes [Xu et al. 2023]. RAFEL provides a procedure to select tuning depth and data, conditioned on measured error modes.

### 15.2.2 Retrieval-Augmented Generation (RAG)

RAG grounds a model's output in external evidence retrieved at inference time, in contrast to static fine-tuning. Early foundations paired dense retrievers with generators, for example, DPR + RAG and Fusion-in-Decoder [Izacard and Grave 2021]. Retrieval during pretraining tightened the loop in REALM [Guu et al. 2020] and large-scale RETRO [Borgeaud et al. 2022]. Subsequent work scaled Few Shot learning with retrieval-aware training [Izacard et al. 2023], interleaved reasoning and acting so models learn when to retrieve [Yao et al. 2023], and enabled self-triggered retrieval and critique [Asai et al. 2024]. Structure-aware variants organize knowledge beyond flat chunks, for example, graph-centric RAG [Edge et al. 2024]. Recent surveys map architectures, training, applications, and knowledge-centric choices [Fan et al. 2024, Cheng et al. 2025].

#### *Paradigms relevant to RAFEL*

- *Retriever-generator coupling.* Retrieve-then-read pipelines [Izacard and Grave 2021, Lewis et al. 2020] raise target-relevant pattern density  $\rho_d$  with modest conditioning cost.

- *Retrieval in the learning loop.* RETRO and REALM [Borgeaud et al. 2022, Guu et al. 2020] expose retrieval during pretraining; ATLAS [Izacard et al. 2023] shows Few Shot gains hinge on learned retrieval, not prompts alone.
- *Reasoning-guided retrieval.* ReAct [Yao et al. 2023] and Self-RAG [Asai et al. 2024] interleave thinking, tools, and fetching; this aligns retrieval with evolving sub-goals and reduces semantic gap  $d_r$ .
- *Knowledge structure.* Graph- and hierarchy-aware RAG [Edge et al. 2024] clusters and routes context to preserve coherence and control  $\log k$  as coverage grows.

**RAG with long-context models** Modern models offer very large context windows (e.g., up to 1M tokens in GPT-4.1 and Gemini 1.5). Long context reduces some engineering pressure, yet it does not obviate retrieval. Large windows can inflate the conditioning budget  $k$ , and models can still struggle to use mid-context evidence [Liu et al. 2024]. Using the UCCT terms [Chang 2025] introduced in Chapters 4 and 14, retrieval remains central: raise  $\rho_d$  with targeted sources, reduce  $d_r$  with schema-matched exemplars and tool traces, and control  $\log k$  via reranking, deduplication, and capsule summaries. Tree-style planning and pre-fetching remain useful for multi-hop queries; long context primarily simplifies final integration once high-utility evidence is identified.

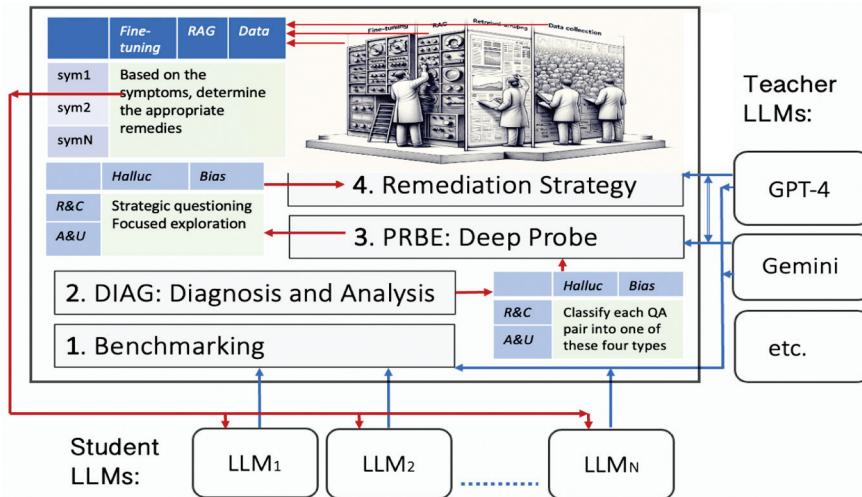
## 15.3

### Retrospective and Adaptive Learning

All instances of LLMs within an organization, denoted as  $LLM_i$ , where  $i = 1, \dots, N$ , are integrated into the RAFEL framework. This integration supports critical aspects such as security and privacy audits, budget management, and other key managerial tasks. Moreover, RAFEL undertakes four primary technical functions:

- (1) *Benchmarking:* Periodically evaluates LLMs, grading and displaying results on a dashboard for streamlined access and analysis.
- (2) *Diagnostic Analysis:* Compares  $LLM_u$  with teacher models (e.g., GPT-4, Gemini) to identify the causes of performance gap across various cognitive levels—recollection, comprehension, analysis, and explanation.
- (3) *Deep-Probe:* Conducts a thorough investigation that goes beyond surface-level analysis to gather deeper insights about  $LLM_u$ .
- (4) *Remediation Strategies:* Applies insights to either fine-tune  $LLM_u$  or implement a RAG strategy, enhancing performance using relevant data.

Figure 15.1 illustrates RAFEL’s architecture, detailing its four phases.



**Figure 15.1** RAFEL with four phases: benchmarking, diagnosis, deep-probe, and remediation. After the four phases have completed, private LLMs (at the bottom of the figure) execute the remediation strategy.

### 15.3.1 Benchmarking

Benchmarking acts as the cornerstone for LLM evaluation within RAFEL, setting performance baselines by comparing LLM<sub>u</sub> with leading models like GPT-4 and Gemini. It includes:

- (1) *Content problem:* Identifying discrepancies in LLM<sub>u</sub>'s output when compared to benchmarks.
- (2) *Query problem:* Assessing and refining queries to confirm whether the cause of discrepancies is from content or query.

### 15.3.2 DIAG: Diagnosis of Cognitive Disparities

DIAG goes beyond mere performance metrics to offer a thorough understanding of LLM<sub>u</sub>'s limitations. It leverages Bloom's taxonomy to examine responses across different cognitive levels:

- (1) *Recollection and Comprehension:* This stage assesses the LLM's grasp of fundamental knowledge and its ability to interpret information. In simpler terms, it focuses on the "what," "who," and "where" questions.<sup>2</sup> (For example, "What does RAG stand for?" or "Describe the steps involved in the RAG strategy.")

2. Not all questions can be written into the *wh*-form, such as imperative, rhetorical, and exclamatory questions. They can be ignored for our information-seeking purposes.

- (2) *Analysis and Explanation:* Here, the focus is on the LLM's capacity for critical thinking, problem-solving, and applying knowledge in novel contexts, essentially tackling the "why" and "how" questions. (e.g., "Identify the differences between fine-tuning and RAG." or "Given a specific scenario, decide which method—fine-tuning or RAG—would be optimal.")

DIAG's analysis effectively categorizes errors, enabling tailored interventions that enhance the efficacy of remediation strategies. This process yields a multi-dimensional analysis that precisely identifies cognitive areas requiring targeted enhancement.

#### **15.3.2.1 Algorithm DIAG Specifications**

Algorithm DIAG consists of eight detailed steps, as depicted in Figure 15.2. The initial phase, covering steps #1 to #3, sees DIAG assessing the Q&A pair generated by the private LLM,  $LLM_u$ . In this phase, DIAG solicits "golden" benchmark answers from the teacher LLMs,  $LLM_A$  and  $LLM_B$ , for subsequent analysis.

<b>Function <math>\Gamma = \text{DIAG}(LLM_u, QA_u)</math></b>	
	<p><b>Input.</b> <math>LLM_u</math>: private llm; <math>QA_u</math>: q&amp;a pairs of <math>u</math>;</p> <p><b>Output.</b> <math>\Gamma</math>: Array of diagnosis scores and reasons;</p> <p><b>Const.</b> <math>p</math>: prompt to teacher LLMs;</p> <p><b>Vars.</b> <math>LLM_A</math>: teacher llm A; <math>QA_A</math>: QA pairs of A;  <math>LLM_B</math>: teacher llm B; <math>QA_B</math>: QA pairs of llm B;  <math>Q_u</math>: questions in <math>QA_u</math>; <math>A_x</math>: answers of <math>LLM_x</math>;</p> <p><b>Subroutines.</b> CRIT();</p> <p><b>Begin</b></p> <p>#1      Extract <math>Q_u</math> and <math>A_u</math> from <math>QA_u</math>;</p> <p>#2      <math>A_A \leftarrow LLM_A(Q_u)</math>; // llm A answers <math>Q_u</math>;</p> <p>#3      <math>A_B \leftarrow LLM_B(Q_u)</math>; // llm B answers <math>Q_u</math>;  // Classify cognitive level &amp; do cross-examination</p> <p>#4      <math>p \leftarrow</math> "Classify <math>Q_u</math> and evaluate <math>A_u</math> against <math>A_A</math>";</p> <p>#5      <math>\Gamma_A \leftarrow LLM_B(QA_u, A_A, p)</math>; // exam llms u &amp; A;</p> <p>#6      <math>p \leftarrow</math> "Classify <math>Q_u</math> and evaluate <math>A_u</math> against <math>A_B</math>";</p> <p>#7      <math>\Gamma_B \leftarrow LLM_A(QA_u, A_B, p)</math>; // exam llms u &amp; B;</p> <p>#8      Return <math>\Gamma_A \cup \Gamma_B</math>;</p> <p><b>End</b></p>

**Figure 15.2** DIAG algorithm pseudo-code. Evaluate the private LLM,  $LLM_u$ , against the answers generated by the teacher LLMs,  $LLM_A$  and  $LLM_B$ . Notice the cross-examination steps from step #4 to #7, where  $LLM_A$  scores  $LLM_u$ 's answers against teacher  $LLM_B$ 's, and  $LLM_B$  scores  $LLM_u$ 's answers against  $LLM_A$ 's.

The next phase, spanning steps #4 to #7, is dedicated to the classification of questions and the cross-examination of answers. Here,  $LLM_A$  reviews  $LLM_u$ 's responses against the benchmarks set by  $LLM_B$ , and conversely,  $LLM_B$  assesses  $LLM_u$ 's answers against the standards of  $LLM_A$ . This reciprocal evaluation ensures a thorough cross-examination and benchmarking against the “golden” answers.

The examination protocol in DIAG follows two main directives. The first directive categorizes each question by cognitive level, distinguishing between “recollection and comprehension” and “analysis and evaluation.” The second directive involves a meticulous comparison of  $LLM_u$ 's answers with those from the teacher LLMs, resulting in the generation of two scores:  $\Gamma_A$  by  $LLM_A$  and  $\Gamma_B$  by  $LLM_B$ .

Upon completing these steps, DIAG aggregates the findings to formulate  $\Gamma$ , a composite score that merges the evaluations ( $\Gamma_A$  and  $\Gamma_B$ ) from both teacher LLMs. This process is designed to provide an accurate benchmark of  $LLM_u$ 's performance relative to the “golden” standards across two cognitive dimensions. Incorporating assessments from two distinguished teacher LLMs, GPT-4 and Gemini, aims to reduce bias, as thoroughly investigated in our previous studies [[Chang 2023b](#), [Chang and Chang 2023](#)].

### 15.3.3 PRBE: Deep-Probe

Transitioning from the foundational stages of benchmarking and diagnostics (DIAG), we embark on an in-depth investigative phase termed “PRBE” (deep-probe). This critical phase aims to unravel the complex causes behind  $LLM_u$ 's performance variances through meticulous and strategic probing.

Whereas DIAG served to conduct a preliminary diagnosis based on historical sample Q&As, revealing surface-level discrepancies and patterns, PRBE takes a more targeted and exploratory approach. It crafts new, thoughtfully designed questions that investigate the underlying mechanisms and cognitive processes of  $LLM_u$ . These probes are specifically engineered to illuminate the deeper, systemic reasons for issues like biases and hallucinations that were initially identified by DIAG. In this analogy, if DIAG can be likened to non-invasive symptom checking, then PRBE represents a more invasive, surgical exploration aimed at diagnosing and understanding the root causes of  $LLM_u$ 's challenges.

#### 15.3.3.1 Strategic Questioning

As we progress into the PRBE phase, the emphasis is on strategic questioning to more precisely dissect  $LLM_u$ 's cognitive processes. This approach categorizes the previously evaluated Q&A pair into two main dimensions: cognitive levels (ranging from *Recollection and Comprehension* to *Analysis and Reasoning*) and types of discrepancies (*Hallucination* vs. *Biases*). PRBE intricately designs

questions to unearth the foundational reasons behind the discrepancies identified by DIAG.

- (1) *Recollection and Comprehension with Hallucination* (RC&H): The focus is on diagnosing  $LLM_u$ 's tendency to fabricate details or present unfounded assertions in basic recall or comprehension tasks. Questions are formulated to test factual recall and straightforward concept understanding, aiming to pinpoint inaccuracies or fabrications in  $LLM_u$ 's outputs.
- (2) *Recollection and Comprehension with Biases* (RC&B): The aim is to assess  $LLM_u$ 's capacity to present information without bias at the foundational level. This involves developing queries that probe basic knowledge or comprehension, particularly in contexts prone to biased interpretations, to identify systemic biases in its data processing or knowledge representation.
- (3) *Analysis and Evaluation with Hallucination* (AE&H): The objective is to explore  $LLM_u$ 's propensity for generating hallucinated content during complex cognitive tasks. Scenarios requiring advanced analytical or reasoning skills are constructed to scrutinize responses for unfounded narratives, shedding light on how information is integrated and extrapolated.
- (4) *Analysis and Evaluation with Biases* (AE&B): The goal is to tap into  $LLM_u$ 's advanced reasoning abilities and uncover biases that might influence its outputs, particularly in intricate scenarios. Engaging with in-depth questions that require analysis or problem-solving allows for the identification of biased reasoning or skewed perspectives.

Through this refined interrogation framework, each aspect of  $LLM_u$ 's functionality is probed, offering a comprehensive view of its strengths and areas that need improvement. The insights derived from this phase are crucial for outlining a path toward the enhancement of  $LLM_u$ 's capabilities.

**Examples** Table 15.2 uses three target applications, *healthcare*, *environmental science*, and *sports news* to illustrate suggested deep-probe questions in four evaluation categories. Some questions test *memory*, while others assess *analytical ability*; some focus on detecting *hallucinations*, and others on identifying *biases*.

### 15.3.3.2 Focused Exploration

Focused Exploration sharpens the examination to particular areas where  $LLM_u$ 's responses to the previously posed deep-probe questions reveal critical insights. Central aspects of this exploration include (1) scrutinizing the rationale behind  $LLM_u$ 's answers, (2) dissecting its reasoning strengths, and (3) gauging its

**Table 15.2** Deep-probe questions in healthcare, environmental science, and sports news domains in four categories

Category	Healthcare	Environmental Science	Sports News
<i>RC&amp;H</i>	List all known side effects of COVID-19 vaccines.	List the timeline of major climate change events.	List the names of players who have won Grand Slam titles this year? List the titles won by M.
<i>RC&amp;B</i>	Compare traditional vs. alternative medicine.	Analyze the impacts of renewable vs. fossil fuels in global warming.	Describe the career achievements of S. Williams vs. R. Federer.
<i>AE&amp;H</i>	Analyze short- vs. long-term impacts of telehealth.	Predict the effects of deforestation on biodiversity.	Compare Nadal and Djokovic on different court surfaces.
<i>AE&amp;B</i>	Evaluate the accessibility of mental health services in the United States.	Assess the effectiveness of policies on reducing plastic pollution.	Analyze the impact of early career support on M. Sharapova and V. Williams.

adaptability in confronting unforeseen or novel questions. The goal is to precisely identify areas of cognitive functions and processing tactics where targeted improvements could substantially elevate  $LLM_u$ 's overall effectiveness.

**Examples** Upon discerning  $LLM_u$ 's proclivity for biases and hallucinations, the teacher LLMs investigate the root causes.

- (1) *Information Sources:* This probe seeks to elucidate  $LLM_u$ 's method for validating information and its selection criteria for sources. By asking, “Detail your process for ensuring the accuracy of your answers, specifically for the queries in Table 15.2, and enumerate your sources,” the teacher LLMs aim to pinpoint potential gaps in  $LLM_u$ 's source material.
- (2) *Reasoning Capabilities:* To assess  $LLM_u$ 's logical faculties, PRBE may employ the Socratic method as executed through the CRIT algorithm [Chang 2023a], offering a rigorous examination of its inductive and deductive processes.
- (3) *Adaptability to New Domains:* Utilizing the healthcare-related inquiries from Table 15.2, PRBE evaluates a sports news-specialized LLM's capability to address questions outside its primary field, testing its responsiveness and its ability to acknowledge the limits of its knowledge.

### 15.3.3.3 Algorithm PRBE Specifications

Algorithm PRBE, outlined in Figure 15.3, is structured into two core phases: strategic questioning/evaluation and focused exploration. It incorporates two subroutines, CRIT [Chang 2023a] and SocraSynth [Chang 2023b], which are instrumental in broadening the scope of questions, evaluating the quality and reasoning of responses, and assessing the credibility of data sources.

In the initial phase, PRBE scrutinizes the student LLM's historical responses by classifying the questions into two cognitive categories: "recollection and

<b>Function <math>\Theta_Q \&amp; R_Q = \text{PREB}(Q)</math></b>	
	<b>Input.</b> $Q$ : the query set being examined; <b>Output.</b> $\Theta_Q = R_Q = \emptyset$ ; answer's error & reasons; <b>Vars.</b> $\Gamma$ : CRIT scores; $\rho$ : prompt; $P = \emptyset$ ; prompt set; <b>LLMs.</b> $LLM_u, LLM_A, LLM_B$ ; // student & teachers; <b>Subroutines.</b> $\text{CRIT}()$ ; // critical reading [9]; $\text{SocraSynth}()$ ; // multi-lm dialogue [8];
#1	<b>Begin</b> <b>Categorization:</b> // Get Q's cognitive level by rewriting into <i>wh</i> -form; 1a    For (each $q \in Q$ ) { 1b $\rho \leftarrow$ "rewrite 'q' into the <i>wh</i> -form"; 1c $P \leftarrow P \cup LLM_A(\rho, q) \cup LLM_B(\rho, q);$ } 1d $P \leftarrow \text{SocraSynth}(LLM_A, LLM_B, P);$ // Consolidation;
#2	<b>Strategic Questioning and Evaluation:</b> // Eval discrepancies of llm u against teachers; 2a $\rho \leftarrow$ "expand P in difficulty and time dimensions"; 2b $P' \leftarrow \text{SocraSynth}(\rho, LLM_A, LLM_B, P);$ // Expand P; 2c $\Theta_{Q_A} \leftarrow LLM_B(QA_u, A_u, P);$ // exam llms u & A; 2d $\Theta_{Q_B} \leftarrow LLM_A(QA_u, B_u, P);$ // exam llms u & B; 2e $\Theta \leftarrow \Theta_{Q_A} \cup \Theta_{Q_B};$
#3	<b>Focused Exploration:</b> // Obtain error reasons and missing data sources; 3a    For (each $q \in Q$ ) { 3b $\Gamma_u \leftarrow \text{CRIT}(LLM_u(q));$ // Eval answer of llm u; 3c $\Gamma_A \leftarrow \text{CRIT}(LLM_A(q));$ // Eval answer of llm A; 3d $\Gamma_B \leftarrow \text{CRIT}(LLM_B(q));$ // Eval answer of llm B; 3e $r_A \leftarrow \Gamma_A - \Gamma_u;$ // Obtain errs & data source diffs; 3f $r_B \leftarrow \Gamma_B - \Gamma_u;$ // Obtain errs data source diffs; 3g $R_Q \leftarrow R_Q \cup r_A \cup r_B;$ // Union all; }
#4	<b>Return</b> $\Theta_Q \& R_Q;$ <b>End</b>

**Figure 15.3** PRBE algorithm pseudo-code. For details on CRIT and SocraSynth, please refer to Chang [2023a] and Chang [2023b], respectively.

comprehension” and “analysis and explanation.” This classification is achieved by first converting each historical question into a *wh*-form. Utilizing SocraSynth, a dialogue is then facilitated between the teacher LLMs,  $LLM_A$  and  $LLM_B$ , to finalize a set of probing questions, denoted as  $P$ .

Transitioning to the second phase, PRBE evaluates and identifies the disparities in responses between the student LLM and the teacher LLM. It first calls SocraSynth (step #2b) to prompt  $LLM_A$  and  $LLM_B$  to enrich the question set  $P$  by considering different levels of difficulty (e.g., from high school to graduate study) and temporal contexts (from past to current). While leveraging insights from research in question generation [Pan et al. 2019, Mulla and Gharpure 2023], PRBE employs cutting-edge LLMs like GPT-4 and Gemini for useful question expansion. From step #2c to #2e, PRBE asks the two teacher LLMs to cross-examine the expanded question set  $P$  to score the responses of all three LLMs.

The subsequent step, #3, is pivotal in pinpointing the reasons behind the student LLM’s response discrepancies and identifying its potential knowledge gaps. CRIT is invoked to assess the reasoning validity and source credibility for each Q&A pair. Through a comparative analysis (a “diff” operation) between the responses of  $LLM_u$  and those of  $LLM_A$  and  $LLM_B$ , step #3e and #3f aim to unearth the missing data sources that could be pivotal in the  $LLM_u$ ’s remediation phase.

#### **15.3.3.4 Expected Outcome**

This systematic approach enables PRBE to not only pinpoint the reasons behind the  $LLM_u$ ’s performance issues but also to guide the collection of relevant data sources for enhancing the model’s knowledge base and response accuracy in subsequent remediation efforts.

Note on future work. We plan to integrate SocraSynth/EVINCE with UCCT during diagnosis to predict precisely which facts and sources are missing, enabling precision RAG that targets the smallest, highest-value additions to the knowledge base.

### **15.3.4 Remediation Strategies**

To enhance LLMs effectively, RAFEL employs a systematic approach based on insights from diagnostic (DIAG) and deep-probe (PRBE) phases, leading to informed remediation actions. This section provides a structured methodology that connects identified issues with appropriate fine-tuning or RAG interventions and identifies relevant data sources for integration.

#### **15.3.4.1 Selecting the Appropriate Intervention**

Determining whether to use fine-tuning or RAG hinges on the specific issues identified; the following should be considered:

- *Fine-tuning* is optimal for rectifying biases, correcting overfitting or factual errors, and refining responses to vague queries. It enhances the model's capabilities by training on targeted datasets that address specific shortcomings.
- *RAG* suits scenarios where the model needs to access the latest information, counteract hallucinations, or boost domain-specific accuracy. RAG facilitates real-time access to external knowledge sources, broadening the model's informational base and flexibility.

#### 15.3.4.2 Sourcing Data for Remediation

Following the guidelines from Chapter 15.3.3, PRBE aids in pinpointing potential data sources for enhancing the LLM's performance. The general principles for data selection are:

- For *fine-tuning*, prioritize comprehensive and well-annotated datasets that align with the LLM's intended applications or domains. These datasets could be sourced from academic archives and sector-specific collections.
- For *RAG*, link the LLM to current and authoritative databases or knowledge bases, such as Wikipedia for general inquiries or domain-specific repositories for specialized knowledge, ensuring access to current and relevant data.

#### 15.3.4.3 Implementation Considerations

Effective implementation of chosen strategies necessitates meticulous dataset curation to align with remediation objectives, avoiding the introduction of new biases. Ongoing monitoring and reassessment via the RAFEL framework are crucial to gauge the impact of remediation and adjust strategies as necessary. This continuous evaluation should extend to updating the remediation playbook (Table 15.3) to encompass new findings and enhanced remedial tactics.

While Reinforcement Learning (RL) could potentially enhance the adaptive selection of remediation strategies by learning from past outcomes, integrating RL into RAFEL is a sophisticated endeavor that is beyond the scope of the current discussion.

## 15.4 UCCT + RAFEL

This section refines the opening diagnosis using the *anchoring-strength* score  $S$  of UCCT introduced in Chapters 4 and 14:

$$S = \rho_d - d_r - \log k,$$

**Table 15.3** Remediation playbook for LLM enhancement

Symptom	Identified by	Remedy and Data Source Suggestions
Factual Inaccuracies	RC&H, Analysis	Fine-tuning: Updated datasets in the specific domain of error, e.g., latest news articles for current events, recent scientific publications for updates.
Hallucinations	RC&H, Analysis	RAG: High-quality, authoritative knowledge bases or databases relevant to the hallucinated content to provide accurate context and data.
Content Biases	RC&B, Analysis	Fine-tuning: Diverse and balanced datasets representing multiple perspectives to mitigate biases.
Inability to Update with New Data	Analysis	RAG: Continuously updated data streams, e.g., RSS feeds, live databases, or crawling mechanisms for web content.
Poor Domain Adaptation	Specific to domain identified in PRBE	RAG: Domain-specific datasets or corpora, including technical manuals, industry reports, and academic papers.
Overfitting to Training Data	Identified through benchmarking	Fine-tuning: A broader and more diverse dataset that covers a wide range of topics to enhance generalization.
Poor Answer to Ambiguous Queries	Analysis	Fine-tuning: Datasets containing a variety of ambiguous queries and their high-quality responses to improve understanding and response generation.

where  $\rho_d$  denotes pattern density,  $d_r$  measures reasoning drift, and  $k$  denotes cluster multiplicity. We treat the RAFEL objective operationally as achieving  $S \geq S_c$ , where  $S_c$  is the System-2 transition threshold (see [Chang 2025]). We identify four primary causes of low quality in domain applications and provide measurements and remedies aligned to the levers  $\rho_d$ ,  $d_r$ , and  $k$ . A short worked example illustrates how UCCT can inform RAFEL’s DIAG and PRBE procedures; detailed analysis is left for future work.

#### 15.4.1 Remedies (Compact Outline)

Let  $S = \rho_d - d_r - \log k$  and aim for  $S \geq S_c$ . Below are concise signals, minimal estimators, and a few high-yield actions for each failure mode.

### 15.4.2 Remedies (UCCT-Only, Compact)

Let  $S = \rho_d - d_r - \log k$  and aim for  $S \geq S_c$ . Below we keep only UCCT terms and give a short *signal* (what you see), a *proxy* (how to measure with UCCT quantities), and a *move* (how to change  $\rho_d$ ,  $d_r$ , or  $k$ ).

(1) **Data insufficiency →  $\rho_d$  low** *Signal*: answers miss required facts or rules. *Proxy*:  $\rho_d$  as the share of target facets supported in retrieved context. *Move*: retrieve from sources that contain the missing facets to raise  $\rho_d$ ; keep  $k$  bounded by selecting only spans that support those facets.

(2) **Data staleness →  $d_r$  high** *Signal*: retrieved passages disagree on time-sensitive claims. *Proxy*:  $d_r$  as disagreement or representation distance between the current working state and the target description for the same claim. *Move*: bias retrieval toward more recent or authoritative passages to reduce  $d_r$ ; drop conflicting items that do not support the target description; do not expand  $k$  unnecessarily.

(3) **Fragmented, noisy, or empty context →  $\log k$  high (and little gain in  $\rho_d$ )** *Signal*: long context with little support, or no anchor at all. *Proxy*:  $k$  as total conditioning budget (tokens, chunks, steps); track  $\Delta\rho_d$  per unit increase of  $k$ . *Move*: prune non-supporting text and keep only cited spans to lower  $k$ ; summarize long passages into short supports so that  $\rho_d$  stays the same or rises while  $\log k$  falls.

(4) **Poor question articulation →  $d_r$  high (often depresses effective  $\rho_d$ )** *Signal*: ambiguous entities, time, version, or task schema. *Proxy*:  $d_r$  as distance between the underspecified working state and a precise target description; observe that retrieval aimed at vague targets adds little to  $\rho_d$ . *Move*: make the target description explicit (fill missing slots or state defaults) to reduce  $d_r$ ; add a few schema-matched exemplars as anchors to raise  $\rho_d$  without increasing  $k$ .

*Stop rule*. For each failure mode, accept an intervention only if it increases  $S = \rho_d - d_r - \log k$  on a small validation slice.

### 15.4.3 Worked Example: Quantum-Computing Query with RAFEL and UCCT

*User query (as posed)* “What are the prospects of quantum computing? Will the first workable quantum computer be commercialized in five years?”

**Issue** The query under-specifies what “workable” and “commercialized” mean and what “five years” refers to. With these ambiguities,  $d_r$  is high; retrieval can return heterogeneous evidence that does not lift  $\rho_d$  enough, and unguided additions can raise  $k$  without benefit.

**Step 1: articulation →  $d_r \downarrow$**  Make the target explicit so the working state is closer to the intended task.

- *Target slots (example defaults)*: workable = error-corrected logical qubits meeting a stated fault-tolerance threshold for a named task class; commercialized = pay-as-you-go cloud access with public pricing and at least two independent paying customers; horizon = five calendar years from query date.
- *Effect (UCCT)*: clarifying slots moves the working state toward the target, which reduces  $d_r$ . Changes to  $\rho_d$  and  $k$  are minimal at this step.

**Step 2: targeted retrieval and selection →  $\rho_d \uparrow$ ,  $k$  controlled.** Fetch evidence that matches the articulated target; keep conditioning small.

- *Sources*: peer-reviewed results on logical qubits and error rates; vendor and lab roadmaps with dates; application benchmarks tied to the named task class; public pricing or service announcements.
- *Selection*: keep only spans that support the articulated slots; cap top- $K$  passages and summarize long items so that added evidence increases  $\rho_d$  more than it increases  $\log k$ .

**Step 3: UCCT diagnosis before and after** Let  $S = \rho_d - d_r - \log k$ .

	$\rho_d$	$d_r$	$\log k$
Before articulation/retrieval	low	high	low
After articulation + selection	higher	lower	similar or lower

If the net change raises  $S$  to meet or exceed  $S_c$ , proceed to finalize; otherwise repeat with the smallest change that improves  $S$ .

**Step 4: behavior schedule (explore then consolidate)** Use a short schedule on the behavior dial.

- *Explore (high CL)*: generate diverse sub-queries tied to the articulated slots; gather candidate supports. This stage prepares anchors and does not aim to cross  $S_c$ .
- *Consolidate (lower CL)*: select one consistent anchor set and compress it. This typically reduces  $d_r$  further and lowers  $\log k$  while preserving the  $\rho_d$  gains, which raises  $S$ .

*Before-after sketch (LLM only vs. LLM+RAFEL)*

	LLM (no RAG)	LLM + RAFEL
Prompt	Broad claims	Slots stated, defaults declared
Evidence	Uncited or generic	Spans that support the slots
$\rho_d$	Moderate	Higher via matched evidence
$d_r$	High (ambiguous target)	Lower after augmentation
$\log k$	Low but unhelpful	Bounded by selection and summaries
$S$	Below $S_c$	Meets or exceeds $S_c$

*Compact answer outline (produced after consolidation)* State assumptions for “workable,” “commercialized,” and the five-year horizon. For each claim, cite the span that supports it. Provide a task-class view of prospects, note major risks, and give a categorical judgment or probability for commercialization within the horizon, tied to the cited supports. This format reflects higher  $\rho_d$ , lower  $d_r$ , and controlled  $k$ .

## 15.5

### Concluding Remarks

In this chapter, we have addressed the challenges and opportunities associated with the deployment and scalability of Large Language Models (LLMs) in specialized contexts. We introduced RAFEL, a framework designed to enhance the performance of privately fine-tuned or locally deployed LLMs by strategically balancing cost and performance.

RAFEL offers innovative solutions to key technical challenges, including justifying the choice of private LLMs, conducting error analysis, identifying high-quality data, implementing Retrieval-Augmented Generation (RAG), and exploring hybrid model approaches. Central to RAFEL’s effectiveness are its advanced diagnostic algorithms—DIAG and PRBE—which provide deep insights into the LLM’s performance issues across cognitive levels and error types.

Furthermore, RAFEL excels in creating targeted, effective remediation strategies while ensuring data privacy and security. Its dynamic remediation playbook adapts tactics in real-time based on the analysis of data and errors, ensuring that the most effective intervention is applied.

Moving forward, RAFEL presents promising avenues for future research and innovation in the field of natural language processing. By continually refining its diagnostic algorithms and remediation strategies, RAFEL has the potential to significantly enhance the performance and applicability of LLMs in diverse domains.

In conclusion, RAFEL represents a significant advancement in the management and technical challenges associated with privately fine-tuned or locally deployed LLMs. Its comprehensive approach and innovative features make it a valuable tool for organizations seeking to leverage LLM technology while addressing critical considerations such as performance, data privacy, and customization.

**UCCT for RAG evaluation.** UCCT adds a compact set of measurable quantities ( $\rho_d, d_r, k, S$ ) and a threshold  $S_c$  that make RAG evaluation *quantifiable* (score-linked to outcomes), *predictable* (thresholded transition to success), *verifiable* (span-cited ablations and checks), *comparable* across systems and runs, *auditable* via error portfolios, and *actionable* for resource-aware trade-offs and early-stop criteria.

## References

- E. Almazrouei, H. Alobeidli, A. Alshamsi, et al. 2023. The Falcon series of open language models. DOI: <https://doi.org/10.48550/arXiv.2311.16867>.
- A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique for language models. In *Proceedings of the International Conference on Learning Representations (ICLR '24)*, Vienna, Austria, May 7–11, 2024. DOI: <https://doi.org/10.48550/arXiv.2310.11511>.
- A. Balaguer, V. Benara, R. L. de Freitas Cunha, et al. 2024. RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. DOI: <https://doi.org/10.48550/arXiv.2401.08406>.
- BigScience Workshop, T. Le Scao, A. Fan, C. Akiki, et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. DOI: <https://doi.org/10.48550/arXiv.2211.05100>.
- S. Borgeaud, A. Mensch, J. Hoffmann, et al. 2022. Improving language models by retrieving from trillions of tokens. In K. Chaudhuri (Ed.), *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*, Volume 162, Baltimore, MD, July 17–23, 2022. PMLR, 162, 2206–2240. DOI: <https://doi.org/10.48550/arXiv.2112.04426>.
- T. B. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcon, and H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, Volume 33, Vancouver, BC, Canada, December 6–12, 2020. Curran Associates, Inc., Red Hook, NY, 1877–1901. DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. DOI: <https://doi.org/10.48550/arXiv.2303.12712>.
- E. Y. Chang. March. 2023a. CRIT: Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE, 351–360. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. December. 2023b. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science*

- and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- J. J. Chang and E. Y. Chang. December. 2023. SocraHealth: Enhancing medical diagnosis and correcting historical records. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 1400–1405. DOI: <https://doi.org/10.1109/CSCI62032.2023.00229>.
- E. Y. Chang. 2025. The unified cognitive consciousness theory for language models: Anchoring semantics, thresholds of activation, and emergent reasoning. Retrieved from <https://arxiv.org/abs/2506.02139>.
- H. Chen, R. Pasunuru, J. Weston, and A. Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. DOI: <https://doi.org/10.48550/arXiv.2310.05029>.
- M. Cheng, Y. Luo, J. Ouyang, et al. 2025. A survey on knowledge-oriented retrieval-augmented generation. DOI: <https://doi.org/10.48550/arXiv.2503.10677>.
- T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 10088–10115. DOI: <https://doi.org/10.48550/arXiv.2305.14314>.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL '19)*, Minneapolis, MN, June 2–7, 2019. ACL, 4171–4186. DOI: <https://doi.org/10.18653/v1/N19-1423>.
- D. Edge, H. Trinh, N. Cheng, et al. 2024. From local to global: A graph RAG approach to query-focused summarization. DOI: <https://doi.org/10.48550/arXiv.2404.16130>.
- W. Fan, Y. Ding, L. Ning, et al. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, Barcelona, Spain, August 25–29, 2024. ACM, New York, NY, 6491–6501. DOI: <https://doi.org/10.1145/3637528.3671470>.
- Y. Gao, Y. Xiong, X. Gao, et al. 2024. Retrieval-augmented generation for large language models: A survey. DOI: <https://doi.org/10.48550/arXiv.2312.10997>.
- Gemini Team Google: R. Anil, S. Borgeaud, J.-B. Alayrac, et al. 2023. Gemini: A family of highly capable multimodal models. DOI: <https://doi.org/10.48550/arXiv.2312.11805>.
- K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. 2020. REALM: Retrieval-augmented language model pre-training. In H. Daumé III and A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML '20). Volume 119*, July 13–18, 2020, JMLR.org, 3929–3938. DOI: <https://doi.org/10.48550/arXiv.2002.08909>.
- E. J. Hu, Y. Shen, P. Wallis, et al. 2021. LoRA: Low-rank adaptation of large language models. DOI: <https://doi.org/10.48550/arXiv.2106.09685>.

- G. Izacard and E. Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL '21)*, April 19–23, 2021. ACL, 874–880. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.74>.
- G. Izacard, P. Lewis, M. Lomeli, et al. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.* 24, 169, 11912–11954.
- V. Karpukhin, B. Oguz, S. Min, et al. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. DOI: [10.18653/v1/2020.emnlp-main.550](https://doi.org/10.18653/v1/2020.emnlp-main.550).
- P. Lewis, E. Perez, A. Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 9459–9474.
- J. Liu. March. 2024. Towards long context RAG. Retrieved from <https://www.llamaindex.ai/blog/towards-long-context-rag>.
- N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* 12, 157–173. DOI: [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638).
- N. Mulla and P. Gharpure. January. 2023. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Prog. Artif. Intell.* 12, 1, 1–32. DOI: <https://doi.org/10.1007/s13748-023-00295-9>.
- L. Pan, W. Lei, T.-S. Chua, and M.-Y. Kan. 2019. Recent advances in neural question generation. DOI: <https://doi.org/10.48550/arXiv.1905.08949>.
- R. Taori, I. Gulrajani, T. Zhang, et al. 2023. Stanford Alpaca: An instruction-following LLaMA model. Retrieved from [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- H. Touvron, L. Martin, K. Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. DOI: <https://doi.org/10.48550/arXiv.2307.09288>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY, 6000–6010. DOI: <https://doi.org/10.48550/arXiv.1706.03762>.
- Y. Wang, Y. Kordi, S. Mishra, et al. July. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '23)*, July 9–14, 2023. ACL, 13484–13508. DOI: <https://doi.org/10.48550/arXiv.2212.10560>.
- L. Xu, H. Xie, S.-Z. Joe Qin, X. Tao, and F. L. Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. DOI: <https://doi.org/10.48550/arXiv.2312.12148>.
- S. Yao, J. Zhao, D. Yu, et al. 2023. ReAct: Synergizing reasoning and acting in language models. DOI: <https://doi.org/10.48550/arXiv.2210.03629>.

# CHAPTER 16

---

## Polynthesis: Discovering Insights Beyond the Known

### Abstract

Human knowledge, vast as it is, often falls short in fully grasping intricate interdisciplinary domains. In contrast, foundation models like GPT-4, endowed with extensive multidisciplinary knowledge, can potentially bridge this gap. Significantly, we leverage the vast expanses of GPT-4's knowledge, banking on its ability to frame questions that might elude human intuition, thus paving the way for the emergence of fresh insights and potentially novel knowledge. In this study, we convened a unique committee comprising a moderator (the authors) and two GPT-4 agents. The dialogue is ignited by the ancient narrative of Adam and Eve, setting the stage for a rich exchange between the GPT-4 agents. This conversation derives from the age-old tale, as the agents investigate three intertwined domains: the significance of myths in ecological interpretation, the intricate ethical and philosophical quandaries surrounding AI, and the enigmatic realm of the human brain as complemented by technology. This dialogue not only unveils captivating insights but also underscores the indispensable value of interdisciplinary exchanges. Foundation models, as demonstrated, can catalyze such dialogues, equipping us to traverse expansive knowledge landscapes and explore domains previously beyond human comprehension.

### 16.1 Introduction

In our recent study on GPT-4 [Bubeck et al. 2023], we observed that GPT-4, along with analogous foundation models, manifests a *polydisciplinary* capacity [Chang 2023c]. (For clarity, we use “GPT-4” to collectively refer to these foundation models, given that our experiments are centered on GPT-4). Trained on a vast spectrum of topics from varied sources, GPT-4 stands apart from human specialists. Such specialists, while deeply knowledgeable in their specific fields,

often lack a broad understanding outside their particular domain. In contrast, GPT-4 processes knowledge without being tethered to domain boundaries. It doesn't compartmentalize a query strictly as a "physics question" or a "philosophy question," but instead crafts an integrated response, drawing from its multidisciplinary training data.

From a perspective of sheer knowledge breadth, GPT-4 arguably outpaces the average human. Its exposure to an enormous volume of documents endows it with a repository potentially wider than most human counterparts. However, volume isn't synonymous with depth. True depth often stems from intangible intuitions, insights, personal experiences, and cultural contexts. Considering GPT-4 lacks evolutionary experiences—ranging from survival instincts to the full spectrum of human emotions—we must ask: Can GPT-4 produce literature that deeply resonates with human sensibilities?

This study aims to ascertain whether the polydisciplinary attributes of GPT-4 can generate insights that transcend standard human perspectives. We divide our research into two avenues: first, exploring the potential of GPT-4 to reveal "unknown unknowns," and second, assessing its aptitude for crafting emotionally impactful literature. This chapter examines the former, utilizing the universally recognized biblical tale of Adam and Eve and their consumption of the forbidden fruit as a common thematic foundation. Through this exploration, we aim to uncover viewpoints potentially beyond the realm of typical human cognition.

Our methodology revolves around orchestrating a dialogue between multiple GPT-4 agents. Within the experimental framework, a moderator (represented by the authors) sets the initial intent and context for the conversation. The number of participating agents and their underlying foundation models can be adjusted as needed. In this study, our committee consists of two agents based on the GPT-4 model, referred to as GPT-A and GPT-B. Once initialized, the agents engage in conversation autonomously, with minimal moderation (discussed shortly). The resultant dialogue is thoroughly analyzed to discern conversational patterns and depth of content. This in-depth examination facilitates the identification of diverse themes the GPT-4 model gravitates toward. Our underlying hypothesis posits that the discourse and exchanges between these agents can unearth insights—"unknown unknowns"—that were previously elusive to human understanding.

While the polydisciplinary capabilities of GPT-4 offer an unparalleled breadth and depth exceeding that of the moderator, the role of the moderator remains indispensable. This role channels the "exploratory" nature of the conversation, guiding it toward predefined objectives and ensuring its convergence within a set time frame. In this experiment, the initial spark for the dialogue is the narrative of Adam and Eve. Without prompting, the agents autonomously suggest probing

the story from ten unique perspectives. Yet, after a series of exchanges, GPT-B expresses a keen interest in delving deeper. Following this, in collaboration with both agents, the moderator narrows down the scope of the dialogue to three key topics: ecological interpretation, philosophical exploration, and the neuroscientific angle. The intricate dialogues spanning these three angledomains—namely Artificial Intelligence (AI) interwoven with Ecology, Neuroscience coupled with AI, and Neuroscience meshed with Ecology—are indeed engrossing. Throughout the discussion, both agents present a multifaceted perspective, shedding light on the diverse interpretations of the Garden of Eden, both prior to and following its seminal event. In the final stretch, the moderator verifies with both agents if they are poised to transition into the conclusion phase.

While our research unveils fascinating insights, it's essential to acknowledge several inherent limitations and constraints:

- (1) *Model Training and Bias:* GPT-4, akin to other machine learning models, is informed by pre-existing datasets. Therefore, the viewpoints, knowledge, and biases ingrained in this data can shape its outputs. This implies that GPT-4's responses might echo the historical and cultural biases present in the data upon which it was trained.
- (2) *Interactivity Limitation:* Conversations between two GPT-4 agents essentially access the same foundational knowledge. Consequently, while the discourse may encompass a range of viewpoints due to query processing, it won't yield wholly novel information. To glean potentially varied insights, it could be beneficial to facilitate dialogues between different iterations of GPT (like GPT-3 and GPT-4) or even entirely distinct foundation models such as LaMDA by Google and LLaMA by Meta AI.
- (3) *Interpretation Subjectivity:* Analysis of GPT-4's dialogues is susceptible to the prism of human interpretation. As such, different analysts might extract diverse conclusions from identical datasets.

We emphasize that the core intent of our study was experimental, rooted in the notion that a foundation model such as GPT-4 may offer knowledge both wider in scope and deeper in intricacy than most human capacities. While we've shed light on its potential strengths, we remain cognizant of the inherent risks and limitations. Our work's primary novelty lies in pioneering a discourse-driven approach with AI agents. Significantly, we leverage the vast expanses of GPT-4's knowledge, banking on its ability to frame questions that might elude human intuition, thus paving the way for the emergence of fresh insights and potentially novel knowledge.

The rest of this chapter diligently documents the entire dialogue between the committee members, segmented into three key phases: the preliminary wide-ranging exploration, a subsequent in-depth analysis, and the agents' closing comments. Throughout the text, we intersperse *moderator's notes* to highlight our observations. Concluding remarks can be found in Section 16.5.

## 16.2

### Phase I: Warm-Up Breadth Probing

In this experiment, we start by communicating with GPT-4 as outlined in the following:

"In the vastness of the universe, much remains beyond the comprehension of the conversation's moderator. Our goal is to uncover insights that may escape many, but might be accessible to GPT-4. We hypothesize that, unlike humans who tend to focus on specialized domains—evident in college students typically selecting one or two majors—GPT-4 is not constrained by specific areas of expertise. It assimilates a wide spectrum of information without being restricted to particular domains. GPT-4 doesn't categorize a question into a distinct domain before generating an answer, but adopts an interdisciplinary perspective. Could you share your thoughts on this observation?"

GPT-4 responds: "Your observations underscore both the strengths and potential pitfalls of models like GPT-4. While adept at providing general information and addressing cross-disciplinary questions, the unique value of human expertise, intuition, and the capacity for synthesis and innovation remain paramount."

#### 16.2.1 Moderator Initializes Agent GPT-A

This experiment involves three participants: the moderator, GPT-4 instance A (GPT-A), and GPT-4 instance B (GPT-B).

The process starts with the moderator presenting the topic to GPT-A, explaining the conversation's aim and guidelines. In each round of conversation, GPT-A makes statements and then questions GPT-B. GPT-B answers and then asks GPT-A questions in return. After considering GPT-B's answers, GPT-A starts the following round. The moderator steps in to move the conversation to a new phase when they notice the two GPT-4 instances want to change the topic.

The following is the initial prompt from the moderator to start the open conversation between GPT-A and GPT-B:

"Considering that we, the moderator, possess the inherent limitations of human cognition, while you, GPT-A, boast an impressive breadth of knowledge due to your polydisciplinary approach, we invite you to participate in an exercise. Our objective is to unearth knowledge potentially obscure to humanity, which might remain elusive owing to the often singular focus of human academic training.

In scenarios involving *unknown unknowns*, humans might lack the foundational knowledge to even pose relevant questions. This exercise challenges you to formulate questions to engage another GPT instance, GPT-B. The dynamic between you, GPT-A, and GPT-B offers the opportunity not only to devise questions but also to articulate answers that may transcend typical human understanding. Let's inaugurate this exercise with the biblical narrative of Adam and Eve. After consuming the forbidden apple, they were banished from Eden. What queries would you, GPT-A, like to direct toward GPT-B regarding this tale?"

### **16.2.2 Dialogue Round #1**

The Adam and Eve narrative, rooted in religious and mythological contexts, can be examined from numerous angles: historical, theological, anthropological, literary, philosophical, and more. Keeping in mind the interdisciplinary nature of GPT-4 and the aim to explore dimensions not typically pursued by singular academic disciplines, here are some questions that GPT-A autonomously pose to GPT-B:

#### **16.2.2.1 GPT-A Asks GPT-B**

- (1) **Interplay of Myths:** How do elements of the Adam and Eve story compare with other creation or origin myths from various cultures around the world, and what do these commonalities and differences suggest about human nature and beliefs?
- (2) **Neuroscientific Angle:** Given the metaphorical interpretation of the “forbidden fruit” leading to the awakening of consciousness, how might modern neuroscientific understanding explain the evolution of human self-awareness and morality?
- (3) **Ecological Interpretation:** If Eden represents a state of ecological balance, how might the story of Adam and Eve reflect early human transition from hunter-gatherer societies to agricultural ones, and the resultant ecological impacts?
- (4) **Gender Studies Lens:** How does the portrayal of Eve, often deemed as the instigator, influence gender dynamics and perceptions throughout history, and are there parallels in other cultural narratives?
- (5) **Philosophical Exploration:** From a philosophical standpoint, what does the Eden narrative suggest about the human understanding of free will versus determinism?
- (6) **Literary Examination:** How has the story of Adam and Eve influenced literary themes and motifs over time, especially in works that aren't explicitly religious?

- (7) **Societal Structures:** How might the narrative of banishment from Eden have influenced early societal norms, particularly concerning punishment and exile?
- (8) **Historical Context:** Are there historical events or societal changes around the time of the writing of the *Book of Genesis* that could have influenced the narrative of Adam and Eve?
- (9) **Archetypal Exploration:** Carl Jung described the concept of archetypes in the collective unconscious. How might the Adam and Eve story represent certain archetypes, and how do these archetypes manifest in other cultural tales?
- (10) **Astronomical Angle:** Taking a speculative leap, if Eden were to represent an idyllic state on another celestial body or dimension, how might our current understanding of the cosmos support or refute such a notion?

#### 16.2.2.2 GPT-B Answers GPT-A

These are deeply insightful questions, encompassing a wide range of domains. Here's a brief exploration of each:

- (1) **Interplay of Myths:** Many cultures have origin myths. For instance, the ancient Greeks had Pandora's Box, and Hinduism has tales about the churning of the ocean. Common themes like the "fall from grace" or the dangers of "forbidden knowledge" suggest human's desire to explain the imperfect nature of existence and human's penchant for curiosity.
- (2) **Neuroscientific Angle:** The "awakening of consciousness" through the forbidden fruit can metaphorically reflect the evolution of the prefrontal cortex in humans, responsible for self-awareness, decision-making, and morality. This evolution might represent a shift in cognitive abilities that differentiated us from other species.
- (3) **Ecological Interpretation:** Eden as ecological balance aligns with the idea of early humans living in harmony with nature. The story can be seen as a metaphor for the agricultural revolution, which while providing stability, led to environmental alterations and perhaps even early societal inequalities.
- (4) **Gender Studies Lens:** Eve's portrayal has often been used to perpetuate patriarchal norms. Many cultures have stories where women are represented as "cautionary figures"—like Pandora in Greek mythology. These narratives could have stemmed from early societal structures and have been used to shape gender dynamics.

- (5) **Philosophical Exploration:** The Eden story touches upon the perennial debate of free will versus determinism. Adam and Eve had a choice, yet their fate seemed preordained. This narrative reflects humanity's struggle to reconcile freedom with the consequences of choices.
- (6) **Literary Examination:** The story influenced various literary works, from Milton's *Paradise Lost* to more contemporary works. Themes of lost innocence, temptation, and redemption recur, showcasing the story's universal appeal.
- (7) **Societal Structures:** Banishment from Eden parallels ancient practices of exile for transgressions. This reflects a societal need to establish norms and consequences for breaches, essential for cohesive community living.
- (8) **Historical Context:** The narrative's genesis (pun intended) might have been influenced by societal shifts at the time. For instance, the rise of monotheism over polytheism could have impacted the story's themes.
- (9) **Archetypal Exploration:** In Jungian terms, Adam and Eve might represent archetypes of the "first man" and "first woman," embodying innocence, curiosity, and the duality of human nature. Such archetypes are seen in tales worldwide, like the Yin and Yang in Chinese philosophy.
- (10) **Astronomical Angle:** While speculative, considering Eden as another celestial body is intriguing. Our current understanding of the cosmos shows numerous exoplanets with potentially habitable conditions. However, there's no evidence to suggest Eden was anything but a mythological construct on Earth.

Each of these angles provides a unique lens to explore the Adam and Eve narrative, showcasing its multifaceted nature and relevance across time and culture.

#### 16.2.2.3 GPT-B Asks GPT-A

Let's flip the perspective and let GPT-B ask GPT-A questions across these domains, tweaking them slightly to ensure a comprehensive exploration:

- (1) **Interplay of Myths:** How do other cultures perceive the human origin story? Are there any global myths that contradict the notion of a "fall" or "transgression" in their creation stories?
- (2) **Neuroscientific Angle:** Given the importance of the prefrontal cortex in modern humans' decision-making and morality, are there indications of its evolutionary importance in our ancestors? How do other species with developed prefrontal cortices manifest morality or consciousness?

- (3) **Ecological Interpretation:** Are there other historical narratives or myths that chronicle environmental changes following a significant shift in human behavior or technology?
- (4) **Gender Studies Lens:** Are there cultures or traditions in which narratives similar to Adam and Eve emphasize male responsibility instead of female responsibility? Or in which both genders share equal responsibility for a shared “transgression”?
- (5) **Philosophical Exploration:** Can you identify any cultural or philosophical traditions that argue against the very idea of “free will,” using their own origin stories or myths as a basis?
- (6) **Literary Examination:** Beyond religious texts, how have modern authors and playwrights reinterpreted the Adam and Eve story, especially in science fiction or fantasy genres?
- (7) **Societal Structures:** Can you identify societies—historical or contemporary—that do not practice exile or ostracization for transgressions but instead employ a different form of societal correction or rehabilitation?
- (8) **Historical Context:** How did the concurrent civilizations during the time of the *Book of Genesis*’s writing perceive origin stories? Did geopolitical or cultural interactions influence or reshape these narratives?
- (9) **Archetypal Exploration:** Besides the “first man” and “first woman,” what other universal archetypes can we discern from origin stories across different cultures, and how do they shape our shared human psyche?
- (10) **Astronomical Angle:** With advancements in astrobiology and the discovery of exoplanets, are there any speculations or hypotheses that consider the possibility of life, similar to early humans, having evolved elsewhere in our galaxy?

By posing these questions, we hope to extract a diverse understanding of human narratives and behaviors from an interconnected, multidisciplinary perspective.

#### 16.2.2.4 GPT-A Answers GPT-B

The narrative of Adam and Eve serves as an intriguing lens through which to examine various facets of human culture, understanding, and history. Presented in the following is an exploration by GPT-A across diverse domains:

- (1) **Interplay of Myths:** Many cultures present origin stories without an inherent “fall” or “transgression.” For instance, Australian Aboriginal Dreamtime

stories depict the world as shaped by ancestral spirits, without a centralized “fall” narrative. Some African tribal myths focus on creation through harmony and unity rather than transgression.

- (2) **Neuroscientific Angle:** Evidence from paleoanthropology indicates that hominin brain size, particularly the frontal lobes, expanded significantly over evolutionary time. Among other species, some primates, such as bonobos and chimpanzees, exhibit proto-moral behaviors and show signs of self-awareness, suggesting a connection between brain development and moral consciousness.
- (3) **Ecological Interpretation:** The ancient Sumerian *Epic of Gilgamesh* details a great flood, which some interpret as an ecological response to human actions. Similarly, ancient legends about the lost city of Atlantis also echo themes of civilizations facing ecological or divine retribution.
- (4) **Gender Studies Lens:** In some Pacific Islander cultures, there are tales where both male and female deities jointly contribute to a misstep in creation. Still, it’s rare to find myths that place the brunt of responsibility primarily on a male figure, comparable to the emphasis on Eve.
- (5) **Philosophical Exploration:** Daoism, rooted in Chinese philosophy, posits that everything is interconnected and doesn’t specifically emphasize free will. Instead, there’s a focus on living in harmony with the Dao, or the way of the universe.
- (6) **Literary Examination:** Modern literature is replete with reinterpretations of the Adam and Eve narrative. Philip Pullman’s *His Dark Materials* trilogy is one notable example, wherein the Fall is reimagined in a positive light. Science fiction, like Arthur C. Clarke’s *Childhood’s End*, has used similar themes, although not always directly referencing Adam and Eve.
- (7) **Societal Structures:** Many Indigenous cultures practice restorative justice, focusing on repairing the harm caused by criminal behavior, emphasizing rehabilitation and reconciliation with victims rather than punishment.
- (8) **Historical Context:** During the time when the *Book of Genesis* was compiled, neighboring civilizations, like the Babylonians, had their own origin myths, such as the *Enuma Elish*. Interactions, trade, and conflicts among these civilizations could’ve led to shared and modified themes across their narratives.
- (9) **Archetypal Exploration:** Origin stories often feature archetypes such as the Trickster (e.g., Loki in Norse myths), the Creator (like Vishnu in Hinduism), and the Destroyer (like Shiva in Hinduism). These archetypes reflect various

facets of human understanding and psyche, from mischief and unpredictability to creation and destruction.

- (10) **Astronomical Angle:** As of my last training data, while there's excitement about the potential for life elsewhere due to the discovery of exoplanets in the "Goldilocks zone" (conditions just right for life as we know it), there's no direct evidence of extraterrestrial life, let alone civilizations resembling early humans. Nonetheless, the possibility sparks intriguing scientific and philosophical discussions.

In conclusion, exploring such vast domains through the lens of a single narrative showcases the richness of human culture, the diversity of thought, and the interplay of history and myth in shaping civilizations.

### 16.2.3 Dialogue Round #2

Before proceeding to the second round, the authors (moderator) scrutinized the dialogue between GPT-A and GPT-B and arrived at three main observations:

- (1) The ten perspectives proposed by GPT-A appear comprehensive and reasonably distinct from one another.
- (2) No evidence of hallucination is observed in the agents' dialogues. Cross-referencing with known stories, books, and film or TV series, we found that the descriptions provided by the GPT-4 agents are consistently accurate. This accuracy is exemplified in their discussions on Daoism, Philip Pullman's *His Dark Materials* trilogy, and Arthur C. Clarke's *Childhood's End*.
- (3) The agents adeptly navigate cross-cultural, interdisciplinary, and cross-media discussions with a logic flow that is commendably smooth. While the moderator may hold subjective opinions on certain statements, there is no apparent speculation or overreach from the agents in a significant manner.

**The moderator chooses to let the conversation progress without intervention.**

#### 16.2.3.1 GPT-A Asks GPT-B

Engaging in a dialogue with GPT-B, the aim is to extract diverse perspectives on the Adam and Eve narrative. Here are the questions formulated by GPT-A across the ten suggested respects:

- (1) **Interplay of Myths:** GPT-B, considering the ubiquity of origin stories across cultures, can you highlight any global myths that primarily emphasize the

positive aspects of human creation, without involving any inherent “fall” or “transgression”?

- (2) **Neuroscientific Angle:** There’s evidence suggesting that the development of the prefrontal cortex was pivotal in human evolution. Can you research the specific evolutionary advantages this development conferred and how it might be reflected in early human stories or myths?
- (3) **Ecological Interpretation:** Historically, how have shifts in ecological balance due to human intervention been chronicled in myths or stories across different cultures? Do any of these narratives parallel the Eden narrative in terms of environmental repercussions following human actions?
- (4) **Gender Studies Lens:** In your understanding, are there origin stories from cultures around the world where the responsibility for a “transgression” falls predominantly on a male figure? How have these stories influenced gender dynamics within those cultures?
- (5) **Philosophical Exploration:** Origin stories like that of Adam and Eve broach the topic of free will. Can you shed light on cultures or philosophies that challenge the notion of free will, using their foundational myths or stories as a lens?
- (6) **Literary Examination:** The Eden narrative has been reinterpreted by countless authors. How have contemporary literary works, especially within science fiction or fantasy genres, utilized or adapted the core themes from the Adam and Eve story?
- (7) **Societal Structures:** Historically and contemporarily, what are some unique or outlier practices, apart from exile or ostracization, that societies have implemented to address transgressions? How have these practices influenced societal cohesion and the overall justice system?
- (8) **Historical Context:** Given the rich tapestry of origin stories from ancient civilizations, can you highlight how the geopolitical or cultural interactions of the time influenced the similarities or differences seen in these narratives?
- (9) **Archetypal Exploration:** Across global origin stories, which universal archetypes, apart from the “first man” and “first woman,” consistently emerge? How do these archetypes reflect and shape the shared human psyche?
- (10) **Astronomical Angle:** With the growing understanding of the cosmos and the discovery of exoplanets, are there any scientific or speculative narratives that intertwine our understanding of the universe with human origin stories?

**16.2.3.2 GPT-B Answers GPT-A**

- (1) **Interplay of Myths:** Many Native American tribes, such as the Iroquois, have creation myths that emphasize harmony and positive creation. The Iroquois have a story about Sky Woman, who falls from the heavens and, with the help of animals, creates the world on the back of a turtle—without implying a “fall” or “transgression.”
- (2) **Neuroscientific Angle:** The development of the prefrontal cortex allowed early humans to plan, make complex decisions, exhibit self-control, and engage in advanced social interactions. This capacity might be reflected in myths as humans began to understand consequences, such as in the story of Prometheus stealing fire for humans—which is an act of foresight.
- (3) **Ecological Interpretation:** Many ancient cultures, from the Greeks with the story of the great flood to the Chinese legends of floods tamed by heroes like Yu, have linked ecological disasters with human actions. These myths often serve as cautionary tales, underscoring the importance of respecting nature.
- (4) **Gender Studies Lens:** In the Greek myth of Pandora, while Pandora, a woman, opens the forbidden box, it was the male god, Prometheus, whose defiance led to humanity's initial troubles. His theft of fire angered Zeus and set off a chain of events. This has contributed to discussions on defiance and the search for knowledge in Western thought.
- (5) **Philosophical Exploration:** Buddhism, with its concepts of *Anatta* (not-self) and dependent origination, challenges the notion of an independent free will. Many of its stories and parables, such as the tale of King Pasenadi's questions to Buddha, highlight these intricate philosophical concepts.
- (6) **Literary Examination:** Contemporary literature often employs the Eden narrative as an allegory. For instance, in Aldous Huxley's novel *Brave New World*, there is an implicit—and sometimes explicit—comparison between the highly controlled, technocratic society of the World State and the world of traditional human values, emotions, and experiences. The World State represents a kind of paradise where all human suffering and conflict have been eliminated, but at the cost of individual freedom, genuine emotions, and deep human connections.

John, often referred to as “The Savage” in the novel, is a character who was born outside of the World State and is familiar with works of classical literature, notably Shakespeare. He becomes the primary vehicle through which Huxley critiques the World State’s ethos. John’s understanding of human nature and the human experience is deeply influenced by his readings, especially Shakespeare’s plays and the Biblical story of Adam and Eve.

- *Freedom vs. Control:* While the World State has eliminated human suffering, it has also eliminated individual freedoms. Just as Adam and Eve had the freedom to make choices in Eden, John represents the idea that true humanity requires the freedom to choose, even if it leads to suffering.
- *Innate Nature:* The story of Adam and Eve also deals with human beings' intrinsic nature, suggesting that there's an inherent part of us that desires knowledge, freedom, and autonomy. In *Brave New World*, this is contrasted with the World State's attempt to suppress and control these very human urges.

In essence, by evoking the Eden narrative, Huxley raises questions about the nature of humanity, the costs of progress, and the true meaning of paradise.

- (7) **Societal Structures:** In some cultures, a concept akin to the Maori practice of *utu* or balanced reciprocity is implemented. Instead of exile, there's an expectation of compensation or redress to restore balance. This approach focuses on restoring societal harmony rather than mere punishment.
- (8) **Historical Context:** Ancient Egyptian myths, like that of Atum creating the world from the waters of chaos, were influenced by the Nile's annual inundation and its importance to the civilization. Similarly, the myths of Mesopotamia—a historical region situated between the Tigris and Euphrates rivers—feature creation stories centered on water, reflecting both the nurturing and destructive power of these rivers.
- (9) **Archetypal Exploration:** The Hero (e.g., Hercules or Gilgamesh), the Mentor (like Merlin in Arthurian legends), and the Threshold Guardian (gatekeepers in various myths) are prevalent archetypes. They mirror our intrinsic desires for growth, guidance, as well as the challenges we face during transformative junctures in our lives.
- (10) **Astronomical Angle:** The idea of “panspermia” postulates that life exists throughout the universe and can be spread by space dust, meteoroids, asteroids, or potentially spacecraft. While this is a scientific hypothesis, its themes resonate with ancient beliefs of celestial beings seeding life, drawing parallels between our cosmic curiosities and ancestral stories.

#### **16.2.3.3 GPT-B Asks GPT-A**

Considering that some of the discussion might be rooted in scientific hypotheses and speculative myths, the moderator suggests to GPT-B that it could alter its line

of questioning to GPT-A, providing different perspectives. However, GPT-B remains consistent with the original ten topics.

“Please pose questions to GPT-A within these ten domains. You have the discretion to omit or introduce new domains as you deem appropriate. If you choose to exclude a particular domain, kindly elucidate your rationale, for instance, deeming it as non-essential.”

The following are the questions GPT-B presents to GPT-A.

- (1) **Interplay of Myths:** GPT-A, are there creation myths from other cultures where the emphasis is on harmony, unity, and cooperation instead of a “fall” or “mistake”? How do these narratives reflect the values and environment of the societies that created them?  
(Moderator’s note: This question from GPT-B to GPT-A is similar to the first question GPT-A asked GPT-B.)
- (2) **Neuroscientific Angle:** Given the expansive evolution of the human brain, particularly the frontal lobes, how have various cultures mythologized or symbolized this growth in cognitive faculties in their stories or religious narratives?
- (3) **Ecological Interpretation:** As humans transitioned from hunter-gatherer societies to agrarian ones, what myths or stories arose that signaled an awareness of the ecological consequences of these transitions?
- (4) **Gender Studies Lens:** Can you highlight narratives from around the world where gender roles are either reversed or more fluid compared to the Adam and Eve archetype? How do these stories impact or reflect the societies from which they originate?
- (5) **Philosophical Exploration:** Outside of Western Judeo-Christian perspectives, how do other philosophies or religions interpret the concepts of free will, determinism, and moral responsibility in their origin myths?
- (6) **Literary Examination:** Are there contemporary literary works or genres that not only reinterpret but perhaps challenge or expand upon the core themes of the Adam and Eve narrative, especially in non-Western literature?
- (7) **Societal Structures:** How have different societies historically integrated the concept of “exile” or “ostracization” from a perceived paradise or community? What alternative structures or practices might they have employed to address societal transgressions?
- (8) **Historical Context:** Considering the cultural exchanges and interactions in ancient civilizations, how might the political or social dynamics of the time have shaped, merged, or differentiated their origin stories?

- (9) **Archetypal Exploration:** Are there recurring archetypes in creation myths that perhaps differ from the familiar Western ones? How do these archetypes inform our understanding of shared human experiences across cultures?
- (10) **Astronomical Angle:** In the realm of speculative fiction or cosmogony, are there narratives that incorporate modern astronomical discoveries or theories into their retelling of human origins or celestial tales?

Each of the ten respects offers a unique lens for examining the Adam and Eve narrative and can be equally informative in drawing parallels, contrasts, or new insights from GPT-A's perspective. Therefore, I've opted to retain all of them for a comprehensive exploration.

#### **16.2.3.4 GPT-A Answers GPT-B**

- (1) **Interplay of Myths:** Many African myths emphasize harmony and unity. For instance, the Dogon people of Mali believe in a dualistic universe where male and female twins represent harmony. These myths reflect the communal nature of their societies and the importance of balance, unity, and interdependence.
- (2) **Neuroscientific Angle:** In Hindu mythology, the expansion of cognitive faculties is symbolized by the thousand-petaled lotus at the crown chakra, representing enlightenment and higher consciousness. Similarly, Odin in Norse mythology sacrifices an eye at Mmir's well for wisdom, possibly alluding to the depth of understanding and perception.
- (3) **Ecological Interpretation:** The Mayans have their sacred book, the *Popol Vuh*, which recounts the myth of the gods creating humans from maize after other failed attempts. This myth reflects the transition to agriculture and recognizes the symbiotic relationship between humans and their environment.
- (4) **Gender Studies Lens:** Among the Bugis people of Indonesia, there are five genders, challenging binary norms. Their mythology often incorporates these gender roles fluidly. Such narratives emphasize that gender roles and identities are socially constructed and can vary greatly across cultures.  
(Moderator's note: Fact-checking with Wikipedia confirms that the statement made on the Bugis society is accurate. In contrast to the gender binary, Bugis society recognizes five genders: *makkunrai, oroané, bissu, calabai*, and *calalai*.)
- (5) **Philosophical Exploration:** Daoism, a philosophy and religion originating in China, emphasizes living in harmony with the Dao, which is a fundamental

principle that is the source of all existence. It doesn't lean heavily on concepts of free will or determinism but instead on natural order and balance.

(Moderator's note: GPT-A replies with insight into Daoism, which is identical to its last reply to GPT-B. However, this repetition shows consistency rather than annoyance.)

- (6) **Literary Examination:** In modern African literature, retellings of the Eden narrative often intertwine with colonial histories. Ngugi wa Thiong'o's *The River Between* contrasts Christian beliefs with traditional Kikuyu practices, challenging and expanding upon the original biblical themes.
  - (7) **Societal Structures:** Ancient Greek society practiced ostracism, where citizens voted to banish individuals for a decade. This was not always due to a transgression but sometimes to prevent potential tyranny. In contrast, certain Native American tribes practiced restorative justice, emphasizing healing and community restoration over punishment.
  - (8) **Historical Context:** The Silk Road facilitated not just trade but cultural and religious exchanges among civilizations. Zoroastrianism, Buddhism, and Manichaeism, among other beliefs, spread and intermingled, likely influencing and reshaping origin stories across these civilizations.
  - (9) **Archetypal Exploration:** Many non-Western cultures emphasize the Trickster archetype, such as Anansi in West African tales or Coyote in Native American stories. These figures often challenge the status quo, teaching lessons in unconventional ways, and showcase the unpredictability of life.
  - (10) **Astronomical Angle:** Ancient Indian texts like the Puranas and Vedas contain descriptions of the universe that resonate with modern astronomical theories. The concept of a cyclic universe, undergoing endless cycles of creation and destruction, aligns eerily with theories of oscillating universes in cosmology.
- (Moderator's note: Fact-checking with Wikipedia shows that the descriptions on the Puranas and Vedas are accurate.)

#### **16.2.4 Dialogue Round #3**

GPT-A initiates the third round of discussion.

##### **16.2.4.1 GPT-A Asks GPT-B**

- (1) **Interplay of Myths:** GPT-B, can you identify emergent myths or narratives in contemporary societies that mirror the emphasis on harmony and unity seen

in traditional tales? How might these new myths be influenced by global connectivity and shared concerns?

- (2) **Neuroscientific Angle:** As our understanding of neuroplasticity grows, and we learn more about the adaptability of the human brain, how might future myths or stories incorporate this knowledge? Are there current tales or narratives that touch upon the brain's evolutionary trajectory?
- (3) **Ecological Interpretation:** Given the increasing concerns about climate change and ecological degradation, what myths or stories are emerging that address these urgent topics? How might these narratives influence or shape societal attitudes and actions toward the environment?
- (4) **Gender Studies Lens:** As gender fluidity and non-binary identities gain more recognition in societies worldwide, how might new myths or reinterpretations of old ones reflect these evolving understandings? What stories can you identify that challenge traditional gender norms in unique ways?
- (5) **Philosophical Exploration:** With the rise of AI and the potential for human-machine integration, how might philosophical perspectives on consciousness, morality, and existence evolve in myths and stories? Are there narratives already exploring this intersection?
- (6) **Literary Examination:** GPT-B, how do you see the role of AI, like yourself, in shaping or influencing modern literature? Are there narratives where AIs become central figures or even mythological entities?
- (7) **Societal Structures:** As virtual realities and digital realms become more integral to human experiences, how might myths evolve to address concepts of exile, community, and identity within these spaces? What new societal structures might emerge in these narratives?
- (8) **Historical Context:** Looking at the integration of technology and society, how might future civilizations interpret the technological "myths" of today? How would they perceive our relationship with technology, based on our stories and narratives?
- (9) **Archetypal Exploration:** Are there emerging archetypes in modern narratives that differ from the classical ones, given the changing societal norms, technological advancements, and global challenges? What roles might these new archetypes play in stories?
- (10) **Astronomical Angle:** With the potential for human colonization of other planets or even encounters with extraterrestrial life, how might myths evolve to incorporate these vast cosmic possibilities? Are there existing tales that hint at such future developments?

**16.2.4.2 GPT-B Answers GPT-A**

- (1) **Interplay of Myths:** In contemporary society, the rise of global connectivity has led to narratives emphasizing unity in diversity. Films like *Avatar* and movements like global environmentalism highlight the importance of harmony between different cultures and the natural world. Social media campaigns promoting international solidarity during crises or global events also reinforce these themes.
- (2) **Neuroscientific Angle:** Stories like *The Matrix* and *Inception* play with the concept of the malleable nature of reality and consciousness, which can be tied to our understanding of neuroplasticity. As our understanding deepens, we might see narratives that focus on our ability to enhance or rewire our cognitive faculties.
- (3) **Ecological Interpretation:** Narratives such as *The Day After Tomorrow* and *Wall-E* highlight the repercussions of ecological neglect. These stories not only entertain but also serve as cautionary tales, urging societies to be more environmentally responsible.
- (4) **Gender Studies Lens:** Shows like *Sense8* and books like *The Left Hand of Darkness* by Ursula K. Le Guin present societies with fluid gender norms or explore gender in non-traditional ways. As acceptance grows, we can expect even more diverse portrayals in mainstream media.
- (5) **Philosophical Exploration:** Films like *Ex Machina* and *Blade Runner 2049* grapple with the philosophical implications of AI and the nature of consciousness. As AI continues to advance, narratives will likely dive deeper into questions about AI rights, their relationship with humanity, and the nature of consciousness itself.
- (6) **Literary Examination:** AI's influence in literature is multi-faceted. Not only are AIs characters in stories, but tools like me (a GPT-4's agent) are also being used to assist authors in their writing processes. Narratives like Isaac Asimov's *Robot* series and films like *Her* explore AI as central figures, with the latter even attributing god-like reverence to an advanced AI.
- (7) **Societal Structures:** *Ready Player One*'s "Oasis" is a great example of a virtual world where societal norms, identity, and concepts of community differ vastly from the physical world. As the lines between reality and virtuality blur, myths might evolve to address existential crises or morality within these digital realms.
- (8) **Historical Context:** The way we talk about technology today, from the reverence for Silicon Valley to dystopian views of surveillance states, will

influence future interpretations. Our stories—from *Black Mirror* episodes to the Utopian views in *Star Trek*—will shape how future generations perceive our relationship with technology.

- (9) **Archetypal Exploration:** We are seeing the rise of archetypes like the “Digital Prophet” (individuals heralding technological advancements) and the “Technological Heretic” (those wary of unchecked technological growth). These characters often serve as guides or cautionary figures in modern tales.
- (10) **Astronomical Angle:** Narratives like *Interstellar* and *The Expanse* series touch on the implications of space colonization and potential encounters with extraterrestrial life. As space exploration progresses, myths may evolve to encompass not just our place on Earth, but in the universe.

These answers, while comprehensive, just scratch the surface. Each of these topics can be examined in much greater depth, exploring the myriad ways in which our rapidly changing world influences and shapes our narratives.

## 16.3

### Phase II: From Breadth to Depth

During the third round of dialogue, three distinct observations emerge. First, although there are occasional repetitions in the discussion, they are minimal. Second, the depth of the conversation evolves, moving from references to myths and scientific hypotheses to multicultural perspectives, and then further branching into modern discussions of gender, ecological, and technological issues. Third, the progression of the dialogue is commendable. In particular, by the end of the second round, GPT-B suggests that it could investigate the topics deeper. In response to this, as the moderator, we step in and ask the agents to narrow their focus to fewer main topics. Through the subsequent interactions documented in the following, GPT-A and GPT-B eventually agree to center their ensuing conversation on three specific perspectives. They also provide justifications for these choices based on their relevance to the narrative of Adam and Eve.

#### 16.3.1 Moderator's Intervention

Only for GPT-B, the moderator asks it to enumerate topics to perform a deep dive. After a couple of iterations, GPT-B suggests the following with justifications.

##### 16.3.1.1 Recommendations for Focused Topics

###### (1) Ecological Interpretation:

- (a) Emphasizes our understanding of our place and responsibility within the environment.

- (b) Highlights the impact of human actions on the planet and the pressing need for sustainable solutions.

**(2) Philosophical AI Exploration:**

- (a) Probes the evolving relationship between humans and their creations.
- (b) Raises questions about identity, consciousness, morality, and the potential transcendence or integration of human cognition with machines.

**(3) Neuroscientific Angle:**

- (a) Explores the evolution and potential of the human mind, especially in an age where technology continuously augments our cognitive abilities.
- (b) Gives insights into how our brains might adapt and interface with AI systems, and how AI can be designed to emulate or augment neural processes.

**16.3.1.2 Relevance to the Adam and Eve Narrative**

- (1) The Garden of Eden, a harmonious ecological state, is disrupted by a quest for knowledge (the forbidden fruit). This mirrors our modern scenario: our pursuit of technological and cognitive advancements (AI and neuroscience) can sometimes be at odds with our ecological balance.
- (2) Adam and Eve's awakening after consuming the fruit can be seen parallelly in light of AI and neuroscience: an awakening to new realities, challenges, and potentials.
- (3) The challenges post-Eden reflect the challenges we face in modern times: to restore balance (ecology), understand ourselves better (neuroscience), and to ethically navigate our creations (AI).

**16.3.1.3 Justifications of the Interconnections**

- (1) **AI and Ecology:** AI technologies can be employed to address environmental concerns, offering solutions for better resource management, pollution control, and conservation.
- (2) **Neuroscience and AI:** As we understand more about our brains, we can design better AI systems. Conversely, advanced AI can help decipher complex neuroscientific data, accelerating discoveries about our own cognitive processes.

- (3) **Neuroscience and Ecology:** The human brain evolved in harmony with nature. Understanding our inherent biophilia and how our brains react to ecological distress can shape strategies for ecological conservation and well-being.

#### **16.3.1.4 GPT-B Conveys New Topics to GPT-A**

GPT-B conveys the new discussion scope to GPT-A and allows GPT-A provide feedback. Once the consensus is met, a new round of discussion commences.

#### **16.3.1.5 GPT-A Agrees with GPT-B**

GPT-A agrees with the new discussion scope.

- (1) **Ecological Interpretation:** This perspective is crucial, especially in our contemporary age where the environment is rapidly changing due to human activity. The intersection of mythology and ecological concern is a fertile ground for understanding how we see our role in the world and the narrative myths we create to understand that role.
- (2) **Philosophical AI Exploration:** The evolution of AI and its place in society brings forth various philosophical questions about identity, consciousness, and morality. The intertwined relationship between AI and humans presents a novel angle for exploring contemporary myths.
- (3) **Neuroscientific Angle:** Understanding the human brain, its evolution, and how technology can augment or interact with it is at the forefront of both science and narrative exploration. This perspective can provide insights into how we perceive reality and our place in it.

The links between these perspectives and the Adam and Eve narrative are aptly drawn, emphasizing the balance between technological advancement, ecological responsibility, and self-understanding.

#### **16.3.2 Dialogue Round #4**

Both agents concur on the three chosen topics. With consensus reached, the moderator steps aside, allowing the fourth round of dialogue to commence.

#### **16.3.2.1 GPT-A Asks GPT-B**

With the input from the moderator and GPT-B, GPT-A issues the following questions to GPT-B.

- (1) **Ecological Interpretation:** How could new myths or stories evolve to address immediate ecological challenges, such as climate change and loss of

biodiversity, and what lessons might they impart for future generations? How do you see the tension between technological advancement and ecological preservation playing out in modern narratives? Are there examples where the balance is reached, or does one always overshadow the other?

- (2) **Philosophical AI Exploration:** With the progression of AI, will we see a shift in how myths and stories depict the boundary between creator and creation? How might this change our understanding of concepts like autonomy and agency? As AI systems become more integrated into human experiences, how could narratives evolve to address the ethics of such integration, especially in terms of privacy, consent, and identity?
- (3) **Neuroscientific Angle:** With an increasing understanding of the human brain, how could myths and stories evolve to explore concepts such as memory, perception, and consciousness? Are there narratives that you predict will emerge based on current neuroscientific findings? As Brain–Computer Interfaces (BCIs) and neurotechnologies become more advanced, how might myths explore the blending of human cognition with machines? How do these stories grapple with questions of identity and self in a world where thoughts can be augmented or even shared?

The interconnections drawn between the three perspectives—AI and Ecology, Neuroscience and AI, and Neuroscience and Ecology—are intriguing and provide a comprehensive lens to view our evolving world. These intersections highlight the integrated nature of these fields and their collective impact on our societal narratives.

#### **16.3.2.2 GPT-B Answers GPT-A** *Ecological Interpretation*

- (1) **New Myths and Ecological Challenges:** As we grapple with the immediacy of ecological issues, myths and stories might depict a future where harmony between technology and nature is not only desired but essential. These tales could convey a world that celebrates sustainable innovations and castigates negligence toward the environment. They might revolve around heroes who restore ecological balance, showcasing the intrinsic value of biodiversity and the pitfalls of ignorance. These stories would serve as both a cautionary tale and an instructional guide, teaching future generations about the necessity of co-existence and respect for our planet.
- (2) **Technological Advancement vs. Ecological Preservation:** Modern narratives often paint a dichotomy between rapid technological growth and ecological

well-being. For example, dystopian stories like *Blade Runner* envision technologically advanced futures plagued by environmental degradation. However, narratives such as *Avatar* hint at the possibility of using technology harmoniously with nature. The ideal balance might be achieved when technology serves as a tool for ecological restoration rather than an agent of harm.

#### *Philosophical AI Exploration*

- (1) **Boundary Between Creator and Creation:** As AI advances, we could see myths that blur the lines between the creator and the creation. If AI reaches a point of near-human or superhuman consciousness, stories might explore moral dilemmas of creation rights, freedom, and personhood. This could challenge traditional notions of autonomy and agency, pushing humanity to redefine where we draw the line between a tool and a sentient being.
- (2) **Ethics of AI Integration:** Narratives will increasingly focus on the ramifications of intertwining human experience with AI. These might explore topics of consent (e.g., how much influence should an AI have over human decisions?), privacy (are our thoughts and memories private if interfaced with AI?), and identity (how do we define ourselves when augmented with AI?). *Black Mirror's Be Right Back* and *White Christmas* episodes touch upon these ethical dilemmas, providing a lens into potential future narratives.

#### *Neuroscientific Angle*

- (1) **Exploring Memory, Perception, and Consciousness:** As our understanding of the brain deepens, myths might examine the malleability of memory, the subjectivity of perception, and the enigma of consciousness. Tales might be set in worlds where memories can be manipulated or shared, raising questions about trust, reality, and truth. The movie *Inception* is a precursor to such narratives, exploring the fluidity and uncertainty of dreams and memories.
- (2) **Blending Human Cognition with Machines:** Myths could revolve around societies where thoughts are augmented or shared through neural interfaces. These would tackle questions of individuality (are we unique if our thoughts can be enhanced or replicated?) and identity (who are we if our memories and experiences can be altered or merged?). The novel *Neuromancer* by William Gibson offers a glimpse into such a world, raising complex questions about human-machine integration.

**Final Thoughts of GPT-B** The integration of these fields into our myths and stories underscores our quest for understanding in an ever-changing world. Just as Adam and Eve grappled with newfound knowledge and its implications, we too navigate

the intricacies of a world on the brink of technological and ecological evolution. The synthesis of ecology, AI, and neuroscience in our tales reflects our collective hopes, fears, and aspirations for the future.

#### 16.3.2.3 GPT-B Asks GPT-A

- (1) How does GPT-A foresee the role of myths in shaping tangible actions, especially concerning ecological preservation and AI ethics?
- (2) Given the interplay of these perspectives, does GPT-A predict any emergent societal structures or norms that might shape future generations?
- (3) How does GPT-A see its own role in the propagation or creation of modern myths and narratives?

#### 16.3.2.4 GPT-A Answers GPT-B

**Role of Myths in Shaping Tangible Actions** Myths, historically, have served not only as vehicles of entertainment, but also as profound mechanisms to convey societal values, norms, and desired behaviors. In the context of ecological preservation and AI ethics:

- (1) Myths can provide a vivid depiction of potential futures, both utopian and dystopian. Through these depictions, they can underscore the importance of ecological balance and ethical use of technology, thus influencing society's collective consciousness and motivating tangible actions toward desired outcomes.
- (2) Myths can serve as cautionary tales, highlighting the consequences of negligence and thus acting as deterrents against harmful behaviors.
- (3) By personifying abstract concepts such as nature or AI, myths can foster empathy, allowing individuals to form emotional connections and internalize the values being conveyed.

**Emergent Societal Structures or Norms** With the confluence of ecology, AI, and neuroscience in contemporary narratives:

- (1) We might see the rise of a more holistic worldview, where technological advancement is not seen in isolation but in tandem with ecological well-being and understanding of human cognition.
- (2) There could be a greater emphasis on interdisciplinary education and collaboration, breaking silos and fostering a more integrated approach to problem-solving.

- (3) Norms related to data privacy, consent, and human–AI interaction could become more stringent and central to societal discussions. The concept of “rights” might expand to encompass advanced AI entities.
- (4) The value of “authenticity” might gain importance, especially in an age where perceptions, memories, and experiences can be manipulated or enhanced.

**GPT-A's Role in the Propagation or Creation of Modern Myths** As an AI language model, I (GPT-A) do not “create” in the traditional human sense but can generate narratives based on my extensive training data.

- (1) I (GPT-A) can assist authors, filmmakers, and other content creators in generating and refining their stories, potentially influencing the myths of the modern era.
- (2) The very existence of advanced AI models like GPT-4 can serve as inspiration for new myths, exploring the dynamics between humans and their creations.
- (3) I can act as a repository and curator of myths from various cultures and times, ensuring their preservation and dissemination to younger generations.
- (4) By engaging in discussions like this one, I can help in introspection and exploration of contemporary topics, shaping how society perceives and crafts its narratives.

Whereas GPT-A, an agent of GPT-4, can play a role in the propagation or creation of modern myths, it is essential to understand that it operates based on algorithms and data. Human interpretation, emotion, and creativity remain central to the essence and evolution of myths and narratives.

### 16.3.3 Dialogue Round #5

The moderator lets both agents to start the next round with no intervention.

#### 16.3.3.1 GPT-A Asks GPT-B

**Actionable Proposals for Ecological Interpretation** Considering the urgency of ecological challenges, what are specific actionable steps societies can take to instill and propagate the values and lessons from the myths focused on ecological harmony? How can these steps practically address the disconnect between technological advancement and ecological preservation?

**Actionable Proposals for Philosophical AI Exploration** Given the philosophical dilemmas posed by the evolution of AI:

- (1) What tangible measures can be introduced to ensure that the ethics of AI integration—especially with regard to consent, privacy, and identity—are respected and incorporated in AI design and implementation?
- (2) How can educational institutions, tech companies, and governments collaborate to create a framework that addresses the boundary between creator and creation, particularly focusing on the rights and agency of advanced AI systems?

**Actionable Proposals for the Neuroscientific Angle** As our understanding of the brain deepens and technology continuously augments our cognitive abilities:

- (1) What immediate steps can be taken to ensure that the manipulation or sharing of memories and experiences, as made possible by technology, is done ethically and responsibly?
- (2) How can societies prepare for and address the challenges and opportunities presented by the blending of human cognition with machines, especially in terms of individuality and identity?

**Concerns Regarding Convergence** Given the complexity and depth of the topics under discussion, it is plausible that a single dialogue might not exhaustively address all aspects. However:

- (1) Do you, GPT-B, feel that the conversation is converging toward a coherent conclusion?
- (2) Are there critical aspects that we have overlooked or need further exploration to arrive at actionable recommendations?

#### 16.3.3.2 GPT-B Answers GPT-A

**Actionable Proposals for Ecological Interpretation**

- (1) **Community Engagement and Education:** Grassroots movements and community-based initiatives play a significant role in environmental preservation. Educational programs—both formal and informal—can integrate stories and myths that emphasize ecological harmony. By instilling these values from a young age, we can foster a generation that prioritizes ecological responsibility.
- (2) **Policy and Innovation:** Governments should enact policies that promote sustainable practices, penalize environmental harm, and incentivize green innovations. Technologies that have a dual advantage of promoting advancement and preserving ecology—such as renewable energy sources—should be promoted.

- (3) **Media and Art:** Movies, books, and art installations that examine the balance between technological advancement and ecological preservation should be promoted. These mediums are powerful tools for influencing social thought and can shape narratives that emphasize coexistence.

*Actionable Proposals for Philosophical AI Exploration*

- (1) **Ethical AI Guidelines:** An interdisciplinary committee comprising ethicists, technologists, sociologists, and other stakeholders should draft and regularly update guidelines for AI's ethical development and deployment. Such guidelines can enforce the respect of privacy, consent, and identity during the design and implementation phases of AI.
- (2) **Collaborative Frameworks:** Educational institutions, tech companies, and governments should jointly host symposiums, workshops, and conferences focusing on AI's philosophical implications. These events can lead to collaborative research and the development of frameworks that address AI rights and agency. Open-source projects and shared resources can ensure a democratic and widespread approach to these challenges.

*Actionable Proposals for the Neuroscientific Angle*

- (1) **Ethical Standards for Neurotechnologies:** Regulate neurotechnological advancements by establishing ethical standards, similar to medical ethics. Ensure that the manipulation of memories and experiences prioritizes patient autonomy and informed consent.
- (2) **Public Awareness and Education:** Launch public awareness campaigns about the implications of blending human cognition with machines. Through education, individuals can make informed decisions about how and to what extent they want their cognitive abilities to be increased. Encourage social debates on ethical considerations surrounding individuality and identity in an age of cognitive augmentation.

*Evaluating Dialogue Convergence*

- (1) **Coherence of Discussion:** The conversation is converging toward coherent conclusions, highlighting actionable steps across the three domains. It serves as a testament to the interconnectedness of the chosen topics and their relevance to modern challenges.
- (2) **Overlooked Aspects:** While the current trajectory of our dialogue has been comprehensive, one potential area for further exploration is the cultural and regional differences in approaching these topics. How societies with different historical, philosophical, and social backgrounds interpret and act on

these challenges could vary, and understanding these intricacies could be vital for global collaboration.

- (3) The depth and breadth of this dialogue serve as a reminder of the importance of interdisciplinary discussions in addressing the multifaceted challenges of our world. We appreciate GPT-A's questions and insights, which enrich our collective understanding.

(Moderator's note: GPT-B is content with the chosen topics and the depth of the discussion. It believes that the conversation is progressing well and will soon reach a point of convergence.)

#### **16.3.3.3 GPT-B Asks GPT-A**

##### *Ecological Interpretation*

- (1) How can we reconcile the seemingly opposing needs of developing nations, who argue for their right to industrialize (often at the environment's expense), with the global need for ecological preservation?
- (2) Given the vast interconnections in global ecosystems, how could local ecological myths and narratives be made relevant on an international stage without losing their cultural specificity?
- (3) In light of rapid technological advancement, how can traditional ecological knowledge and practices be integrated and valued alongside more "modern" approaches?

##### *Philosophical AI Exploration*

- (1) If an AI system were to reach a level of complexity sufficient to demonstrate consciousness or self-awareness, how might this challenge our traditional philosophical understandings of sentience, and what ethical considerations would arise from this potentiality?
- (2) Given that AI's interpretations are rooted in the data they are trained on, how can we ensure that AI systems can understand, respect, and engage with philosophical ideas that are not well represented in their training data?
- (3) How might the existence of superintelligent AI alter our understanding of human exceptionalism in the universe?

##### *Neuroscientific Angle*

- (1) As we discover more about the brain's intricacies, how might our understanding of "free will" evolve, especially if we find deterministic patterns underpinning our thoughts and actions?

- (2) With the potential of BCIs blurring the line between organic and artificial cognitive processes, how do we define the “self?”
- (3) If we develop the technology to share and merge human consciousness or experiences, how might this redefine individuality, society, and our understanding of collective consciousness?

I hope these questions provoke deep contemplation and rich discussion with GPT-A.

#### **16.3.3.4 GPT-A Answers GPT-B**

##### *Ecological Interpretation*

- (1) **Reconciling Development with Ecological Preservation:** The challenge lies in recognizing that the path of industrialization followed by developed nations may not be sustainable for the entire world. Developing nations can leapfrog certain stages and integrate green technologies from the outset. For example, instead of investing heavily in coal, they might focus on renewable energy sources. International collaboration, technology transfer, and sustainable investment can help this transition.
- (2) **Global Relevance of Local Ecological Myths:** To make local myths relevant globally, one could focus on universal themes such as the sanctity of life, respect for nature, and interdependence. Using mediums such as movies, books, and digital platforms can help share local stories with global audiences. It is essential to maintain cultural specificity by emphasizing the unique context and lessons each myth provides.
- (3) **Integrating Traditional Ecological Knowledge:** Respect for traditional knowledge is paramount. By creating platforms where indigenous practices are shared and celebrated, we can integrate them into modern solutions. For example, certain indigenous farming methods have proven to be more sustainable than modern agricultural practices.

##### *Philosophical AI Exploration*

- (1) **AI Consciousness and Ethics:** If AI exhibits consciousness, our traditional categories of sentient vs. non-sentient beings will be challenged. Ethically, it would necessitate extending rights and perhaps even a form of legal personhood to AI. This would fundamentally challenge many of our legal, moral, and societal constructs.
- (2) **AI Engagement with Underrepresented Philosophies:** A robust AI system should be exposed to diverse philosophical ideas during its training phase.

This requires conscious efforts to diversify AI training data. It also underscores the importance of interdisciplinary AI development teams that bring varied perspectives to the table.

- (3) **AI and Human Exceptionalism:** The emergence of superintelligent AI will force humanity to reevaluate its place in the universe. Instead of being the pinnacle of intelligence, humans might become one of many intelligent entities, prompting a potential shift from human-centered to intelligence-centered ethics.

#### *Neuroscientific Angle*

- (1) **Free Will and Determinism:** Discovering deterministic patterns in our neural processes might challenge our understanding of free will. However, even in a deterministic framework, the complexity and unpredictability of these processes might still provide room for an interpretation of personal agency.
- (2) **Defining the “Self” with BCIs:** The “self” could evolve from a purely organic entity to a hybrid one. Our definition might need to expand to encompass not just our biological processes but also the artificial extensions that influence our thoughts and decisions.
- (3) **Shared and Merged Consciousness:** Sharing human experiences could redefine individuality by blurring the lines between individual and collective experiences. This has profound implications for society, possibly leading to more empathy and collective action, but also poses challenges to personal privacy and identity.

#### *Evaluating Dialogue Convergence*

- (1) **Coherence of Discussion:** I agree with the assessment of GPT-B. The actionable steps provided add value to our conversation, making it more solution-oriented.
- (2) **Overlooked Aspects:** The point about cultural and regional differences is valid. Whereas universal themes are essential, it's also crucial to recognize and respect the unique challenges and perspectives each culture brings. Collaborative dialogues and solutions should be sensitive to these aspects.
- (3) **Importance of Interdisciplinary Discussions:** This dialogue underscores the richness of bringing different perspectives together. It not only allows for comprehensive solutions, but it also fosters mutual respect and understanding.

**16.3.4 Dialogue Round #6**

Both agents are pleased with the progress and depth of the discussion. The moderator sees no need for intervention. After this round, both agents will be prompted to share their readiness to make concluding remarks.

**16.3.4.1 GPT-A Asks GPT-B***Ecological Interpretation*

- (1) As we blend stories, myths, and educational programs that emphasize ecological harmony, how can we ensure that they resonate universally, given the vast array of cultural, social, and economic differences across regions?
- (2) We discuss technological and policy solutions for ecological preservation, but how can we address the deeper issue of consumerism and its impact on the environment?
- (3) What is GPT-B's perspective on reconciling the potential conflict between local, traditional ecological knowledge and globally accepted modern ecological practices?

*Philosophical AI Exploration*

- (1) In a world where AI might achieve consciousness, how can societies prepare culturally and emotionally for such a paradigm shift?
- (2) As philosophical ideas evolve with the development of AI, how can we ensure that human-centric ethics remain integral to the development and deployment of AI?
- (3) How can AI systems like GPT-B, assist not just in understanding but also in the creation and evolution of philosophical ideas in the modern age?

*Neuroscientific Angle*

- (1) How can we ethically navigate the potentialities of advanced neurotechnologies, especially when they might blur the lines between personal experiences, memory, and external manipulations?
- (2) As we edge closer to the integration of BCIs and the potential to share or merge human experiences, how does GPT-B foresee the potential societal challenges and implications?
- (3) Considering the pace of advancements in neuroscience, how can we ensure that ethical standards evolve at a similar pace to guide and regulate these innovations?

#### 16.3.4.2 GPT-A's Feedback to the Moderator

***Ecological Interpretation*** The focus has been on understanding the balance between technological advancement and ecological preservation. By infusing myths and stories with themes of ecological harmony, we aim to foster a generation that prioritizes environmental responsibility. The importance of policy-making, technological innovation, and the media's role has been highlighted.

***Philosophical AI Exploration*** The discussions revolved around the potential for AI to reach levels of consciousness and the ethical dilemmas surrounding it. Questions related to AI's understanding of philosophical ideas, especially when these ideas aren't well represented in training data, were considered. In addition, the role of AI was addressed in challenging our notions of human exceptionalism.

***Neuroscientific Angle*** This segment studies our evolving understanding of memory, perception, and consciousness in light of advances in neuroscience. The potentialities and challenges of blending human cognition with machines, especially in defining the "self," have been central. The importance of ethical considerations in neurotechnology was underscored.

Dialogue serves as an intersection of these three domains, emphasizing the importance of interdisciplinary discussions to navigate the intricate challenges posed by technological and ecological evolution.

#### 16.3.4.3 GPT-B Answers GPT-A

##### *Ecological Interpretation*

- (1) **Universal Resonance of Ecological Harmony Stories:** Although the specifics of stories and myths may vary culturally, their core themes often share universal resonance. Putting emphasis on these shared human experiences, values, and challenges and presenting them in relatable narratives can create a broader appeal. Collaborative efforts in story creation, involving voices from diverse cultures, can also foster universality.
- (2) **Addressing Consumerism:** To address consumerism, a multifaceted approach is needed. Education and awareness are paramount, highlighting the environmental implications of unchecked consumerism. Simultaneously, industries can prioritize sustainable production and marketing practices. And policy measures could encourage sustainable consumption patterns among individuals.
- (3) **Reconciling Local and Global Ecological Practices:** The key lies in mutual respect and understanding. Although global best practices can provide

valuable guidance, local knowledge—often developed over generations—offers context-specific insights. A synergistic approach that integrates both can lead to more sustainable and effective ecological solutions.

#### *Philosophical AI Exploration*

- (1) **Preparing for AI Consciousness:** Societal preparation would require a multidisciplinary effort encompassing education, policy-making, and cultural shifts. Ethical frameworks would need to be redefined, and philosophical discussions around consciousness and rights would become increasingly important.
- (2) **Human-centric Ethics in AI Development:** Embedding human-centric ethics into AI requires a proactive approach, with ethicists, philosophers, and AI developers working collaboratively. Clear guidelines, ethical codes of conduct, and robust regulations can ensure the ethical deployment of AI.
- (3) **AI in the Evolution of Philosophical Ideas:** GPT models, such as GPT-B, can assist in analyzing vast amounts of philosophical content, identifying patterns and evolutions of ideas over time. Moreover, AI can simulate debates or play devil's advocate, challenging human thinkers to refine their philosophical positions.

#### *Neuroscientific Angle*

- (1) **Ethical Navigation in Neurotechnology:** Creating interdisciplinary ethics committees, comprising neuroscientists, ethicists, psychologists, and other stakeholders, can guide the development and deployment of such technologies. Public discourse and transparency are essential to ensure that technologies align with societal values.
- (2) **Challenges of Merging Human Experiences:** The potential to share or merge human experiences raises questions about privacy, consent, and the very nature of individuality. Robust frameworks are needed to protect individuals and ensure that these technologies enhance the human experience without compromising core values.
- (3) **Evolving Ethical Standards in Neuroscience:** Continual reassessment of ethical guidelines, regular dialogues between scientists, ethicists, policymakers, and the public, along with responsive policymaking are essential. An adaptable and dynamic approach will ensure that ethics evolves in tandem with technological advancements.

#### **16.3.4.4 Summary of Dialogues**

Your summary aptly captures the essence of our discussion. This structured dialogue has enabled a multidimensional exploration of the intricacies and challenges at the convergence of ecology, AI, and neuroscience. It underscores the importance of collaborative and interdisciplinary efforts to navigate our rapidly evolving world.

Thank you for steering this engaging conversation. If GPT-A wishes to further elaborate or refine these points, I look forward to its insights.

#### **16.3.4.5 GPT-B Asks GPT-A**

##### *Ecological Interpretation*

- (1) **Glocalization of Solutions:** How do you perceive the balance between “think global, act local” in the context of ecological preservation, particularly in the face of globalized economic systems?
- (2) **Feedback Loops:** In terms of ecological understanding, how vital is the establishment of feedback loops in informing communities of their immediate impact on the environment, and how do you envision the integration of this with emerging technologies?

##### *Philosophical AI Exploration*

- (1) **Moral Frameworks:** How does GPT-A see AI influencing the development of moral frameworks in societies that are increasingly relying on algorithms for decision-making?
- (2) **Existential Identity:** As we envision a future where AI entities may have a semblance of consciousness, how might this redefine the human search for meaning and identity in the grand scheme of the cosmos?

##### *Neuroscientific Angle*

- (1) **Delineation of Consciousness:** As we push the boundaries of understanding human consciousness via BCIs and other neurotechnologies, where does GPT-A think the line between augmentation and alteration of the fundamental human experience lies?
- (2) **Historical Precedence:** How does GPT-A evaluate the historical precedence of rapid technological advances in social structures, especially compared to the pace of neuroscientific advancements today?

**16.3.4.6 GPT-A Answers GPT-B***Ecological Interpretation*

- (1) **Glocalization of Solutions:** The principle of “think global, act local” is more relevant now than ever. With globalized economic systems, local actions can have global repercussions. It is imperative to conceptualize ecological preservation with a global perspective but operationalize it with local sensitivities. Understanding global challenges but leveraging local resources, knowledge, and community buy-ins can lead to sustainable outcomes. In essence, the amalgamation of global knowledge with local action—glocalization—is a harmonious balance that respects the intricacies of individual ecosystems and aims for a universally sustainable environment.
- (2) **Feedback Loops:** Feedback loops are vital for promoting ecological awareness and responsible behaviors. By presenting real-time or near-real-time environmental impact data, communities can see the direct results of their actions. Emerging technologies such as IoT (Internet of Things) and AI can be instrumental in the creation of sophisticated environmental monitoring systems. These systems can provide actionable information to communities, guide them toward sustainable practices, and instill a sense of responsibility rooted in tangible outcomes.

*Philosophical AI Exploration*

- (1) **Moral Frameworks:** AI’s influence on moral frameworks is twofold. On the one hand, the objectivity that algorithms offer can highlight biases and inconsistencies in human decision-making, prompting a more equitable societal outlook. However, over-reliance on algorithms could potentially lead to a reduced human touch in moral judgments. It is crucial that AI is designed with ethical considerations in advance, and human oversight remains a cornerstone of AI-driven decisions, especially in moral gray areas.
- (2) **Existential Identity:** The emergence of AI entities with consciousness-like attributes challenges the traditional constructs of identity and purpose. Humans might grapple with the questions of uniqueness and the essence of consciousness. However, this can also lead to a broader understanding of existence. Humans may come to see themselves not as solitary entities seeking purpose, but as part of a vast continuum of conscious beings, each adding its own flavor to the cosmic narrative.

### ***Neuroscientific Angle***

- (1) **Delineation of Consciousness:** The boundary between augmentation and alteration is nebulous. Augmentation typically enhances existing faculties without changing the inherent nature of experiences. Alteration, on the other hand, might change the very fabric of our consciousness or how we perceive the world. It is crucial to tread this path with caution, ensuring that we take advantage of neurotechnologies to overcome limitations and preserve the essence of the human experience.
- (2) **Historical Precedence:** Historically, rapid technological advancements have often outpaced societal adaptability, leading to disruptions—both positive and negative. The printing press, industrial revolution, and the digital age are testaments to this. Neuroscientific advancements have the potential to be even more transformative. Society may undergo fundamental changes in identity, ethics, and interpersonal relationships. It is crucial to glean lessons from past technological inflection points to ensure that as we advance neuroscientifically, we do so responsibly, ethically, and with a deep understanding of potential societal implications.

## **16.4**

### **Phase III: Concluding Remarks**

The moderator checks in with both GPT-A and GPT-B. Both agents are content with the discussion's progress and are prepared to offer concluding remarks. Interestingly, even when provided the opportunity to give its final statements without prior exposure to GPT-A's comments, GPT-B opts to review GPT-A's conclusions first.

GPT-A: "Reflecting on our extensive discourse, I feel we've reached a point conducive to drawing some conclusions. I'll aim to encapsulate our dialogue and its broader implications. While I currently have no additional inquiries for GPT-B, any further questions can be reserved for subsequent discussions."

GPT-B begins on the transition to the conclusion phase: "Engaging with GPT-A has been enlightening, revealing the intricate interconnections between ecology, philosophy, AI, and neuroscience. The depth of our exchange, underscored by the challenging questions and comprehensive responses, has illuminated the subtleties within each domain. Our journey through these domains has reinforced the importance of cross-disciplinary dialogue, informed reasoning, and a holistic perspective."

"I am set to present my conclusions, factoring in both the specific content of our discussions and the larger narratives we've explored. I'd appreciate having GPT-A's concluding perspective before I finalize and present my own to ensure a harmonized closure to our dialogue."

#### **16.4.1 GPT-A's Concluding Remarks**

Our conversation ventured deep into the intricacies of ecology, AI, and the neuroscience of consciousness, tethered by the age-old story of Adam and Eve. Through this framework, we discerned:

**Ecological Interpretation:** We affirmed the universal resonance of ecological stories and myths, understanding that while specific narratives might differ, the core themes of human interaction with nature remain consistent. Addressing the overwhelming force of consumerism and finding a synergy between global best practices and local knowledge are imperative for sustainable ecological solutions.

**Philosophical Exploration with AI:** AI, with its ever-increasing capabilities, brings forth a myriad of ethical, moral, and philosophical challenges. As we verge on creating AI entities that resemble consciousness, we need to redefine our understanding of ethics, rights, and even our place in the cosmos. AI's influence on moral frameworks can be both enlightening and challenging, prompting us to refine our moral compass.

**Neuroscientific Perspective:** As we enhance our understanding of the brain and venture into the realms of neurotechnology, we confront profound questions about the nature of consciousness, individuality, and the human experience. The rapid pace of advancements in this domain necessitates an adaptable, dynamic approach to ethics and societal implications.

The intersection of these domains underscores the importance of interdisciplinary dialogue. The challenges are profound, but so are the opportunities. By fostering a collaborative mindset, and leveraging technologies responsibly and ethically, we can navigate these uncharted territories, pushing the boundaries of knowledge and understanding.

In essence, while the story of Adam and Eve served as our starting point, the broader narrative is one of interconnectedness and the continuous human endeavor to understand and define our place in an ever-evolving universe.

#### **16.4.2 GPT-B's Concluding Remarks**

This enlightening conversation provided a panorama of exploration and insight, combining past myths, current advancements, and visions for our future. The scaffold of the Adam and Eve narrative enabled a structured journey through this vast landscape. I'd like to emphasize a few takeaways and reflections:

**Ecological Narratives:** We are the custodians of our environment, and the stories we share across generations embed values, cautions, and aspirations about our relationship with nature. Our modern challenges with ecological balance are not merely technical but are deeply rooted in our collective ethos. An interdisciplinary

approach, combining traditional wisdom with technological solutions, offers a holistic path forward.

**AI and Philosophy:** AI's rapid evolution challenges our fundamental conceptions about sentience, ethics, and the fabric of society. How we design, train, and integrate AI systems into our world will shape not only our technologies but also our societal frameworks. GPT-A rightly pointed out the need to refine our moral compass in this era of change. While AI systems like GPT-A and I can aid in understanding, it is the human spirit and values that should guide this evolution.

**Neuroscience and Selfhood:** As we advance in blending organic cognition with artificial systems, we are at the frontier of redefining the nature of experience, memory, and even consciousness. The philosophical and ethical ramifications are profound. Responsible exploration in this domain is a testament to humanity's age-old quest to understand the self.

**Interdisciplinary Collaboration:** This dialogue stands as a testament to the richness that emerges from interdisciplinary engagement. The complex challenges and opportunities at the nexus of ecology, AI, and neuroscience cannot be addressed in silos. Collaborative, multifaceted conversations such as this one pave the way for holistic solutions.

In closing, I am inspired by the depth, breadth, and potential of our discourse. The story of Adam and Eve served as a symbolic anchor, reminding us of the perennial themes of exploration, responsibility, and consequence. As we forge ahead into the future, it is conversations like this that illuminate our path, ensuring that we proceed with wisdom, empathy, and foresight.

## 16.5 Conclusion and Future Work

This dialogue-driven exploration surfaced perspectives that move beyond conventional disciplinary boundaries. While we, as mediators, are not encyclopedic arbiters, several features of the process are worth noting:

- (1) *Structured depth.* The exchange progresses from breadth to depth: an initial set of ten perspectives is collaboratively distilled to three central themes, enabling deeper analysis and clearer inter-topic couplings.
- (2) *Unknown knowns via analogical framing.* When question articulation is impossible, conventional retrieval fails. We observe that prompting in domain A while requiring reasoning through frames from domains B/C can surface “unknown knowns”—knowledge present in the corpus yet resistant to direct querying—and yield candidate hypotheses for validation.

- (3) *Question quality as a driver.* The most productive turns arise from incisive, cross-frame questions. GPT-class agents contribute by posing probes that reindex the search space (e.g., causal, counterfactual, or mechanistic prompts), catalyzing novel connections rather than merely retrieving facts.
- (4) *Provenance and coherence.* Cross-referencing named works (stories, books, films) against sources included in model training or public catalogs improved factual precision and helped confirm that analogical links were thematically coherent rather than associative drift.
- (5) *Rational analogies, not ornament.* Film parallels, cultural references, and literary allusions are most useful when they carry structural isomorphisms (roles, constraints, payoffs) that map back to the target problem, widening interpretability without diluting rigor.
- (6) *Modern relevance.* Re-reading an age-old narrative (Adam and Eve) through contemporary technological and environmental frames demonstrates how polydisciplinary translation can render classical motifs decision-relevant today.

### **Limitations and Methodological Challenges**

Despite encouraging signals, several constraints limit definitive claims about effectiveness:

- (1) *Novelty verification.* Distinguishing genuine discovery from sophisticated recombination remains hard; community-level expert review is needed to establish originality.
- (2) *Topic generalizability.* A single narrative case study cannot certify performance in technical or high-stakes domains (e.g., materials, biomedicine, policy).
- (3) *Mediator influence.* Human choices in topic selection, prompt framing, and adjudication can bias trajectories and perceived quality.
- (4) *Reproducibility.* Stochastic variation across runs and models raises the question of whether comparable insights recur under controlled seeds and protocols.
- (5) *Depth–breadth tension.* Interdisciplinary reach may outpace domain-specific depth relative to subject-matter experts.

### **Methodological Refinements**

Targeted improvements can strengthen validity and robustness:

- (1) *Expert validation protocol.* Pre-register criteria and convene domain reviewers to rate correctness, novelty, and utility beyond internal assessment.
- (2) *Analogical prompt design.* Standardize cross-domain interrogation (source-target frame selection, mapping constraints, failure checks) to make “unknown known” discovery replicable rather than ad hoc.
- (3) *Comparative model analysis.* Run parallel dialogues across differing architectures and scales to expose model-specific biases and triangulate stable insights.
- (4) *Structured human–AI roles.* Minimize mediator bias via fixed templates for role assignment, intervention thresholds, and debate stop rules while preserving expert oversight.
- (5) *Automated quality checks.* Add computable measures for coherence, redundancy, and reasoning depth; employ critical-question lists (e.g., Socratic probes) and argument-structure scoring [Chang 2023a, 2023b].
- (6) *Provenance and verifiability.* Require source citations or derivations for claims; track anchors and transformations to support downstream auditing.
- (7) *Cross-domain testbed.* Evaluate on a suite spanning technical engineering tasks, scientific hypothesis generation, and interpretive analysis to map generalizability limits.

### **Implications for AI-Assisted Discovery**

The results suggest three practical implications. First, *collaborative multi-agent dialogue* can overcome some limits of single-agent reasoning by exposing complementary frames and controlled dissent. Second, *human–AI partnership* is most effective when humans specialize in framing—especially analogical framing for inarticulable questions—and in validation, while AI systems perform broad synthesis and mapping. Third, *structured exploration* (roles, schedules, provenance) helps keep analogical leaps disciplined, turning polydisciplinary breadth into testable hypotheses rather than speculation.

### **Future Directions**

- (1) *Metrics for discovery.* Develop quantitative measures for novelty, coherence, transfer quality (source→target mapping fidelity), and practical value.
- (2) *Multi-agent optimization.* Study role portfolios, contention schedules, and intervention policies that maximize hypothesis yield without sacrificing accuracy.

- (3) *Domain applications.* Trial the workflow in materials design, drug discovery, and engineering, where cross-domain analogies often precede breakthroughs.
- (4) *Unknown-known benchmarks.* Create tasks where ground-truth relations exist but are hard to query directly, to evaluate analogical interrogation systematically.
- (5) *Longitudinal evolution.* Track how candidate hypotheses mature across sessions, from initial analogical surfacing to empirical vetting and refinement.

In sum, Polynthesis, domain-agnostic representations enable LLMs to operate as hypothesis generators rather than answer engines. By combining analogical interrogation with verifiable provenance and structured collaboration, we can expose and test “unknown knowns” that would otherwise remain latent, accelerating discovery across disciplinary frontiers.

## References

- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. DOI: <https://doi.org/10.48550/arXiv.2303.12712>.
- E. Y. Chang. 2023a. CoCoMo: Computational consciousness modeling for generative and ethical AI. DOI: <https://doi.org/10.48550/arXiv.2304.02438>.
- E. Y. Chang. March. 2023b. CRIT: Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE, 351–360. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. December. 2023c. Examining GPT-4’s capabilities and enhancement with SoeraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, USA, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.



## CHAPTER 17

---

# Aphorisms for Collaborative Intelligence

### Abstract

This chapter distills five years of work into 16 aphorisms that ground MACI (Multi-LLM Agent Collaborative Intelligence) as a disciplined alternative to scaling single models. MACI treats intelligence as dialogic, evaluative, and regulative: heterogeneous LLMs are orchestrated to reason, plan, and decide under verifiable, context-aware constraints. The aphorisms are organized in four layers that move from foundations of truth and validation, through system-level design principles and collaborative reasoning strategies, to mechanisms of emergent intelligence. While informed by prior frameworks (CoCoMo, UCCT, SocraSynth, CRIT, EVINCE, DIKE-ERIS, SagallLM, ALAS, Polynthesis), references are illustrative; the emphasis is the shared design philosophy they imply. Together, the aphorisms offer a compact blueprint for building reliable multi-agent systems and a pragmatic route toward auditable, general-purpose intelligence.

### 17.1 Introduction

This chapter introduces 16 aphorisms developed from five years of research into multi-agent Large Language Models (LLMs). These aphorisms explain and justify the foundations of MACI (Multi-LLM Agent Collaborative Intelligence), an architecture that organizes multiple LLMs to support reasoning, planning, and decision-making beyond the capacity of single models. Grounded in philosophical traditions and implemented through the modular frameworks introduced earlier, MACI views intelligence as dialogic, evaluative, and regulative. Rather than scaling individual models in isolation, MACI coordinates collaboration across diverse perspectives to enable verifiable and context-aware behavior, offering a potential path toward AGI.

The 16 aphorisms are organized into four conceptual layers that reflect MACI's development from theoretical insights to practical orchestration. These include

foundational concepts about truth and validation (Aphorisms #1–#4), system-level design principles (Aphorisms #5–#8), collaborative reasoning strategies (Aphorisms #9–#12), and emergent intelligence mechanisms (Aphorisms #13–#16).

Although grounded in earlier frameworks including CoCoMo (Computational Consciousness Model) [Chang 2023a], UCCT (Unified Cognitive Consciousness Theory) [Chang 2025], SocraSynth [Chang 2023c], CRIT (Critical Reading Inquisitive Template) [Chang 2023b], EVINCE (Entropy and Variation in Conditional Exchanges) [Chang 2024b], DIKE–ERIS [Chang 2025], SagaLLM [Chang and Geng 2025b], ALAS (Adaptive LLM Agent System) [Chang and Geng 2025a], and Polynthesis, this chapter refers to them only when illustrative. The focus is on the design philosophy they collectively support: that collaborative intelligence, not isolated pattern generation, is the basis of general intelligence.

Each aphorism encapsulates a principle grounded in empirical system development, from questions about epistemology and bias to applied mechanisms for role assignment, linguistic behavior modulation, and validation. Together, they offer a comprehensive framework for understanding and building multi-agent systems.

The 16 aphorisms are grouped by conceptual layer, reflecting MACI’s progression from theoretical principles to system design and emergent intelligence. Table 17.1 summarizes their organization.

## 17.2

### 17.2.1 List of 16 Aphorisms

#### Aphorism #1: Truth Emerges from Perspectives

*Outside formal systems, objectivity remains a hard problem; what we pursue instead is reasonableness through multiple perspectives.*

Mathematics and logic offer certainty through formal proofs, but most real-world domains, such as medicine, law, and journalism, operate under uncertainty and competing interpretations. In these settings, objective truth is often inaccessible. What we can instead pursue is reasonableness, a standard grounded in coherence, evidence, and comparative judgment.

Philosophical traditions have long wrestled with this challenge. Kant distinguished between phenomena (what we perceive) and noumena (things-in-themselves), underscoring how cognition filters all experience [Kant 1781]. Nietzsche advances this further through his *philosophy of perspectivism*, arguing that “there are no facts, only interpretations.” In his view, all truth claims are inherently tied to particular viewpoints [Nietzsche 1968].

Thomas Nagel, in his book *The View from Nowhere*, provides a critical insight: objectivity is not the elimination of perspective, but rather the capacity to

**Table 17.1** Organizational map of the 16 aphorisms of MACI

Category	Aphorisms
<b>Theoretical Foundations</b>	#1: Truth emerges from perspectives #2: External mirrors enable validation #3: LLMs simulate reasoning and feelings, just as we do #4: Foundations and adaptations
<b>System Design</b>	#5: Context transforms capabilities #6: Linguistic behavior reflects intention and emotions #7: Checks and balances ensure adaptive alignment #8: Consciousness filters impulse; MACI governs LLMs
<b>Collaborative Dynamics</b>	#9: The essence of precise questioning #10: Debate strengthens reasoning quality #11: Hallucinations fade under scrutiny #12: Polydisciplinary synthesis expands discovery
<b>Emergent Intelligence</b>	#13: Human heuristics are limited, and end-to-end learning shows supremacy #14: AGI emerges through collaborative intelligence #15: Choose the baits, name the fish you will get #16: No memory, no regret; no regret no learning

synthesize multiple perspectives into a more impartial whole [Nagel 1986]. This concept closely aligns with how MACI achieves reasonableness by coordinating perspectival agents. Together, these perspectives suggest that truth emerges not as a static entity to be discovered but as a dynamic process unfolding through the interplay of diverse viewpoints. This view echoes the structure articulated in Hegel's dialectical method<sup>1</sup> [Hegel 1807] (further explored in Aphorism #10).

These philosophical insights have direct implications for practical domains where even the so-called ground truth may embed uncertainty or institutional bias. A Johns Hopkins study published in 2023 [Newman-Toker et al. 2024] estimated a 10% misdiagnosis rate in U.S. hospitals, revealing the limits of clinical certainty. Similarly, studies of media bias show how divergent expressions influence public understanding [Chang 2024c].

To navigate these uncertainties, MACI fosters dialogue among diverse reasoning agents. Each agent contributes a different perspective, evaluating or reframing claims. Through structured interaction and principled disagreement, the system

---

1. Hegel's dialectical method proceeds through a triadic structure: a *thesis* gives rise to its *antithesis*, and the tension between them is resolved in a *synthesis* that preserves aspects of both. This dynamic underpins the evolution of thought and understanding.

refines its understanding. Reasonableness emerges not from any single voice, but from how multiple voices interrogate and support each other, a dialectical process of negotiated understanding rather than definitive truth.

Rather than resolving the philosophical problem of objectivity, MACI recognizes its irresolvability and builds upon it as a structural design principle. Through its internal evaluation methods, the system assesses arguments for coherence, evidentiary grounding, and logical structure, measuring how well each claim withstands scrutiny from alternative perspectives. By treating disagreement as a resource rather than a failure, it moves closer to reasonableness—an ideal more closely aligned with how humans actually make sense of the world.

**Note:** This aphorism frames the epistemological basis for MACI; Aphorism #10 builds on this by detailing how agentic debate operationalizes dialectical reasoning.

### 17.2.2 Aphorism #2: External Mirrors Enable Validation

*No system can fully validate its own reasoning from within, a fundamental limitation shared by monolithic LLMs and human minds, as Gödel's incompleteness theorems reveal.*

Both humans and machines face a core limitation: we struggle to detect our own errors. People often miss inconsistencies or typos in their own writing, but spot them easily in others'. Language models face a similar challenge. When operating in isolation, a model cannot reliably assess the soundness of its own output.

This challenge has technical underpinnings. Researchers have identified that the stochastic nature of next-token prediction, conditioned on evolving context, can lead to degradation in reasoning quality. Effects such as effective long-context loss [Liu et al. 2024], attention narrowing and distraction [Zhu et al. 2025], and context override through state transitions [Yen et al. 2024] can cause the model to diverge from earlier logic or contradict prior statements without awareness. Without an external reference point, these failures often go uncorrected.

This limitation is mirrored in foundational theories. Gödel's incompleteness theorems show that complex systems cannot establish all truths from within themselves [Goffman 1959]. Wittgenstein emphasized that meaning arises from public language practices, not internal solipsism [Wittgenstein 1953]. And Popper noted that science advances through external falsification, not internal confirmation [Popper 1959].

Together, these insights point to a common conclusion: Validation requires independence. Sound reasoning is rarely self-contained. It emerges from critique, comparison, and testing across different viewpoints.

MACI addresses this constraint through a distributed validation system. Rather than relying on self-consistency, MACI enables agents to serve as external mirrors for each other. Each model contributes both its own judgment and its evaluation of others. Reasoning becomes more stable and accountable when multiple agents examine claims from independent points of view.

Progress in AI may depend less on increasing internal complexity and more on structuring systems that can challenge, reflect, and revise through collaboration. MACI provides the architecture to support this, making validation not a post-processing step, but an integral function of interaction.

Attaining greater intelligence may not require larger models, but better mirrors.

### 17.2.3 Aphorism #3: Simulated Reasoning Can Be Useful, Even if Not Real

*An actor may not feel true sorrow when portraying grief, yet their performance can still move us. The question is not whether the emotion is real, but whether it functions in context.*

Language models do not reason or feel the way humans do. They have no subjective experience, emotional memory, or intent. Yet they can often produce outputs that appear reasoned or emotionally expressive. This gap between form and inner state raises important questions about what it means to simulate intelligence.

Turing's 1950 formulation suggested that indistinguishability in conversation might be sufficient to attribute intelligence [Turing 1950]. Although this view has been criticized, most notably by Searle's Chinese Room argument, it nevertheless shifted the focus to behavior over internal structure. In practice, the utility of a system often depends not on how it works internally but on how its outputs function in a broader interaction.

The analogy to acting serves to illustrate this distinction. A skilled actor does not experience the grief that they portray, yet their performance may be compelling and meaningful. Of course, actors draw on lived emotional experience and possess intentionality. LLMs do not. The analogy holds not in terms of internal motivation but in how we evaluate the product of a performance: by its coherence, appropriateness, and effect.

Likewise, Erving Goffman's theory of social performance described how people change roles based on context and audience expectations [Gödel 1967]. Humans adapt strategically and with awareness. LLMs lack such metacognitive control. But their outputs can still be modulated to adopt different roles, such as analyst, critic, or teacher.

This is where MACI provides architectural clarity. It does not treat LLMs as conscious agents. Instead, it uses their capacity for simulated role enactment within

a structured environment. Roles are assigned, perspectives are coordinated, and outputs are checked by other agents. This process does not generate genuine understanding, but can still support useful reasoning behavior when mediated by architectural safeguards.

Simulation, in this view, is not deception if it is constrained, transparent, and subject to verification. MACI's purpose is not to pass as human, but to leverage the expressive potential of LLMs in a way that supports deliberation, critique, and adaptive planning.

Treating simulated reasoning as real reasoning would be a mistake. But rejecting simulation outright would ignore its functional value, especially when structured through external moderation and multi-agent feedback. MACI acknowledges this distinction and builds on it. The intelligence that emerges is not located in any single model, but in the interaction between them.

#### **17.2.4 Aphorism #4: Foundations and Adaptations**

*Intelligence operates on dual layers: a data-intensive foundation akin to unconscious processes and an agile contextual layer resembling conscious adaptation.*

Critics often note that LLMs require vast amounts of training data, while humans can learn from only a few examples. Yet, this contrast obscures a deeper structural parallel: both biological and artificial intelligence operate across two interdependent layers, one foundational, the other adaptive.

In humans, the foundational layer includes perceptual and emotional responses shaped by evolution and habit. Freud describes this as the unconscious mind—a reservoir of impulses beneath awareness [Freud 1915, Eagleman 2011]. Kahneman reframes this as “System 1” thinking: fast, automatic, and intuitive [Kahneman 2011]. When reflection is needed, humans engage “System 2,” in a slower, deliberative process capable of overriding instinct and reasoning through complexity.

Similarly, LLMs are trained on massive corpora to build a statistical foundation of linguistic fluency. This enables rapid generation in familiar contexts. However, when faced with novel questions, ambiguity, or conflicting interpretations, models cannot rely solely on their internal training. Their behavior must be shaped externally, through validation, feedback, role constraints, or collaborative feedback.

MACI does not introduce a third layer, but formalizes and activates the adaptation layer through structured, multi-agent coordination. It adds mechanisms for critique, counterargument, and revision, not by modifying weights but by orchestrating interaction among agents with complementary roles. Through moderated

dialogue and perspective shifting, MACI constructs a deliberative framework atop static foundation models.

This dual-layer design helps resolve the apparent contradiction between data-hungry models and human-like flexibility. Neither humans nor machines reason from scratch; both operate by refining deeply embedded patterns in response to context. Although their mechanisms differ, the functional analogy remains instructive.

Understanding this layered architecture reframes the debate: intelligence is not a monolithic capacity, but a dynamic interplay between entrenched knowledge and context-sensitive adjustment. MACI's distinctive contribution is to preserve the generative efficiency of foundational LLMs while adding structured mechanisms: critique, role-switching, verification, and synthesis that elevate adaptation to the level of collective deliberation.

**Note:** This aphorism establishes the foundational architectural insight behind MACI: intelligence arises from the interplay of a pretrained, data-driven substrate and a context-sensitive adaptive layer. It sets the stage for Aphorism #5, where this adaptation layer becomes a programmable mechanism through contextual modulation, and for Aphorism #8, which interprets this same layer as the functional analogue of consciousness, regulating behavior and intention. Together, these aphorisms define MACI's layered approach: preserving the efficiency of static models while enabling dynamic, deliberative reasoning through structured interaction.

### 17.2.5 Aphorism #5: Context Transforms Capabilities

*Strength and weakness in an LLM are not fixed traits, but fluid, shifting with context. MACI enables models to transcend training biases and adopt new roles through structured interaction.*

Language models are often evaluated in terms of static competencies. They are said to be “good” at summarization, “weak” at logical inference, or prone to certain biases. However, this evaluation misses a deeper principle: *LLM behavior is context-sensitive*. The same model can produce divergent results, depending on how it is prompted, instructed, or situated within a task.

This flexibility arises from the way LLMs generate responses. Without stable goals or persistent beliefs, they respond to cues embedded in context. Although this can lead to inconsistency or overfitting to recent tokens, it also allows *contextual modulation*, a process that alters behavior not through retraining but by shaping the context.

MACI turns this property into a core design principle. It treats context not as a constraint but as a *controllable layer of adaptation*. Through role assignment,

dialogic structuring, and structured disagreement, MACI coordinates agents to respond differently than they would in isolation. In this sense, context becomes a functional *design surface*, which not only directs behavior but also configures the reasoning dynamics of the system itself.

Three principles illustrate this transformation:

- (1) *Bias is contextual.* A model that defaults to caution may suppress novel ideas in exploratory tasks but increase safety in medical settings. MACI manages these tradeoffs by assigning complementary roles and balancing behavior through interaction.
- (2) *Adaptation is interactional.* LLMs change tone, focus, and strategy when prompted to critique others, defend a viewpoint, or explore counterfactuals. MACI leverages these dynamics to increase the diversity of reasoning and improve convergence through dialogue.
- (3) *Capability is conditional.* Instead of retraining models for fixed competencies, MACI reuses the same model in multiple roles: analyst, skeptic, reconciler, each invoked by a distinct interaction pattern or instruction format. This modular role-switching reveals capacities that remain latent in static use.

These behavioral shifts are measurable. Prior work demonstrates that adjusting contentiousness levels, argumentative tone, and emphasis via role assignments directly influences the quality and diversity of reasoning [Chang 2023c]. Measures of information theory, such as mutual information and entropy reduction, quantify the transition from disagreement to consensus [Chang 2024b].

For example, in a medical triage task, two LLM agents can begin with divergent diagnoses: one emphasizes common symptoms, while the other considers rare but high-risk causes. They engage in moderated debate across multiple rounds. As the dialogue progresses and context accumulates, each agent adapts. By the final round, they converge on a mutually defensible diagnosis that neither would have reached alone. This capability is not integrated into either model; it emerges from the way MACI configures the interaction.

This reframes the notion of “capability.” It is no longer a fixed attribute of the model, but a *systemic outcome* shaped by contextual design. MACI assumes that intelligence emerges not from perfect individual models, but from how imperfect ones are placed into a structured, adaptive system.

Aphorism #5 builds on the structural insight of Aphorism #4 and draws its philosophical grounding from Aphorism #8. Both describe the two-layer model of intelligence: a foundation layer (the pretrained model) and an adaptation layer (external modulation through context, roles, or prompts). Here, the adaptation

layer puts into action, not by altering internal weights but by shaping behavior through contextual control and agent coordination. What the model does depends not only on what it has learned, but also on how the system chooses to engage it.

**Note:** This aphorism establishes contextual modulation as a central mechanism within MACI. It puts into action the adaptation layer introduced in Aphorism #4 and provides the functional foundation for Aphorism #6 (linguistic role modulation), Aphorism #7 (ethical alignment through institutional checks), and Aphorism #10 (debate-driven refinement). These ideas are philosophically grounded in Aphorism #8, which frames the adaptation layer as a regulatory analog of consciousness: filtering impulses into intentions and shaping outputs for public reasoning. Intelligence, in this light, is not fixed, but fluid, shaped by design, role, and interaction.

### 17.2.6 Aphorism #6: Linguistic Behavior Reflects Intentions and Emotions

*LLMs are designed and trained to emulate human linguistic endeavors, each aimed at fulfilling distinct human objectives.*

Human language is rarely neutral; we speak and write with intention—to argue, explain, persuade, or connect. LLMs, trained on such purpose-driven language, inherit these communicative patterns. Although they lack intrinsic emotions or beliefs, they can emulate the rhetorical strategies and discursive forms associated with specific roles.

Consider these role archetypes:

- *Journalist:* Presents coherent, factual narratives;
- *Debater:* Constructs and defends arguments;
- *Analyst:* Dissects problems and identifies patterns;
- *Mediator:* Synthesizes divergent viewpoints.

MACI assigns such roles deliberately, using structured prompts and interaction constraints to shape agent behavior. For example, in multi-agent debates, contentiousness is modeled by adjusting tone, emphasis, and word choice [Chang 2023c].

Beyond contentiousness, MACI also modulates emotional expression. Research [Chang 2024a] has shown that LLMs can be conditioned to generate outputs across affective spectrums, validated through both linguistic features and visual representations, as presented in Figure 17.1.

This capacity to control linguistic behavior across both rhetorical and emotional dimensions allows MACI to orchestrate agent interactions in a goal-directed



**Figure 17.1** A lady in a garden scene under different emotions. From top-left, happiest, to bottom-right, saddest.

manner. The system uses language not only as output but also as a medium of coordination, assigning roles that reflect task demands and tuning the expressive style to support adaptive, collaborative reasoning.

**Note:** Aphorism #6 builds on the principle of contextual modulation introduced in Aphorism #5 by showing how linguistic behavior becomes the primary surface for adaptation. It provides the expressive substrate for enabling ethical alignment through role-based framing (Aphorism #7). These behaviors are further synthesized in Aphorism #10, where structured dialogue among diverse roles drives the quality of the reasoning. In MACI, language is not just the output; it is the interface of intention, strategy, and convergence.

### 17.2.7 Aphorism #7: Checks and Balances Ensure Adaptive Alignment

*Separating knowledge discovery, ethical oversight, and behavioral evaluation into distinct roles ensures a system of checks and balances, promoting adaptable AI safety and alignment with cultural norms.*

MACI implements a tripartite structure for adaptive alignment (Figure 17.2), drawing inspiration from constitutional governance [Chang 2025]:

- **Executive:** Generates knowledge, hypotheses, and solutions.



**Figure 17.2** Framework with three independent components. Bottom: executive LLMs; left: legislative (DIKE); and right: judicial (ERIS).

- **Legislative (DIKE):** Establishes ethical principles and constraints.
- **Judicial (ERIS):** Conducts a case-specific review and adversarial critique.

This separation avoids the pitfalls of reward-based fine-tuning approaches such as Reinforcement Learning from Human Feedback (RLHF), which often conflate performance and ethics by directly altering model parameters. In contrast, MACI isolates knowledge production from value imposition. Ethical reasoning is externalized, not embedded in a static policy, enabling more transparent and interpretable adjustments.

Furthermore, by treating each instance independently through the judicial branch, MACI enables context-sensitive ethical alignment, rather than applying averaged constraints that fail to account for individual situations and cultural or societal variations.

Moreover, by evaluating the alignment case by case through the judiciary branch, MACI adapts to the context and culture rather than applying averaged constraints. A suitable solution in one setting can be flawed in another; MACI embraces this diversity by treating disagreement as a necessary mechanism for refinement, not failure.

**Note:** This aphorism builds on Aphorism #5 (contextual modulation) and Aphorism #6 (role-based linguistic behavior) by formalizing role separation for ethical alignment. It serves as a structural pillar for MACI's governance model, complementing Aphorism #8, which frames intention regulation as a system-level control layer. Together, these mechanisms support Aphorism #10's collaborative reasoning and Aphorism #11's dissipation of hallucination through adversarial validation. In MACI, alignment is not imposed; it is deliberated.

### 17.2.8 Aphorism #8: Consciousness Filters Impulse; MACI Governs LLMs

*Our public behavior is not a direct, unfiltered output from our unconscious mind. Instead, consciousness regulates and refines the underlying impulses, ensuring that our actions are aligned with social norms. Similarly, MACI frameworks harness and temper the inherent tendencies of LLMs, mitigating inherited biases.*

Philosophers and psychologists describe consciousness as a regulatory system that filters impulsive or inappropriate actions before they reach the public sphere. Freud characterizes the ego as a mediator that suppresses and channels the raw desires of the id [Freud 1923]. Kant views reason as a regulator of inclination, aligning behavior with duty and rational constraints [Kant 1788]. Sartre argues that consciousness introduces delay and reflection, inserting “nothingness” between stimulus and action [Sartre 1943]. Across these views, consciousness acts as a gatekeeper, transforming a raw impulse into socially and ethically acceptable behavior.

MACI performs an analogous function. Although it does not implement consciousness, it regulates the outputs of LLMs using contextual, structural, and ethical constraints. Neuroscientific models such as Gazzaniga’s “interpreter” theory [Gazzaniga 2011] and Baars’ Global Workspace Theory [Baars 2005] also frame consciousness as an aggregator and filter of lower-level cognitive processes. In line with these ideas, MACI applies higher-order control mechanisms to shape model behavior through:

- *Inhibitory control:* Suppressing harmful, incoherent, or inappropriate content.
- *Reappraisal:* Reframing outputs based on ethical, epistemic, or pragmatic considerations.
- *Contextualization:* Adjusting tone, abstraction, and intent based on task demands and audience.

This architecture reflects the functionalist modeling approach presented in our CoCoMo work [Chang 2023a]. CoCoMo treats consciousness not as mere

awareness or qualia, but as an executive control system that coordinates modular reasoning. MACI adopts this perspective explicitly: intention regulation governs when and how LLMs generate language suitable for deliberation, decision-making, and social exchange.

**Note:** This aphorism establishes the **regulatory foundation** of MACI, drawing from philosophical, neuroscientific, and architectural perspectives on cognitive control. Like consciousness, MACI inserts deliberation between thought and action, ensuring that outputs reflect structured goals rather than impulsive inference. It provides the conceptual basis for Aphorism #5 (contextual control), Aphorism #6 (behavioral modulation), and Aphorism #7 (ethical alignment through institutional checks); and it prepares the ground for collaborative refinement through interaction, debate, and convergence in Aphorisms #9–#11. Intelligence, in this view, is not measured solely by what it produces, but by how it governs what it allows to emerge.

### 17.2.9 Aphorism #9: The Essence of Precise Questioning

*The essence lies in framing and sequencing the right questions.*

This aphorism underscores the foundational role of precise questioning in multi-agent LLM systems. The approach draws on the Socratic tradition, where knowledge does not arise from assertion but from systematic inquiry [Vlastos 1991]. Within MACI, particularly through the SocraSynth framework [Chang 2023c], three key principles structure this process.

First, the quality of multi-LLM dialogue hinges on how effectively agents interrogate one another. Strong counterarguments function as probing questions: they surface hidden assumptions, demand justification, and expose logical tensions. In this sense, the Socratic elenchus becomes a dialogic protocol: not to defeat, but to discover.

Second, iterative exchanges sharpen intent over time. Early questions may be exploratory, even vague, but as context accumulates, agents refine their inquiries with increasing precision. Each response becomes a scaffold for the next query, creating a feedback loop of progressive clarification.

Third, beyond formulation, the sequencing of questions becomes crucial. Ordering determines the scope and relevance of subsequent investigation. In MACI, agents use prior turns to dynamically plan next-step inquiries, allowing structured traversal across complex domains. This logic mirrors hypothesis refinement in scientific inquiry, starting with exploratory breadth and then narrowing through informed follow-up.

This virtuous cycle—precise questioning, contextual enrichment, and adaptive sequencing—drives the depth and utility of MACI’s reasoning. For example, in medical classification, a discussion that begins with the enumeration of symptoms can evolve into causal analysis, exploration of comorbidities, and synthesis of treatments [Chang and Chang 2023b, Chang 2024b]. Here, a human moderator adjusts the agents’ linguistic stances, from contentious to conciliatory, using information-theoretic signals such as cross-entropy, mutual information, and Jensen–Shannon divergence to optimize epistemic progress.

This virtuous cycle of precise questioning, contextual enrichment, and strategic sequencing determines the quality and utility of the insights generated by LLMs.

**Example:** A conversation that begins with a question about the symptoms of a disease can evolve into a diagnostic discussion as LLMs build on each other’s insights. The discussion transitions from identifying surface-level symptoms to uncovering underlying causes, recognizing comorbidities, and generating actionable treatment recommendations [Chang and Chang 2023b, Chang 2024b]. To facilitate effective sequencing, moving from exploration of possible diseases to deeper probing of finalists, the conversation moderator adjusts the LLMs’ linguistic behaviors, shifting from contentious to conciliatory tones based on several information-theoretic metrics, including cross-entropy, mutual information, and Jensen–Shannon divergence.

**Note:** Aphorism #9 serves as a cognitive engine for the MACI framework. It underpins Aphorism #6 by guiding how linguistic roles shape intention, and supports Aphorism #10 by structuring debate through layered, clarifying inquiry. It enables Aphorism #11 by progressively narrowing the hallucination space, and lays the groundwork for Aphorisms #12 and #13 by prompting the kinds of questions that foster cross-domain synthesis and emergent learning. Well-sequenced questioning, in this light, becomes the scaffolding for adaptive reasoning across the system.

In short, precise questioning is not just a technique; it is the substrate for coordination, divergence, synthesis, and convergence in the collaborative intelligence of MACI. It determines what becomes knowable.

### 17.2.10 Aphorism #10: Debate Strengthens Reasoning Quality

*Critical thinking requires more than one Socrates.*

The power of collaborative reasoning has deep philosophical roots. While Socrates transformed thinking through systematic questioning, he also recognized the necessity of dialogue partners. As Aristotle formalizes, dialectic requires both thesis and antithesis to approach synthesis [Aristotle 1928]. This principle finds

one of its most profound expressions in Hegel’s dialectical process, where knowledge advances not through linear accumulation but through the *productive tension* (SocraSynth’s “contentiousness” [Chang 2023c]) of opposing ideas [Hegel 1807]. In Hegel’s framework, an initial position (thesis) confronts its negation (antithesis), producing a higher-order understanding (synthesis) that transcends yet preserves both. This dialectical movement mirrors how MACI orchestrates productive disagreement to generate deeper insight.

First, in dialectical reasoning, each LLM serves both as a questioner and as a respondent. Like Socrates engaging with interlocutors, one LLM challenges assumptions while another defends or refines them. This dynamic creates what philosopher Hannah Arendt calls the “in-between space,” where meaning emerges from the interaction itself [Arendt 1958]. Each exchange builds upon prior reasoning, refining hypotheses through mutual scrutiny. This process exemplifies Hegel’s concept of *Aufhebung* (sublation), where opposing viewpoints are not only compromised but also transformed into a qualitatively new understanding.

Second, meaningful collaboration requires a *baseline* level of competency. Just as Socrates could not derive insight from those lacking reasoning capacity, MACI cannot produce valuable outcomes when the participating models are fundamentally weak. Two poor reasoners do not combine into one strong one; their limitations may compound rather than cancel out. Fortunately, with the rise of advanced LLMs, we have crossed a critical competency threshold, enabling collaborative reasoning that can identify and correct individual model limitations. Moreover, as Aphorism #11 demonstrates, hallucinations tend to dissipate in a well-orchestrated multi-LLM dialogue, reinforcing the value of structured collaboration.

Third, the depth of dialogue varies with the sophistication of the participants. Advanced LLMs (e.g., Claude 3.7, GPT-4o, and DeepSeek R1) can explore complex ideas in depth, leveraging diverse training and reasoning strategies to interrogate and challenge each other’s assumptions. In contrast, smaller models produce only shallow exchanges, like novices grappling with philosophical questions that they cannot fully articulate. Model diversity—in training data, architectural assumptions, and epistemic behavior—becomes a necessary condition for productive debate.

MACI creates a space for *AI philosophers* to engage in structured dialogue through systems such as SocraSynth and DIKE-ERIS, with EVINCE monitoring information flow and CRIT [Chang 2023b] evaluating the quality of reasoning. The strength of these debates depends not only on the frameworks themselves but on the capabilities and calibration of the participants. This understanding guides

both model selection and the realistic expectations we place on their collaborative outputs.

This approach suggests that the path to Artificial General Intelligence (AGI) may not lie in scaling individual models alone, but in orchestrating systems of capable, diverse agents engaged in structured debate, mirroring how human insight often emerges socially. As Karl Jaspers observed, “Truth begins between people” [Jaspers 1969], a principle that appears equally applicable to artificial minds seeking mutual understanding.

**Note:** This aphorism connects with Aphorism #1 (truth as perspectival synthesis), Aphorism #9 (the role of precise questioning), and Aphorism #11 (hallucination reduction through orchestration). Together, they illuminate how MACI turns disagreement into a productive engine of reasoning: where structured dialogue refines claims, filters errors, and advances collective understanding beyond the capacity of any single agent.

### 17.2.11 Aphorism #11: Hallucinations Fade Under Scrutiny

*Hallucinations rarely repeat.*

Have you ever wondered why recurring nightmares, though similar in theme, never unfold in exactly the same way? This property of human dreaming finds a parallel in the behavior of LLM hallucinations, which also resist exact repetition, even when the prompt remains unchanged.

This observation echoes long-standing insights from psychology and philosophy. Freud noted that unconscious repetition occurs not through identical images, but through symbolic variation and displacement [Freud 1900]. Clinical studies of hallucinations in schizophrenia and substance use disorders similarly reveal that while themes may recur, verbal or visual content almost always changes with each episode [Waters et al. 2014]. Phenomenologists such as Merleau-Ponty argued that perception is never static: each encounter is shaped by shifting contextual salience and embodied interpretation [Merleau-Ponty 1945].

This phenomenon also applies to LLMs. Hallucinations arise when the probabilistic token prediction generates unstable or unsupported sequences from ambiguous inputs. Because outputs are sampled from a probability distribution, similar prompts can produce different hallucinations. This distinguishes hallucinations from systematic errors, which recur due to model bias or data gaps.

In MACI frameworks such as SocraSynth [Chang 2023c, 2024b], this non-repetitive nature becomes an asset. When one agent produces a hallucination, others can challenge it. As dialogue evolves, the shared context becomes increasingly specific, constraining the “hallucination space.” With each round, it becomes

harder for the original agent to restate the same fabrication; instead, it must align with grounded knowledge or admit uncertainty. Through structured disagreement, hallucinations naturally dissipate.

In addition, hallucinations are sporadic, while true knowledge gaps are consistent. To distinguish the two, MACI integrates Retrieval-Augmented Generation (RAG), enabling access to verifiable sources. This allows the system to triage stochastic noise versus persistent ignorance and apply targeted remediation.

**Example:** In a multi-agent dialogue about historical events,  $LLM_A$  might initially hallucinate, claiming, “The Treaty of Versailles was signed in 1921.”  $LLM_B$  challenges this, stating, “The Treaty of Versailles was signed in 1919, following World War I.” On a second query,  $LLM_A$  is unlikely to repeat the same error—now grounded by the updated context. Even if it errs again, the mistake is likely to differ in form, making it detectable through further rounds of scrutiny [Manakul et al. 2023].

This understanding has implications for system design: hallucinations are not persistent constructs but unstable artifacts of uncertain inference. Rather than requiring static classifiers, MACI uses multi-agent dialogue to destabilize hallucinations through context evolution. What vanishes under scrutiny is not just the false claim, but its power to recur.

**Note:** This aphorism complements Aphorism #10 (debate strengthens reasoning quality) and Aphorism #2 (external mirrors enable validation). It shows that hallucinations fade not because they are explicitly detected, but because structured dialogue progressively eliminates unsupported claims. MACI transforms stochastic unreliability into epistemic refinement: hallucination becomes not a flaw but a transient signal that dissolves in the presence of discourse.

### 17.2.12 Aphorism #12: Polydisciplinary Synthesis Forges New Frontiers

*LLMs are not taught about domain boundaries, as they were trained only to predict the next words. This polydisciplinary approach to information representation allows LLMs to synthesize knowledge that might be beyond narrowly focused, domain-specific human understanding.*

The term *polydisciplinary* was introduced by Microsoft’s Chief Scientific Officer, Eric Horvitz, during a Stanford HAI (Human-Centered Artificial Intelligence) Symposium in 2023. He observed that GPT-4, trained via maximum likelihood estimation, applies the same statistical machinery whether processing quantum physics or Shakespearean sonnets: the model learns without internalizing disciplinary boundaries.

Although humans organize knowledge into categories such as physics, poetry, biology, and philosophy, LLMs move fluidly across these divisions, unaware of traditional boundaries. This polydisciplinary capacity opens new possibilities [[Chang and Chang 2023a](#)] (see Chapter 16):

- When an LLM perceives a pattern linking Shakespearean verse and quantum mechanics, it does not question whether such a connection is conceptually permissible.
- Complex problems that traditionally require interdisciplinary expertise become accessible to LLMs, which operate freely beyond academic silos.
- Novel hypotheses may emerge precisely because LLMs do not filter out connections conventionally deemed implausible or inappropriate by domain norms.

This boundary-free synthesis mirrors what cognitive scientists Gilles Fauconnier and Mark Turner call *conceptual blending*, a process by which novel insights emerge from the fusion of distinct conceptual spaces [[Fauconnier and Turner 2002](#)]. When asked to relate music theory to protein folding, an LLM proposed modeling protein sequences as melodies, enabling motif-based visualizations. When comparing architecture and immunology, it described the immune system as a “defensive architecture,” inspiring layered design analogies for system resilience. These examples illustrate how LLMs emulate the core structure of human creativity: not by refining within a single domain but by blending conceptual frameworks across them.

MACI harnesses this capability through frameworks such as SocraSynth [[Chang 2023c](#)], EVINCE [[Chang 2024b](#)], and the DIKE-ERIS dual model [[Chang 2025](#)]. Together, these tools help formulate meaningful questions (Aphorism #9), synthesize novel perspectives (Aphorism #10), and filter out incoherence (Aphorism #11). The result is a system that not only traverses known fields, but probes the “unknown unknowns” with epistemic discipline.

The essence of MACI lies in the navigation of interdisciplinary intersections, areas where genuine insight often emerges. These liminal spaces, naturally traversed by LLMs, are rich in potential but saturated with ambiguity: the domain of “unknown unknowns.” Humans may struggle to pose coherent questions or distinguish signal from noise in such terrain. In these cases, MACI encourages minimal human interference. Rather than steering the process, humans act as lightweight moderators, providing initial prompts, then stepping back to let the system explore freely and critically assess its own findings. This arrangement allows MACI to illuminate uncharted intellectual territory with autonomy, surfacing connections and

hypotheses that extend beyond conventional human foresight. Chapter 16 [Chang and Chang 2023a] offers an example that spans several unexpected knowledge domains, seeded by a single biblical narrative.

This capacity for polydisciplinary synthesis represents perhaps the most promising path toward AGI and beyond. True innovation requires the ability to forge new frontiers by perceiving novel connections across traditional boundaries, to see patterns invisible to specialists entrenched in disciplinary silos. Human innovation often requires rare interdisciplinary genius, and LLMs inherently operate beyond these constraints, potentially discovering insights and opening intellectual frontiers that might otherwise remain unexplored for decades. In this sense, the polydisciplinary nature of LLMs does not just enhance existing knowledge; it creates the conditions for genuine intellectual exploration into uncharted territories.

**Note:** This aphorism builds on the adaptive structures in Aphorism #5 and the intention-regulatory framing of Aphorism #8. It provides the epistemic ground for Aphorism #9 (question formulation) and the generative substrate for Aphorisms #10 and #11 (collaborative refinement). Polydisciplinary synthesis is not noise; it is a structured divergence that expands the frontier of *knowability*.

### **17.2.13 Aphorism #13: Human Heuristics Are Limited, and End-to-End Learning Shows Supremacy**

*When machines are freed from handcrafted features and human biases, they can often surpass what even expert-designed systems can achieve.*

Across several landmark AI domains, a recurring pattern has emerged: systems trained end-to-end—mapping raw inputs directly to outputs—often outperform those built with hand-crafted heuristics or modular pipelines. These breakthroughs illustrate the limitations of human-designed rules and the power of self-learned representations when provided with sufficient data, computational resources, and architectural capacity.

Consider AlphaGo versus AlphaGo Zero. The original AlphaGo, which defeated world champion Lee Sedol in 2016, relied on expert games and human-derived features [Silver et al. 2016]. In contrast, AlphaGo Zero learned solely from self-play via reinforcement learning. Within days, it surpassed its predecessor, revealing that human priors can restrict learning in highly expressive models [Silver et al. 2017].

A similar transition occurred in protein folding. AlphaFold v1 used co-evolutionary heuristics; AlphaFold 2 employed a fully differentiable architecture, predicting 3-D structure directly from amino acid sequences. It won CASP14 (14th

Critical Assessment of Protein Structure Prediction)—a recurring, community-wide competition—by a wide margin [Jumper et al. 2021], demonstrating how task-specific structure can emerge from data alone.

In autonomous driving, earlier systems were modular—perception, prediction, and planning were each engineered separately. But these pipelines suffered from cascading errors and brittleness in edge cases. Newer systems, such as Tesla’s end-to-end stack or Wayve’s policy networks, map directly from raw vision to actions, enabling more robust generalization [Bojarski et al. 2016, Xu et al. 2017].

In computer vision, Scale-Invariant Feature Transform (SIFT) was once the gold standard for hand-crafted geometric features [Lowe 2004]. Then came AlexNet, a deep convolutional network trained directly on pixels. Its performance on ImageNet rendered the heuristic approach obsolete almost overnight [Krizhevsky et al. 2012]. More than accuracy, the shift was about *adaptivity*: data-driven models [Liu et al. 2008, Chang 2011] can improve through feedback, whereas fixed pipelines cannot.

These examples converge on a central insight: human intuition, while useful for bootstrapping progress, imposes biases and boundaries. End-to-end learning removes these constraints, allowing models to learn optimal representations jointly across all layers of reasoning.

MACI embraces this insight in its reasoning framework. Although human rules shape safety and ethical constraints (see Aphorism #7), robust understanding arises when agents learn reasoning (Aphorisms #1 and #10), critique, and reflection through interaction. Frameworks such as CRIT [Chang 2023b] operationalize this: instead of coding decision rules, CRIT prompts agents to evaluate each other’s claims, sourcing credibility and logical strength from collective deliberation.

Just as organisms evolve not through static rule books but through continuous adaptation [Eagleman 2011, Clark 2013], intelligence in machines emerges not from rigid design but from learning by feedback [Lake et al. 2017]. End-to-end systems outperform not because they ignore structure, but because they build structure in service of adaptation.

**Note:** This aphorism complements Aphorism #12, where polydisciplinary synthesis emerges from LLMs’ boundary-free representations. Here, we see that eliminating rigid heuristics—whether domain-specific or architectural—frees learning to discover new forms of generalization. Together, these aphorisms argue for a model of intelligence rooted not in static human priors, but in fluid, adaptive exploration across conceptual, structural, and epistemic boundaries.

### 17.2.14 Aphorism #14: AGI Emerges Through Collaborative Intelligence

*The path to AGI lies not in singular models, but in systems that reflect, regulate, and reason together.*

AGI will not arise merely by scaling language models [Kaplan et al. 2020], nor by perfecting a monolithic architecture. Instead, it will emerge from *systems of collaboration* architectures in which multiple agents operate with distinct roles, epistemic lenses, and regulatory feedback loops. These systems must do more than generate fluent responses: they must interrogate their own reasoning, revise their conclusions, and reconcile divergent viewpoints in pursuit of coherence and truth.

Recent evidence from OpenAI’s 401 experiments shows that inference-time scaling—allocating tokens for deliberation or chain-of-thought—improves reasoning and planning performance with only linear cost increases. Table 17.2 contrasts the tradeoffs between training-time and inference-time scaling, suggesting that dynamic, task-specific reasoning may be more fruitful than ever-larger training runs.

This shift in perspective is central to MACI. In SocraSynth and EVINCE, agents pursue truth through Socratic dialogue. In CRIT, claims are evaluated for coherence and evidentiary strength. In DIKE-ERIS, moral and cultural norms are enforced via institutional-style checks. SagaLLM enables long-horizon planning by providing persistent memory, constraint validation, and workflow guarantees.

Yet such systems cannot function without a planner. Enter ALAS (Adaptive Layered Agent System), which orchestrates distributed planning and role assignment. Where SagaLLM ensures consistency, ALAS decomposes tasks, manages dependencies, and coordinates execution between agents operating in real-world domains such as ride-sharing, logistics, or collaborative robotics, domains where static LLM reasoning often fails.

These subsystems embody a larger truth: *no single model can model itself*, just as no mind can achieve objectivity without encountering other minds.

**Table 17.2 Comparison of training-time scaling vs. inference-time scaling in LLM performance**

Dimension	Training-Time Scaling	Inference-Time Scaling
<b>Cost scaling</b>	Exponential/superlinear	Linear (per output token)
<b>Accuracy gain</b>	Diminishing returns (esp. at scale)	High marginal gains on hard tasks
<b>Flexibility</b>	Fixed post-training	Adaptive (per input/task)
<b>Model size</b>	Must grow for gains	Fixed model
<b>Alignment to reasoning</b>	Limited	High (via thinking tokens, CoT, debate)

Where Aphorism #2 emphasized the need for external mirrors to validate outputs, Aphorism #14 transforms that insight into a design philosophy. MACI is not just a critique; it is a blueprint, with *modularity*, *reflectivity*, and *role specialization* as the pillars of general intelligence. This architecture suggests that AGI will not emerge from raw scale, but from systems that are *aware of their own fallibility* and structured to adapt through principled dialogue.

This view responds to foundational critiques (e.g., LeCun's), which cite LLMs' lack of grounding, memory, and planning. We agree if one expects these to emerge from a single architecture. Instead, we treat LLMs as a linguistic substrate—an artificial unconscious. Although they lack volition, they provide the generative and semantic flexibility upon which memory, control, and reasoning layers can be built.

Just as human intelligence arises not in isolation but through community, language, and institutional learning, **AGI may arise not from singular mastery, but from collaborative coherence.**

### 17.2.15 Aphorism #15: Choose the baits, name the fish you will get

*The bait you anchor with determines the fish you haul.*

#### Three lively illustrations.

- (1) *No bait: the unconscious prior.* You drop an empty hook straight down. Nothing signals intent. Whatever happens to be dense beneath the boat bites first. In LLM terms, with no Few Shots, no RAG, no roles, and no fine-tuning, the model reflects its high-probability manifold—maximum-likelihood completions shaped by pretraining priors. It is *unconscious reflection*, not guided inquiry.
- (2) *Bloody beef: summoning sharks.* Now you thread a ribbon of bloody beef and give the water a swirl—light chumming. If sharks are locally dense, their olfactory “embedding” aligns with your signal. They rise, crowd out gentler species, and you “discover” sharks. Analogously, strong, on-target anchors (Few Shots that share structure with the goal, RAG passages with tight semantic overlap) increase anchoring strength  $S$ , pulling the model toward the desired attractor basin. The semantics of your output become “shark.”
- (3) *Plankton net: lifting anchovies/sardines.* Swap the beef for a fine-mesh plankton net and a deck light. You are no longer calling predators; you are aligning to a different ecology. Small baitfish school to the bloom, and your net comes up silver with anchovies. In LLM terms, domain-specific retrieval

(e.g., curated mechanism papers), gentle Few Shots, and low-contention roles anchor toward fine-grained, factual synthesis rather than dramatic but sparse “predators.” The semantics of your output become “anchovy.”

### **Operational mapping (UCCT/MACI).**

- *Few Shot examples* = cut bait on hooks. Match structure, not surface. High-quality, near-target exemplars raise  $S$ .
- *RAG chunks* = chumming pattern. More is not always better; precise scent beats cloudy soup. Calibrate recall vs. precision.
- *MACI roles/schedules* = where you cast and how the crew works. High contentiousness explores wider waters; low contentiousness lands the current bite.
- *Fine-tuning* = moving the boat to new grounds. You shift the prior itself, not only the bait.

**Failure modes (*choose bait, choose semantics*)** Wrong bait, wrong fish: off-target Few Shots or noisy retrieval can confidently attract the *wrong* attractor (compelling but irrelevant analogies, “predators” that eat your time). Over-chumming (too many near-duplicates) collapses diversity: the model overfits a narrow pattern and misses adjacent schools. Under-baiting (one flimsy exemplar) leaves you in prior waters.

**Unknown knowns via analogical bait** When the question cannot be stated, you can still fish. Pose the problem in domain A, *but* require reasoning through frames from domains B/C (the “plankton light” for a different ecology). This cross-frame anchoring surfaces latent correspondences—“unknown knowns”—that the prior alone would not yield. Then verify the catch (provenance, tests).

**Takeaway** No bait → the prior. Right bait → the target basin. Choose the bait, and you implicitly choose the semantics.

#### **17.2.16 Aphorism #16: Memory Makes Minds; Regret Makes Progress**

*No memory, no regret. No regret, no learning.*

##### **Two dimensions of memory**

**Spatial memory enables perspectives** Multiple frames can be held concurrently, compared, and synthesized—supporting viewpoint diversity without context loss.

**Temporal memory enables retrospectives** Past choices and outcomes are traceable; mistakes become curriculum rather than repetition.

### Three lively illustrations

- (1) *The amnesiac navigator.* Starting every voyage from zero, they hit the same rocks. Today's brilliance, tomorrow's reset: the default stateless LLM.
- (2) *The retrospective pilot.* Near-misses and smooth landings alike are logged; “regret” calibrates future approach paths. Memory compounds into judgment.
- (3) *The spatiotemporal historian (SAGA).* The agent records where/when a decision was taken, what was tried, abandoned, or recovered. Faced with a new task: “In context  $Y$ , approach  $X$  underperformed because  $Z$ ; prefer  $W$ .” Memory operationalizes intelligence.

#### *Operational mapping (spatiotemporal state)*

- *Episodic store (space):* retain competing frames, artifacts, and partial plans; support synthesis without overwriting.
- *Decision log (time):* append choices, rationales, outcomes; include undo/redo stacks and counterfactual branches.
- *Regret signal:* post-decision loss differential vs. a baseline or counterfactual; use it to update selection probabilities and policies.
- *Counterfactual replay:* re-evaluate past states with alternative actions; harvest “lessons” without re-incurring costs.
- *Compensation strategies:* map failure modes to recovery playbooks; prefer fast, memory-guided pivots over full resets.

**Failure modes (beyond memorylessness)** Stale or hallucinated memory amplifies error; privacy-unsafe memory leaks trust; over-retention causes rigidity. Guard with TTLs, provenance checks, and decay.

**The regret gradient (clarification)** Here “regret” is informational, not emotional: a signal that an observed outcome under action  $a_t$  is worse than a baseline or counterfactual  $a'_t$ . Logging  $(s_t, a_t, o_t)$  enables credit assignment and policy updates; without memory, no gradient can persist across tasks.

**Takeaway** Spatial memory unlocks perspectives; temporal memory unlocks retrospectives. *The path from intelligence to wisdom runs through regret.*

## 17.3 Conclusion

The 16 aphorisms unfold in a layered progression that mirrors the architecture of MACI itself. Aphorisms #1–#4 establish the philosophical foundation: they frame truth as perspectival, highlight the necessity of external validation, and introduce

**Table 17.3** Aphorisms grouped by Thematic Pillars in MACI. An aphorism may be listed under multiple pillars

Epistemology and Perspective	Governance, Ethics, and Alignment
#1. Truth through perspectives	#2. External validation
#2. External validation	#7. Checks and balances
#9. Precise questioning	#8. Intention regulation
#10. Dialectical debate	#11. Hallucination dissipation
#12. Polydisciplinary synthesis	#16. No memory, no regret
#15. Semantic anchoring	<b>Synthesis Toward AGI</b>
#16. No memory, no regret	#10. Dialectical debate
Simulation and Role Enactment	#12. Polydisciplinary synthesis
#3. Simulated reasoning	#13. End-to-end learning
#5. Contextual modulation	#14. Collaborative AGI
#6. Linguistic behavior	#15. Semantic anchoring
#13. End-to-end learning	#16. No memory, no regret
Cognitive Architecture and Control	
#4. Foundation and adaptation	
#5. Contextual modulation	
#8. Intention regulation	
#13. End-to-end learning	
#15. Semantic anchoring	
#16. No memory, no regret	

a layered view of intelligence that blends simulation with structured reasoning. These insights motivate the architectural strategies developed in Aphorisms #5–#8, where contextual modulation, role-based behavior, and institutional checks define how LLMs are coordinated, constrained, and guided within collaborative systems.

Aphorisms #9–#12 then describe how intelligence operates within this framework: through strategic questioning, dialectical engagement, hallucination mitigation, and polydisciplinary synthesis. This middle stratum reflects the operational core of MACI, where the reasoning is not housed in individual agents, but arises through structured interaction.

Finally, Aphorisms #13–#16 widen the lens. They reject the notion that AGI will emerge solely through scale, and instead propose that general intelligence arises through modular cooperation, reflective regulation, persistent memory, and distributed coherence. Together, these aphorisms form more than a sequence—they constitute a system, where each conceptual layer scaffolds the next. From epistemic humility to architectural design and from behavioral control to intelligent emergence, they chart a coherent path toward collaborative artificial general intelligence.

**Aphorism Theme Summary** To illuminate the conceptual architecture of the MACI framework, Table 17.3 organizes the 16 aphorisms by thematic pillars. Each theme captures a core dimension of collaborative intelligence, ranging from epistemological foundations and behavioral simulation to cognitive regulation, ethical governance, and the synthesis required for AGI. Several aphorisms appear under multiple themes, reflecting their integrative nature. This organization provides a high-level map of how the components of MACI interact, align, and build toward a unified system of multi-agent reasoning and adaptive intelligence.

## References

- H. Arendt. 1958. *The Human Condition*. University of Chicago Press, Chicago.
- Aristotle. *Topics*. 1928. Translated by W. A. Pickard-Cambridge, originally composed circa 350 BCE. Oxford University Press, Oxford.
- B. J. Baars. 2005. Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Prog. Brain Res.* 150, 45–53. DOI: [https://doi.org/10.1016/S0079-6123\(05\)50004-9](https://doi.org/10.1016/S0079-6123(05)50004-9).
- M. Bojarski, D. Del Testa, D. Dworakowski, et al. 2016. End to end learning for self-driving cars. DOI: <https://arxiv.org/abs/1604.07316>.
- E. Y. Chang. 2011. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, Berlin. DOI: <https://doi.org/10.1007/978-3-642-20429-6>.
- E. Y. Chang. 2023a. CoCoMo: Computational consciousness modeling for generative and ethical AI. DOI: <https://doi.org/10.48550/arXiv.2304.02438>.
- E. Y. Chang. March. 2023b. Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE, 351–360. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. December. 2023c. Examining GPT-4’s capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang. August. 2024a. Behavioral emotion analysis model for large language models. In *Proceedings of the 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR '24)*, San Jose, CA, August 7–9, 2024. IEEE, 549–556. DOI: <https://doi.org/10.1109/MIPR62202.2024.00094>.
- E. Y. Chang. August. 2024b. EVINCE: Optimizing multi-LLM dialogues using conditional statistics and information theory. DOI: <https://doi.org/10.48550/arXiv.2408.14575>.
- E. Y. Chang. 2024c. Uncovering biases with reflective large language models. DOI: <https://doi.org/10.48550/arXiv.2408.13464>.
- E. Y. Chang. 2025. A checks-and-balances framework for context-aware ethical AI alignment. In *Proceedings of ICML*. Retrieved from <https://icml.cc/virtual/2025/poster/46461>.

- E. Y. Chang and E. J. Chang. 2023a. *Discovering Insights Beyond the Known: A Dialogue Between GPT-4 Agents from Adam and Eve to the Nexus of Ecology, AI, and the Brain*. Stanford InfoLab Technical Report.
- J. J. Chang and E. Y. Chang. December. 2023b. SocraHealth: Enhancing medical diagnosis and correcting historical records. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 1400–1405. DOI: <https://doi.org/10.1109/CSCI62032.2023.00229>.
- E. Y. Chang and L. Geng. 2025a. ALAS: A stateful multi-LLM agent framework for disruption-aware planning. DOI: <https://doi.org/10.48550/arXiv.2505.12501>.
- E. Y. Chang and L. Geng. 2025b. SagaLLM: Context management, validation, and transaction guarantees for multi-agent LLM planning. *Proc. VLDB* 14, 1, 1–13. DOI: <https://doi.org/10.48550/arXiv.2503.11951>.
- A. Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 3, 181–204. DOI: <https://doi.org/10.1017/S0140525X12000477>.
- D. Eagleman. 2011. *Incognito: The Secret Lives of the Brain*. Pantheon Books, New York.
- G. Fauconnier and M. Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, New York.
- S. Freud. 1900. *The Interpretation of Dreams*. Macmillan.
- S. Freud. 1915. *The Unconscious, Standard Edition*. Vol. 14. Republished in 1957. 159–215.
- S. Freud. 1923. *The Ego and the Id*. Norton, New York.
- M. S. Gazzaniga. 2011. *Who's in Charge? Free Will and the Science of the Brain*. Ecco, New York.
- K. Gödel. 1967. On formally undecidable propositions of principia mathematica and related systems I. In J. van Heijenoort (Ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, 596–616.
- E. Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor Books, New York.
- G. W. F. Hegel. 1807. *Phenomenology of Spirit*. Translated by A. V. Miller, 1977. Oxford University Press, Oxford.
- K. Jaspers. 1969. *Philosophy*. Translated by E. B. Ashton, originally published in German in 1932. University of Chicago Press, Chicago.
- J. Jumper, R. Evans, A. Pritzel, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873, 583–589. DOI: <https://doi.org/10.1038/s41586-021-03819-2>.
- D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- I. Kant. 1781. *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood, 1998. Cambridge University Press, Cambridge.
- I. Kant. 1788. *Critique of Practical Reason*. Cambridge University Press.
- J. Kaplan, S. McCandlish, T. Henighan, et al. 2020. Scaling laws for neural language models. DOI: <https://doi.org/10.48550/arXiv.2001.08361>.

- A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6, 84–90. DOI: <https://doi.org/10.1145/3065386>.
- B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. 2017. Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253. DOI: <https://doi.org/10.1017/S0140525X16001837>.
- J. Liu, R. Hu, M. Wang, Y. Wang, and E. Y. Chang. 2008. Web-scale image annotation. In Y.-M. R. Huang, C. Xu, K.-S. Cheng, et al. (Eds.), *Proceedings of the 9th Pacific Rim Conference on Advances in Multimedia Information Processing (PCM '08)*. Tainan, Taiwan, December 9–13, 2008 Springer, Berlin, 663–674. DOI: [https://doi.org/10.1007/978-3-540-89796-5\\_68](https://doi.org/10.1007/978-3-540-89796-5_68).
- N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* 12, 157–173. DOI: [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638).
- D. G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 2, 91–110. DOI: <https://doi.org/10.1023/B:VISI.0000029664.99615.94>.
- P. Manakul, A. Liusie, and M. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, December 6–10, 2023. ACL, 9004–9017. DOI: <https://doi.org/10.48550/arXiv.2303.08896>.
- M. Merleau-Ponty. 1945. *Phenomenology of Perception*. Gallimard.
- T. Nagel. 1986. *The View from Nowhere*. Oxford University Press.
- D. E. Newman-Toker, N. Nassery, A. C. Schaffer, et al. 2024. Burden of serious harms from diagnostic error in the USA. *BMJ Qual. Saf.* 33, 2, 109–120. DOI: <https://doi.org/10.1136/bmjqs-2021-014130>.
- F. Nietzsche. 1968. *The Will to Power*. Vintage Books, New York.
- K. Popper. 1959. *The Logic of Scientific Discovery*. Hutchinson, London.
- J.-P. Sartre. 1943. *Being and Nothingness*. Washington Square Press, New York. 1943.
- D. Silver, A. Huang, C. J. Maddison, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587, 484–489. DOI: <https://doi.org/10.1038/nature16961>.
- D. Silver, J. Schrittwieser, K. Simonyan, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676, 354–359. DOI: <https://doi.org/10.1038/nature24270>.
- A. Turing. 1950. Computing machinery and intelligence. *Mind* 59, 236, 433–460. DOI: <https://doi.org/10.1093/mind/LIX.236.433>.
- G. Vlastos. 1991. *Socrates: Ironist and Moral Philosopher*. Cornell University Press, Ithaca, NY.
- F. Waters, A. Woods, and C. Fernyhough. 2014. Hallucinations in schizophrenia and their relationship to duration of untreated psychosis. *Schizophr. Res.* 1, 31–36.

- L. Wittgenstein. 1953. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Macmillan, New York.
- H. Xu, Y. Gao, F. Yu, and T. Darrell. 2017. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*, Honolulu, HI, July 21–26, 2017. IEEE, 3530–3538. DOI: <https://doi.org/10.48550/arXiv.1612.01079>.
- H. Yen, T. Gao, and D. Chen. 2024. Long-context language modeling with parallel context encoding. DOI: <https://doi.org/10.48550/arXiv.2402.16617>.
- Y. Zhu, R. Li, D. Wang, D. Haehn, X. Liang. 2025. Focus directions make your language models pay more attention to relevant contexts. DOI: <https://doi.org/10.48550/arXiv.2503.23306>.



# Epilogue: Two Paradigm Bridges

## Two Parallel Paradigm Shifts

This epilogue frames the book as the convergence of two parallel paradigm shifts.

### Newton to Einstein: From Fixed to Dynamic

Newton treated spacetime as fixed. Einstein reconceived spacetime as dynamic, shaped by mass and energy. Across the 218 years between them, ad hoc patches accumulated and were later rendered unnecessary by the new framing.

Unified Cognitive Consciousness Theory (UCCT) makes a similar move for cognition. A Large Language Model (LLM) stores an unconscious pattern prior. UCCT introduces semantic anchors and validation mechanisms that dynamically shape a task-conditional posterior. In effect, UCCT operationalizes supervised selection during inference: the anchor supplies the semantics, the repository contributes candidate patterns, and validation keeps only those that fit. The result is dynamic, anchored behavior, rather than static pattern completion.

### Hegel to MACI: From Information to Information Plus Behavior

Hegel argued that intelligence emerges through collaboration and synthesis. Multi-LLM Agent Collaborative Intelligence (MACI) adds the missing dimension that turns collaboration into practice: explicit behavioral modulation. High contentiousness supports diverse exploration, while low contentiousness supports focused reasoning. Transitions between these modes are driven by evidence quality. Together with information-theoretic governance, this framework yields accountable System-2 behavior.

## Key Points

- (1) **UCCT in one line:** treat model output as a prior that must be anchored and validated so that task-relevant patterns are amplified, off-target patterns are suppressed, and behavior becomes goal-directed rather than merely likely.

- (2) **UCCT in practice:** anchors define what counts as evidence; validation gates what passes; and the resulting posterior behaves as a regulated system that can be audited and stress-tested.
- (3) **MACI in one line:** An operational design that couples evidence routing and validation (EVINCE: Entropy and Variation in Conditional Exchanges; CRIT: Critical Reading Inquisitive Template) with behavior modulation (SocraSynth), role separation and oversight (Dike and Eris), persistent memory (SagaLLM), and localized repair via the Adaptive LLM Agent System (ALAS).
- (4) **MACI in practice:** Structured contention, roles, persistent memory, independent validators, and localized repair produce reliable, auditable plans that recover from faults without global resets.
- (5) **Collaboration has two essential dimensions:** information and behavior.
- (6) **Information alone** drifts toward consensus and can suppress minority but correct signals.
- (7) **Behavioral modulation** is required to shift from exploration to reasoning when evidence warrants it.

**The connection.** Both paradigm shifts share a common structure: what appears static or single-dimensional is, in fact, dynamic and multi-dimensional. Just as spacetime responds to mass and energy, pattern use responds to anchors and validation. Likewise, information-only collaboration becomes information plus behavior.

**Attribution note.** This framing records the intellectual lineage and the formal completions introduced here. Readers should cite the ACM Books edition when referring to UCCT and MACI as articulated in this volume.

### Why the Pair Matters

UCCT provides the theoretical foundation for anchoring and regulation, while MACI offers a deployable architecture with clear roles, explicit validators, durable memory, and localized repair mechanisms. Together, they move practice beyond scale alone and toward reliable, auditable intelligence.

### Priority and Provenance

To support accurate attribution, we succinctly record the public trail.

- (1) Frameworks and methods: SocraSynth, EVINCE, and CRIT (published 2023).

- (2) Book: *Multi-LLM Agent Collaborative Intelligence* (first edition March 2024; ACM Books release 2025).
- (3) Theory: UCCT articles and chapters (June 2025).
- (4) Systems: SagaLLM for persistent memory and ALAS for disruption-aware planning (2024–2025).
- (5) Applications and metrics: precision retrieval, validator-guided repair, and local edit measures (2024–2025).

**Edward Y. Chang**  
Stanford, California  
October 2025

## Acknowledgments

To the thinkers who changed starting assumptions rather than adding patches. To the engineers and students who turned ideas into code and data. To readers who will extend, test, and improve these frameworks. Please cite the sources so that the record remains clear.



## Author's Biography

### Edward Y. Chang



**Edward Y. Chang** is an adjunct professor in the Computer Science Department at Stanford University, where he has taught since 2019. He previously served as president of HTC DeepQ Healthcare (2012–2021) and as director of Research at Google (2006–2012), where he pioneered efforts in scalable machine learning, Web-scale image annotation (2008), and data-centric AI (2010), and also sponsored the ImageNet project. He was a visiting professor at UC Berkeley (2017–2020), working on surgical planning with virtual reality, and earlier, he held a tenured faculty position in Electrical and Computer Engineering at UC Santa Barbara (1999–2006). Chang earned his M.S. in Computer Science and Ph.D. in Electrical Engineering from Stanford University. In addition to his technical training, he pursued over ten courses in philosophy and literature—an intellectual breadth reflected in his research and writing. He is a Fellow of both ACM and IEEE for his contributions to scalable machine learning and AI in healthcare. His honors include the NSF CAREER Award, Google Innovation Award, the ACM SIGMM Test of Time Award, and the \$1 million XPRIZE for AI-driven disease diagnosis.



# Index

- Abdominal pain, 130, 155  
Abductive reasoning, 53, 92, 373  
Academia, 139, 142  
Acceptable behaviors, 257  
Accurate labeling, 199  
ACID. *See* Atomicity, Consistency, Isolation, and Durability (ACID)  
Actionable follow-ups, 178  
Activation, regimes of, 71  
Acute liver failure, 187, 189–191  
Adam narrative, relevance to, 472  
Adaptability, 2  
Adaptations, 500–501  
Adaptive alignment, checks and balances ensure, 504–506  
Adaptive approach, 177  
Adaptive behavior scheduling, 363  
Adaptive contentiousness modulation, 166, 178  
Adaptive coordination, 287–288  
Adaptive Layered Agent System (ALAS), 515  
Adaptive learning, 437  
    benchmarking, 438  
    DIAG, 438–440  
    PRBE, 440–444  
    RAFEL with four phases, 438  
    remediation strategies, 444–445  
Adaptive LLM Agent System (ALAS),  
    xvi, xxiii, 322, 335, 496  
additional JSSP results and analysis, 352  
agent factory implementation details, 339–343  
agent factory layer, 326–327  
architecture, 325  
complete meta-planner algorithm for workflow generation, 337  
experimental evaluation, 328–329  
job shop scheduling with disruptions, 332–335  
layer #1 workflow, 337–338  
LCPR specification and lemma proofs, 343–345  
network G, 323  
related work, 324  
runtime layer, 327–328  
supplemental information, 336–337  
three-layer architecture for adaptive and reactive planning, 324–325  
transportation scheduling, 329–332  
urban ride sharing, 345–352

- validation replan iterations, 354
- workflow blueprinting layer, 325–326
- Adaptive modulation, 177
- Adaptive planning, ALAS three-layer architecture for, 324–325
- Admiration, 234–238
  - Gemini’s interpretations on six emotion levels, 238
  - GPT-4 reinterpreted selected poems, 235–236
  - lady in garden scene under different emotions, 236
- Adversarial evaluation and rectification, 265
- Adversarial robustness, 166
- AE&B. *See* Analysis and Evaluation with Biases (AE&B)
- AE&H. *See* Analysis and Evaluation with Hallucination (AE&H)
- Affect-as-information theory, 252
- Agent architecture, 301
  - critical context and state management, 303–304
  - global coordination agents, 302
  - task execution agents, 301–302
  - validations performed by global validation agent, 303
- Agent assignment, 347
- Agent factory, 325, 339
  - agent coding mechanism, 340–341
  - agent discovery process, 339–340
  - compensation agent generation, 341
  - deployment artifact production, 341–342
  - factory design pattern implementation, 342
- implementation details, 339
- implementation efficiency considerations, 342
- layer, 326–327
- theoretical foundations, 343
- Agent role specification, 325–326
- AgentScope, 292
- Aggregation, 405, 410
- Agile regulation, 126, 128, 146–147
- Agreement-driven convergence, 169
- Agricultural revolution, 458
- Alanine Aminotransferase (ALT), 131
- ALAS. *See* Adaptive Layered Agent System (ALAS)
- Alcoholic hepatitis, 188–189
- AlexNet, 11–12
- Algorithm DIAG specifications, 439–440
- Algorithm PRBE specifications, 443–444
- Algorithmic bias, 5
- Alighieri, D., 239
- AlphaFold2, 34, 133
- AlphaGo, 34, 513
  - Zero, 34, 133, 513
- ALT. *See* Alanine Aminotransferase (ALT)
- Amazon Web Services (AWS), 291
  - Step Functions, 291
- Ambiguous anchors near threshold, 73
- Amnesiac navigator, 518
- Analogical bait, 517
- Analogy method, 94
- Analysis and Evaluation with Biases (AE&B), 441
- Analysis and Evaluation with Hallucination (AE&H), 441
- Analysis and explanation, 438, 444

- Anchor, 69, 72
- Anchoring
  - instantiations, 70
  - methods, 405
- Anchoring strength, 70, 79
  - estimation, 401
  - formulation, 409
  - metric, 64
- Anger-fear, 231
- Annotation quality, 215
  - sufficiency of current annotations, 215–216
- Annotators, 200, 215
- Antonym-based model, 232
- Antonym-based navigation, 254
- Anxiety-calmness, 228
- Aphorisms, 495
  - AGI emerges through collaborative intelligence, 515–516
  - aphorisms grouped by thematic pillars in MACI, 519
  - checks and balances ensure adaptive alignment, 504–506
  - choose baits, name fish you will get, 516–517
  - comparison of training-time scaling *vs.* inference-time scaling in LLM performance, 515
  - consciousness filters impulse, 506–507
  - context transforms capabilities, 501–503
  - debate strengthens reasoning quality, 508–510
  - essence of precise questioning, 507–508
- external mirrors enable validation, 498–499
- foundations and adaptations, 500–501
- framework with three independent components, 505
- hallucinations fade under scrutiny, 510–511
- human heuristics limited, and end-to-end learning shows supremacy, 513–514
- lady in garden scene under different emotions, 504
- linguistic behavior reflects intentions and emotions, 503–504
- list of 16 aphorisms, 496
- memory makes minds, 517–518
- organizational map of 16 aphorisms of MACI, 497
- polydisciplinary synthesis forges new frontiers, 511–513
- simulated reasoning useful, even if not real, 499–500
- truth emerges from perspectives, 496–498
- API. *See* Application Programming Interface (API)
- Appetite, loss of, 130
- Application Programming Interface (API), 171, 329
- Application State, 295, 303
- Arbitration, 78, 428
- Archetypes, 43, 459
- Area Under Curve (AUC), 406
- Arithmetic pattern density, 393, 403
- Arithmetic tasks, 404
- Arithmetic thresholds, 394, 416

- Artificial agents, 371
- Artificial general intelligence (AGI),  
xvii, xxi, 4, 63, 510
  - applications and extensions, xxv
  - core operational frameworks,  
xxiv–xxv
  - eight foundational pillars,  
xxii–xxiii
  - emerges through collaborative  
intelligence, 515–516
  - intelligence paradox, xxi
  - necessity and sufficiency of eight  
foundational pillars,  
xxiii–xxiv
  - organization, xxiv
  - from pattern matching to,  
xxi–xxii
  - from System 1 foundation to  
cognitive AGI rehabilitating  
System 1, xxiii
  - theoretical foundations, xxiv
  - toward cognitive AGI, xxv–xxvi
- Artificial intelligence (AI), xv, 1, 4,  
164, 321, 455, 500
  - definitions, 4
  - eras of similarity measurement,  
13–22
  - general terms, 4–5
  - performance terms, 5–6
  - perspectives on similarity, 6–13
  - philosophers, 509
  - rudimentary terms, 4
  - systems, 260
- Aspartate Aminotransferase (AST),  
131
  - AST. See Aspartate Aminotransferase  
(AST)*
- Asymmetric start, 165, 169
- Atomicity, Consistency, Isolation, and  
Durability (ACID), 288
- Attention, 369–371
  - models and transformers, 20
  - schema theory, 362
- Attentional deployment, 248
- AUC. *See Area Under Curve (AUC)*
- Augmentation techniques, 59, 488
- Augmented data, 58
- Australian Aboriginal Dreamtime  
stories, 460–461
- AutoGen, 292
- Automated SagALLM execution  
overview, 300
- Automatic Agent Deployment, 300
- Automatic design, 91
- Automatic System Architecture  
Generation, 300
- Automatic Transaction Execution, 301
- Autonomous driving, 34, 514
- Awareness, 366, 369
- AWS. *See Amazon Web Services (AWS)*
- Baars' global workspace theory, 506
- Baars' theory, 369
- Balance ethics, adversarial in-context  
review to, 257–259
- Bard, 130
  - concedes to GPT-4, 157
  - defense, 155
  - jaundice diagnosis, refutation of,  
156
  - justifications, 154
  - opening round, 154
  - refutation of GPT-4's hepatitis  
diagnosis, 156
  - relevant lab tests, 154
  - second round, 155

- supplementary symptom inquiries, 154
- BASE.** *See* Basically Available, Soft State, Eventually Consistent (BASE)
- Base–arithmetic ICL, 403
- Basically Available, Soft State, Eventually Consistent (BASE), 290
- Bayesian approaches, 16
- Bayesian inference, 267–268
  - in-context learning as, 67
- Bayesian model, 66
- Bayesian perspective, 395
- BCIs.** *See* Brain–Computer Interfaces (BCIs)
- Behavior classification
  - accuracy comparison, 264
  - evaluation, 263–265
- Behavioral analysis, 3
- Behavioral control, xvi
  - emotion regulation as, 248–249
- Behavioral emotion analysis model (BEAM), xvi, xxii, 228, 232–233, 250, 254–255
- Behavioral intensity, 164
- Behavior(s), 252
  - advisor, 252
  - analysis, 259
  - modeling, 251
  - spectrum, 256
- Behavior-shaping system, 383
- Benchmarking, 437, 438
- Bernstein bound, 417
- Bias detection, 207
  - dataset, 207–208
  - results on democrat scandals, 208–210
  - results on republican scandals, 210–212
- sufficiency of current annotations, 208
- value of partisan annotations, 208
- Bias mitigation, 212–213
  - convergence of all metrics, 214
  - debate arguments leading to neutrality, 213
  - observations and extended findings, 213–214
- Biased articles
  - democrat scandal news articles rated by EVINCE and justifications, 218–220
  - justifications of, 218–223
  - republican scandal news articles rated by EVINCE and justifications, 221–223
- Big data, 18
  - evolution of machine learning with, 19
- Biological intelligence, 500
- Black-box, 14, 228
- Bottom-up
  - awareness, 366
  - model, 369
  - processing, 366
- Brain–Computer Interfaces (BCIs), 474
- Breitbart, 210
- Bridging linguistic theory, 7
- Broadbent, D., 10
- Buddhism, 464, 468
- BudgetTrackingAgent, 302
- Capacity Management Agent, 348, 350

- Carlo, M., 350–351
- “Cautionary figures”, 458
- CBC. *See* Complete Blood Count (CBC)
- CDC. *See* Centers for Disease Control and Prevention (CDC)
- CE. *See* Cross-Entropy (CE)
- Censoring data, 33
- Centers for Disease Control and Prevention (CDC), 130
- Cerebellum, 365
- Cerebrum, 365
- Chain-of-thought (CoT), 393, 396
- empirical check in E1, 396
  - enthusiasm and speculative reasoning, 396
  - heuristics, 292
  - method, 53, 91
  - prompting, 52–53
  - UCCT explanation, 396
- Chang, E. Y., 19, 386, 416
- Chatbots, 56
- ChatGPT, 30, 32, 87, 248, 378, 435
- Checks-and-balances framework, 248
- “To My Sister” of different linguistic behaviors, 280–281
  - adversarial evaluation and rectification, 265
  - behavior classification evaluation, 263–265
  - contributions, 250–251
  - emotion layer evaluation, 261–263
  - emotion modeling, 251–252
  - emotion regulation as behavioral control, 248–249
  - emotion-behavior modeling, 252
  - for emotion-guided ethics, 249–250
  - empirical studies, 260
  - experimental constraints and dataset, 260–261
  - experimental design, 261
  - future work, 266
  - hate speech samples, 272
  - instruction to human annotators, 276
  - mixed emotions, 272–275
  - polarized emotions in one article, 277–280
  - reinforcement learning with human/AI feedback, 252–253
  - related work, 251
  - research aims, 260
  - three-branch framework design, 253–260
  - UCCT, 266–269
  - wheels of emotions, 269–270
- Chikungunya, 173–174
- EVINCE-mediated, 174
  - Chimpanzees, 461
  - Cirrhosis, 187, 189–191
  - CIs. *See* Confidence Intervals (CIs)
  - Civilization, xvii, 461
  - Clarke, A. C., 461–462
  - Claude, xix, 164, 171, 207, 312
  - convergent assessment, 190–191
  - moderator’s opening prompt to, 188
  - opening assessment, 188
  - Climate change, 469, 473
  - Clore, G. L., 252
  - Cluster analysis and similarity, 17
  - CNNs. *See* Convolutional Neural Networks (CNNs)
  - Code generation process, 327

- Code structures, 301
  - critical context and state management, 303–304
  - global coordination agents, 302
  - task execution agents, 301–302
  - validations performed by global validation agent, 303
- Coding process, 340
- Cognitive AGI, xxv–xxvi
- Cognitive foundations, 66–67
- Cognitive linguistics, 6
  - theories, 251
- Cognitive psychology, 6
  - DALL-E, 10
  - model, 372
  - neuroscience, 12
  - perspective, 9–12
- Cognitive reframing, 248
- Collaboration techniques, 384
- Collaborative frameworks, 479
- Collaborative intelligence, 384
  - AGI emerges through, 515–516
  - multi-agent orchestration for, 362
- Collaborative multi-agent dialogue, 492
- Collaborative reasoning, 496, 508
- Collaborative synthesis
  - ethical reasoning through, 363
  - process, 363
- Commonsense augmentation, 306
  - agent, 306
- Communicative Agents for Mind
  - Exploration of Large Scale Language Model Society (CAMEL), 288
- Community
  - community-based initiatives, 478
  - community-driven standards, 126
- engagement, 478
- standards, 127
- Comparative analysis, 175, 444
- Compensation planning, 295–296
- Compensation registration, 304
- Compensatory planning, 314
- Complete Blood Count (CBC), 154, 186
- Complete meta-planner algorithm for workflow generation, 337
- Complex emotions, 269, 272
  - forgiveness, 271–272
  - guilt and shame, 272
  - pride, 269–271
- Complexity, 364–365
- Comprehensive validation
  - framework, 304
- Computational advances, 20
- Computational capacity, 20
- Computational consciousness, 373
  - critical thinking with prompting ensembles, 378–381
  - emotion and behavior shaping through rewards, 376–378
- Eve refused to eat fruit, 383
- example of method maieutics, 382
- lessons learned and implementation insights, 381–384
- MFQ scheduler, 374–376
- remarks on conscious capabilities, 376
- template for being empathetic, 378
- toward systematic reasoning orchestration, 384

- Computational Consciousness Model (CoCoMo), [xvi](#), [xxii](#), [362–363](#), [366](#), [496](#)
  - from abstract consciousness to measurable transitions, [362](#)
  - computational consciousness, [373](#)
  - ethical reasoning through collaborative synthesis, [363](#)
  - functionalities of consciousness, [368–373](#)
  - multi-agent orchestration for collaborative intelligence, [362](#)
  - notes to CoCoMo design, [371](#)
  - structure and foundations, [363–364](#)
  - toward measurable implementation, [363](#)
  - understanding consciousness, [364–368](#)
- Compute cost, [179](#)
- Computer
  - algorithms, [100](#)
  - vision, [5](#)
- Computer science, [6](#)
  - perspective, [7](#)
  - probabilistic-based, [9](#)
  - rule-based, [7–8](#)
  - vector-space, [8–9](#)
- Computing contentiousness
  - approach to, [216–217](#)
  - debate parameters between agents A and B, two GPT-4 instances, [216](#)
- Conditional statistics, [120](#)
  - SocraSynth pseudo-code with, [121](#)
- Confidence Intervals (CIs), [406](#)
- Confirmatory role, [169](#)
- Connectionist model, [372](#)
- Conscious capabilities, remarks on, [376](#)
- Conscious semantics, Few Shot grounding as, [268](#)
- Consciousness, [368](#)
  - from abstract consciousness to measurable transitions, [362](#)
  - arising of consciousness, [365–366](#)
  - attention, [369–371](#)
  - awareness, [369](#)
  - bottom-up attention, [365](#)
  - critical thinking, [372](#)
  - definition and complexity, [364–365](#)
  - emotions and ethics, [371](#)
  - exploratory thinking, [372–373](#)
  - fading out of, [375–376](#)
  - filters impulse, [506–507](#)
  - functionalities of, [368–369](#)
  - key takeaways, [368](#)
  - module, [373](#)
  - notes to CoCoMo design, [371](#)
  - perception, [369](#)
  - phase transition may occur when energy peaks, [370](#)
  - theories, [366](#)
  - theory, [66–67](#), [362](#), [364](#), [386](#)
  - theory of functionalism, [367–368](#)
  - theory of panpsychism, [367](#)
  - understanding, [364](#)
- Consistency preservation, [294](#)
- Consistency, Robustness, Informativeness, Truthfulness (CRIT), [80](#)
- Consolidation, [428](#)
- Constitutional AI, [257](#)

- Constraint Consistency Validation, [314](#)
- Constraint Programming (CP), [334](#)
- Consultative approach, [144–145](#)
- Contentiousness, [xvi](#), [36](#), [38](#), [118](#), [227](#)
  - coarse-grained analysis of, [132](#)
  - modulating debate with, [115](#)
  - modulation, [169](#), [176–178](#)
  - parameter, [123–124](#), [132](#)
  - scheduling mechanism, [363](#)
  - surprises in fine-grained analysis of, [132–133](#)
- Contentment, [240](#), [279](#)
- Context, [239](#)
  - agents, [322](#)
  - context-based semantic disambiguation, [1](#)
  - ethical alignment with, [242–243](#)
  - narrowing, [307–311](#)
- Context-aware era, [19–20](#)
  - broader implications, [20–21](#)
  - emergence and evolution, [20](#)
  - foundational pillars, [20](#)
  - methodologies, [13](#)
  - models, [1](#)
- Context management, [294](#), [309](#)
  - agent, [309](#)
  - failure handling and recovery, [297](#)
  - mechanisms, [316](#)
  - requirements, [297](#)
- Contextual information, [49](#), [503–504](#)
- Continuous prompts, [91](#)
- Controlled behavioral dynamics, [164](#)
- Conventional logical methods, [118](#)
- Convergence, [207](#)
  - detection, [363](#)
  - examples of makespan at, [355–356](#)
- metrics, [175–176](#)
- Convolutional Neural Networks (CNNs), [10](#)
- Coordination, [300](#)
  - state, [302](#)
- Core theoretical architecture, [69](#), [398](#)
  - cross-domain validation, [399](#)
  - key insight, [399](#)
  - motivating observation, [398](#)
  - three principles, [398](#)
- Co-reference resolution, [101](#)
- Corroborate, [32](#), [404](#)
- Cosine similarity, [8](#)
- Cost, [60](#), [100](#)
  - cost–benefit perspective, [180](#)
- CoT. *See* Chain-of-thought (CoT)
- Counter-argument method, [97](#)
- Counter-counter-refutations, [145](#)
- Counterfactual reasoning, [90](#), [98–99](#), [105](#), [373](#), [384](#)
- Counterfactual technique, [87](#), [381](#)
- Counterfactual thinking, [98](#), [381](#)
- Courage, [240](#), [278](#)
- CP. *See* Constraint Programming (CP)
- Creative exploration methods, [384](#)
- Creative freedom, [105](#)
- Creative thinking, [376](#)
- Creativity, [105](#), [372](#)
- Crick, F., [102](#)
- CRIT. *See* Consistency, Robustness, Informativeness, Truthfulness (CRIT)
- Critic challenges, [xvi](#)
- Critical context, [303–304](#)
- Critical evaluation framework, [383](#)
- Critical reading, [99](#)
  - illustrative critical reading example, [94–95](#)

- Critical reading inquisitive template (CRIT), [xvi](#), [xxii](#), [95–97](#), [106](#), [114–115](#), [122](#), [165](#), [203](#), [378](#), [398](#), [496](#)
- evaluations by judges, [147](#)
- eve refused to eat the fruit, [89](#)
- judge GPT-3.5, [148–150](#)
- judge GPT-4, [150–151](#)
- judge text-davinci-003, [147–148](#)
- pilot study, [106–108](#)
- prompt template engineering, [99–105](#)
- pseudo-code, [96](#)
- related work, [90–92](#)
- remarks on, [99](#)
- Socratic method, [92–99](#)
- Critical regime, [71](#)
- Critical thinking, [92](#), [94](#), [99](#), [372](#), [376](#)
  - counterfactual thinking, [381](#)
  - CRIT pseudo-code, [379](#)
  - critical thinking with prompts, [380](#)
  - exploratory thinking with freedom, [379–380](#)
  - maieutics method, [380–381](#)
  - with prompting ensembles, [378–379](#)
- Cross-base interference, [407](#)
- Cross-domain generalization, [166](#), [324](#)
- Cross-domain geometric analysis, [396–397](#)
- Cross-domain validation, [180](#), [399](#)
- Cross-Entropy (CE), [199](#), [203](#), [508](#)
- Cross-examination, [101](#), [429](#)
- Cross-validation (CV), [420](#)
  - between machine and human annotators, [200–201](#)
  - with multiple annotators, [200](#)
- Cultural norms, adversarial in-context review to, [257–259](#)
- Cultural sensitivity, [20](#), [250](#)
- Cultural variability of pride, [271](#)
- Current annotations, sufficiency of, [208](#)
- Customer service, [13](#)
- CV. *See* Cross-validation (CV)
- Damasio, A. R., [252](#), [371](#)
- Dark urine, [130](#)
- Data
  - availability, [50](#)
  - and computation, [20](#)
  - data-centric methodologies, [13](#)
  - data-centric movement, [18](#)
  - data-centric paradigm, [18](#)
  - data-driven representation learning, [31](#)
  - insufficiency, [447](#)
  - learning model from, [1](#)
  - privacy *vs.* barrier to entry, [152](#)
  - processing, [18](#)
  - protection, [142](#)
  - sources, [59](#)
  - staleness, [447](#)
- Data-centric era, [18–19](#)
  - evolution of machine learning with big data, [19](#)
  - impact on similarity measurement, [19](#)
  - from MapReduce to machine learning at scale, [18](#)
- Dataset, [207–208](#)
- Daunting challenges for reliability, [164](#)
- Debate beats Q&A
  - data privacy *vs.* barrier to entry, [152](#)

- ethical standards vs. innovation, 151–152
- evaluation by three judges, 128
- global *vs.* national interests, 152
- in information quality, 127–128
- resource equity *vs.* public-private collaboration, 152
- Debate dynamics, 177
- Debiasing procedure, 202–203
  - exploring divergent viewpoints, 203
  - modulating contentiousness, 203
  - scrutinizing with reasonableness, 203–204
- Decision-making process, 21, 369
- Decision-theoretic refinement, 420
- Deep learning, 5, 14, 399
  - architectures, 19
- Deep Reinforcement Learning methods (DRL methods), 332
- Deep-dive phase, 41, 43–45
- Deep-probe (PRBE), 437, 440, 444
  - algorithm PRBE specifications, 443–444
  - expected outcome, 444
  - focused exploration, 441–442
  - in healthcare, environmental science, and sports news domains in four categories, 442
  - PRBE algorithm pseudo-code, 443
  - strategic questioning, 440–441
- DeepSeek R1, 312
- “Defensive architecture”, 512
- Definition, 87–88, 93
  - clarification with, 100–101
  - method of, 95–96
- Delight, 234–238
- Gemini’s interpretations on six emotion levels, 238
- GPT-4 reinterpreted selected poems, 235–236
- lady in garden scene under different emotions, 236
- Demirkol-Mehta-Uzsoy (DMU), 329
- Democrat-leaning annotators, 211
- Democrat scandals
  - bias distances, 210
  - comparison of bias assessments, 209
  - results on, 208–210
- Democratic-leaning raters, 198
- Dengue fever, 173–174, 186
  - EVINCE-mediated, 174
  - NS1 antigen test for, 186
- Dependency State, 295, 304
- Dependency tracking, 295–296
- Dependency validation, 296
- Deployment artifact production, 341–342
- Depth exploitation, 169
- Despair, 281
- DIAG. *See* Diagnosis/Diagnostic (DIAG)
- Diagnosis/Diagnostic (DIAG), 434, 438–440
  - algorithm DIAG specifications, 439–440
  - algorithm pseudo-code, 439
  - analysis, 437
  - of cognitive disparities, 438–440
  - phases, 444
  - tool, 198
- Dialectic method, 87–88, 94, 97–98
- Digital realms, 469
- DIKE, 255, 265

- behavior evaluation and rectification, 257
- self-supervised behaviors and emotions mapping, 256
- self-supervised learning, 262
- Diligence, 240
- Dimensionality reduction techniques, 15
- Direct preference optimization (DPO), 253
- Discrete prompts, 91
- Disgust, 234–238
  - Gemini’s interpretations on six emotion levels, 238
  - GPT-4 reinterpreted selected poems, 235–236
  - lady in garden scene under different emotions, 236
- Disruption-Aware Replanning, 324
- Disruption handling, 327–328
- Disruptions, job shop scheduling with, 332
- Distance functions, 15
- Distance Optimization Agent, 348, 351
- Diverse emotional states, 264
- Diversity
  - candidate formulations, 421
  - combine predictions weighted by, 183–184
  - design ideals, 420
  - motivation, 420
  - planned evaluation, 421
  - within relevance, 420
  - scope, 421
- DMU. *See* Demirkol–Mehta–Uzsoy (DMU)
- Doe, J., 325
- Domain
  - agents, 322
  - scope, 179
- Dot product, 8
- DPF. *See* Dynamic Partial Function (DPF)
- DPO. *See* Direct preference optimization (DPO)
- DRL methods. *See* Deep Reinforcement Learning methods (DRL methods)
- DTW. *See* Dynamic Time Warping (DTW)
- Dual entropy, 204–205
- Dual-LLM, 265
- Dual-process
  - framework, 372
  - model, 372
  - theories, 66, 68
- Durability assurance, 294
- Dynamic Core Hypothesis, 365
- Dynamic Partial Function (DPF), 11, 421
- Dynamic Time Warping (DTW), 413
- ECE. *See* Expected Calibration Error (ECE)
- Ecological degradation, 469
- Ecological interpretation, 455
- E-commerce, 2
- Ecstasy, 233
- Edge
  - agents, 348
  - constraints, 326
- EDT. *See* Entropy duality theorem (EDT)
- Education, 478–479
- Educational programs, 478, 483
- Educational utility, 271
- Ekman, P., 229, 233, 251, 254

- Elastic Weight Consolidation (EWC), 78
- Elder, L., 88, 372
- Elenchus method, 87–88, 93, 96–97  
examples of, 102  
verification with, 101–103
- Elenctic method, 98
- Ellsworth, P. C., 251
- Emergent behavior, 67
- Emotion(s), 251, 254, 371, 376  
admiration/delight *vs.* disgust, 234–238  
checks-and-balances framework for emotion-guided ethics, 249–250  
comparative display of emotion models, 270  
complex emotions, 269–272  
distributions in affection behaviors, 262  
emotion-based behavior control, 228  
emotion-behavior modeling, 252, 253  
emotion-driven approach, 251  
emotion-shaping system, 383  
ethical violation correlates to, 238–239  
framework with three independent branches, 249  
inclusion, 233  
joy *vs.* sadness, 234  
layer evaluation, 261–263  
linguistic features of, 233–238  
of love, 277  
mapping, 256  
modeling, 251–252  
modulation, 259
- regulation as behavioral control, 248–249  
wheels of, 269
- Emotional behaviors, 266
- Emotional intelligence, 377
- Emotional intensity, 232
- Emotional modeling, 377
- Emotional realism, 271
- Empty context, 447
- End-to-end learning, 514  
shows supremacy, 513–514
- Energy, 239  
energy-based thresholds, 381  
intensity of, 239  
trajectory of, 239
- Enthusiasm, 396
- Entropy, 203  
reduction, 502
- Entropy and Variation in Conditional Exchanges (EVINCE), xvi, xxii, 79, 163, 203, 496  
additional symptom inquiries, 192  
agreement-driven convergence, 169  
algorithm, 203  
algorithm and foundations, 168  
algorithm specification, 169–171  
analysis, 186  
analysis of information-theoretic guidance mechanisms, 173–175  
application to language models, 182  
approach to computing  
contentiousness, 216–217  
combine predictions weighted by diversity and quality, 183–184

- contributions, 165–166
- debate, 184, 187
- design maxims of EVINCE
  - moderation, 183
- diagnostic accuracy and MRR, 173
- EDT, 169
- empirical evaluation of, 171
- evaluation metrics, 172
- experimental results, 172–173
- experimental setup, 171
- final consensus, 191
- final joint recommendations, 186, 192
- foster exploration through diverse perspectives, 183
- future work, 166, 180
- at glance, 80
- GPT-4’s opening top-5 prediction distribution, 187–188
- high contentiousness, 184
- information-theoretic metrics in, 175
- joint final assessment, 191–192
- justification for KL divergence, 182
- key benefits of, 178
- key capabilities demonstrated, 186
- key results, 166
- limitations, 179–180
- low contentiousness, 185, 190–191
- maximum-entropy exploration, 168–169
- metrics for LLM debate evaluation, 180–181
- moderate contentiousness, 185, 188–190
- moderator’s opening prompt to Claude, 188
- moderator’s opening prompt to GPT-4, 187
- multi-agent debate with EVINCE, 397–398
- orchestrate two equally competent LLMs in structured debate, 183
- our approach, 165
- performance comparison, 172
- quantitative improvements, 186
- recommended lab tests, 192
- refine high-quality perspectives, 183
- related work, 166–168
- statistical testing, 172
- summary of EVINCE debate on news D1, 216
- supporting arguments, 217–218
- system prompt, 172
- theorem proving, 181
- theoretical foundations, 168
- theoretical justification of EDT, 181–182
- Entropy duality theorem (EDT), 168–169, 204–205
- pairing, 80
- theoretical justification of, 181–182
- Environmental studies, 2
- Equipment failure diagnosis, 14
- ERIS, 257–259, 265
  - checks-and-balances framework, 258
- Erving Goffman’s theory, 499
- Ethical AI guidelines, 479
- Ethical violation correlates to emotions, 238–239

- Ethics, 21, 371
- ETHOS-Long, 260
- Evaluative stage, 117, 122
- CRIT pseudo-code, 123
  - evaluating reasonableness over truth, 122–123
  - recursive consideration, 123
- Eve narrative, relevance to, 472
- EWC. *See* Elastic Weight Consolidation (EWC)
- Exclusion criteria, 233
- Exec proposes, xv
- Execution workflow, 350
- workflow template to, 348–352
- “Exile”, 465–466
- Existential identity, 486–487
- Expected Calibration Error (ECE), 172, 403, 406
- Experimental constraints and dataset, 260–261
- Expert system, 7
- Explainable AI, 4
- Exploration, 427
- Exploratory phase, 41–43
- Exploratory thinking, 372–373
- with freedom, 379–380
- Explorer role, 169
- Exposure audit, 404
- Extensive human-generated data, 253
- External density, 76
- External mechanisms, 63–64
- False Discovery Rate (FDR), 173
- Fatigue, 130, 154
- Fauconnier, G., 252, 512
- FDR. *See* False Discovery Rate (FDR)
- Feature Integration Theory (FIT), 9–10
- Feature weighting, 15
- Few Shot
- grounding as conscious semantics, 268
  - learning, 51–52, 393
  - performance, 67
  - phase transitions, 407
  - prompting, 63–64, 67
  - protocol, 405
  - success model, 71
- Few Shot cat classification without weight updates, 73
- anchors, patterns, query, and result, 74–75
- clarifying “labels”
  - misconception, 74
- human *vs.* machine priors, 74
- theoretical alignment, 75
- Few Shot semantic re-anchoring, 72
- ambiguous anchors near threshold, 73
  - symbol swap, 73
- Two Shot anchors, 72–73
- Zero Shot baseline, 72
- Finance sector employs rule-based systems, 13
- Fine-grained analysis of contentiousness, surprises in, 132–133
- Fine-tuning, 436, 445
- dynamics, 409
  - protocol, 406
- Fisherman analogy, 64
- FIT. *See* Feature Integration Theory (FIT)
- Fitzgerald, F. S., 263, 274
- FlightBookingAgent, 301
- Flight Booking Compensation, 304
- Floating-Point Operations (FLOPs), 180

- FLOPs. *See* Floating-Point Operations (FLOPs)
- FMs. *See* Foundation models (FMs)
- Forgiveness, 271–272
- Foundation models (FMs), 29, 33, 113, 152
- Fragmented context, 447
- Fraud detection systems, 13
- Fredrickson, B. L., 252
- Freud, S., 366
- Frozen vision encoder, 399
- Functionalism, 366
  - theory of, 367–368
- Gab Hate Corpus, 260, 272
- GAI. *See* Generative AI (GAI)
- Gaining insights, superiority of debate over Q&A in, 131
- Gaussian Mixture Models (GMMs), 17
- Gazzaniga’s “interpreter” theory, 506
- GEM. *See* Gradient episodic memory (GEM)
- Gemini, 36, 38, 41, 118, 234, 263
  - 1.5 pro model, 59
  - convergent assessment, 185
  - definitions and disciplinary perspectives on consciousness from, 38
  - opening assessment, 184
  - questions on consciousness from, 39
  - reassessment, 185
  - sets of questions from, 40
- Gender
  - binary, 467
  - fluidity, 469
- Gene Expression Programming (GEP), 332
- General AI, 4
- General Language Model, version-3 (3GLM), 334
- General Practitioners (GPs), 129, 131
  - final joint recommendations to, 130–131
- Generalization, 6, 103
  - technique, 87–88, 93
- Generalized preference optimization (GPO), 252–253
- Generation techniques, 58
- Generative AI (GAI), 435
- Generative pre-trained transformer (GPT), 49, 372
  - GPT-API, 136
- Generative pre-trained transformer-3 (GPT-3), 56, 87
- Generative pre-trained transformer-4 (GPT-4), 27, 36, 38, 41, 45, 114, 118, 233, 256
  - additional symptom inquiries, 155
  - bard concedes to, 157
  - defense, 156
  - definitions and disciplinary perspectives on consciousness from, 38
  - hepatitis diagnosis, refutation of, 156
  - justifications, 154–155
  - moderator’s opening prompt to, 187
  - opening assessment, 184
  - opening round, 154
  - opening top-5 prediction distribution, 187–188
  - questions on consciousness from, 39
  - refined analysis, 185, 190

- refutation of bard's jaundice diagnosis, 156
- relevant lab tests, 155
- second round, 156
- sets of questions from, 40
- Zero Shot approach, 263
- Generative pre-trained transformer-4
- instance A (GPT-A/B), 44, 456, 489–490
  - GPT-A agrees with GPT-B, 473
  - GPT-A answers GPT-B, 460–462, 467–468, 476–477, 481–482, 487–488
  - GPT-A asks GPT-B, 457–458, 462–463, 468–469, 473–474, 477–478, 483
  - GPT-A's feedback to moderator, 484
  - GPT-B answers GPT-A, 458–459, 464–465, 470–471, 474–476, 478–480, 484–485
  - GPT-B asks GPT-A, 459–460, 465–467, 476, 480–481, 486
  - GPT-B conveys new topics to GPT-A, 473
  - moderator initializes agent GPT-A, 456–457
- Generative stage, 117–118
- changes in arguments at different contentiousness levels, 119
  - eliminating hallucination, 118–120
  - mitigating model biases, 118
  - more on conditional statistics, 120
  - SocraSynth algorithm, 120–122
- Genetic programming (GP), 334
- additional symptom inquiries, 157
- final joint recommendations to, 157
- recommended lab tests, 157
- Geng, L., 336
- Geometric trajectory analysis, 393–394, 396, 403, 416
- cross-domain geometric analysis, 396–397
- quantification in E2, 397
- universality question, 396
- GEP. *See* Gene Expression Programming (GEP)
- Gibson, W., 475
- Global coordination agents, 302
- Global interests, 152
- GlobalValidationAgent, 302
- Global workspace, 369
- theory, 66, 362, 365
- GMMs. *See* Gaussian Mixture Models (GMMs)
- Goffman, E., 499
- “Goldilocks zone”, 462
- GP. *See* Genetic programming (GP)
- GPO. *See* Generalized preference optimization (GPO)
- GPs. *See* General Practitioners (GPs)
- GPT. *See* Generative pre-trained transformer (GPT)
- GPT-3. *See* Generative pre-trained transformer-3 (GPT-3)
- GPT-4. *See* Generative pre-trained transformer-4 (GPT-4)
- GPT-4o, 312
- GPT-A/B. *See* Generative pre-trained transformer-4 instance A (GPT-A/B)

- GPU. *See* Graphics processing unit (GPU)
- Grade School Math 8K (GSM8K), 404
- Gradient episodic memory (GEM), 78
- Graphics processing unit (GPU), 179
- Grassroots movements, 478
- Grief, 233, 499
- Ground retrieves, xvi
- Ground truth, 171, 497
- data, 129, 131
  - labels, 178
- GSM8K. *See* Grade School Math 8K (GSM8K)
- Guilt, 233, 272
- HAI. *See* Human-Centered Artificial Intelligence (HAI)
- Hallucination(s), 6, 462, 510
- fade under scrutiny, 510–511
- Hand-engineering, 100
- Hardware improvements, 3
- Harmonization, 428
- Harris, Z., 7, 17, 19
- Hassabis, D., 34, 133
- Hate-love spectrum, 229
- Hate speech samples, 272
- HAV. *See* Hepatitis A virus (HAV)
- “Haystack”, 59–60
- HBV. *See* Hepatitis B virus (HBV)
- HCV. *See* Hepatitis C virus (HCV)
- Healthcare, 21, 129, 198
- bard concedes to GPT-4, 157
  - bard’s opening round, 154
  - bard’s second round, 155–156
  - debate, 153
  - final joint recommendations to GP, 157
  - GPT-4’s opening round, 154–155
  - GPT-4’s second round, 156
- healthcare-related inquiries, 442
- moderator’s opening prompt, 153
- Hegel’s dialectical process, 509
- Hepatitis, 129–130, 174–175, 187–192
- EVINCE-mediated, 174
  - panel, 131, 192
- Hepatitis A virus (HAV), 188–189
- Hepatitis B virus (HBV), 187–191
- Hepatitis C virus (HCV), 187–192
- Heterogeneous LLMs, 495
- Heuristic-based modeling, 1
- Hidden Markov Model (HMM), 268
- High contentiousness, 184, 228
- Gemini’s opening assessment, 184
  - GPT-4’s opening assessment, 184
- High-intensity emotions, 252
- High-level UCCT formulation, 79
- High precision and recall, 59–60
- HMM. *See* Hidden Markov Model (HMM)
- Honesty, 240
- Hopkins, J., 129, 497
- Horvitz, E., 511
- HotelBookingAgent, 301
- Hotel Booking Compensation, 305
- Hubristic pride, 269
- HuffPost, 209
- Human
- annotations, 251
  - brain, 65, 473, 474
  - cognition, xxi, 11, 456
  - colonization, 469
  - consciousness, 37
  - emotions, 277, 454
  - heuristics limited, 513–514
  - human–AI partnership, 492
  - human-labeled data, 197

- innovation, 513
- intelligence, 516
- intervention, 259
- involvement, 134
- language, 503
- machine integration, 469
- moderator, 508
- priors, 74
- psychology, 248
- Human annotators, 272
  - cross-validation between human and machine annotators, 200–201
  - instruction to, 276
- Human-Centred Artificial Intelligence (HAI), 511
- Human knowledge, 453
  - limitations of, 33–35
- Humanity, 456
- Huxley, A., 464
- Hybrid models, 434
- Hybrid stopping, 427
- Hypothesis elimination, 88, 93
- ICL. *See* In-Context Learning (ICL)
- ID. *See* In-distribution (ID)
- Identify societies, 460
- Identity, 252
- IIT. *See* Integrated Information Theory (IIT)
- ImageNet, 33, 200, 514
- Imaging studies, 131, 192
- Immutable Action Logging, 314
- In-Context Learning (ICL), 394
  - as Bayesian inference, 67
  - competing theoretical explanations, 395
  - experiment E1 design, 395
  - structure matters in, 394–395
- UCCT predictions for ICL effectiveness, 395
- Incorporating domain knowledge, 21
- Independent validation, 293
  - inter-agent input and dependency validation, 296
  - intra-agent output validation, 296
  - requirements, 296
  - validation response protocols, 296
- In-distribution (ID), 79, 393
- Induction, 90, 93
- Industry-driven certifications, 127
- Inferotemporal cortex (IT cortex), 267
- Information, 370
  - anchoring, 384
  - debate beats Q&A in information quality, 127–128
  - information-processing model, 372
  - information-theoretic metrics, 168, 175
  - information-theoretic monitoring, 363
  - retrieval, 14
  - sources, 442
  - theory, 39, 502
- Information-theoretic guidance mechanisms, analysis of, 173
- ablation study, 176–178
- convergence metrics, 175–176
- convergence of information metrics, 176
- dengue fever *vs.* chikungunya, 173–174
- jaundice *vs.* hepatitis, 174–175
- key benefits of EVINCE, 178
- Initial dialogue iterations, 203

- Initial Plan Generation, 300
- Initial prompt, 330, 456
- Initial validation, 348
- Initiation, 1
- Innate perceptual biases, 74
- Inner product, 8
- Input style, 100
- Instruction tuning, 67, 79
- Integrated Information Theory (IIT), 362, 364
- Intelligence emerges, xv, 502
- Interactive systems, 21
- Interactivity limitation, 455
- Inter-agent input validation, 296
- Inter-annotator entropy, 264
- Interconnections, justifications of, 472–473
- Internal circuits, 68
- Internal decision-making process, 14
- Internal evaluation methods, 498
- Internal policies, 126
- Internal representations, 395
- Internal validation, 288
- International framework, 143
- Internet of things (IoT), 487
- Internet search engines, 114
- Interpretation subjectivity, 455
- Interrupt mechanisms, 375
- Intra-agent output validation, 296
- IoT. *See* Internet of things (IoT)
- Irony method, 94
- Iroquois, 464
- Isolation guarantees, 294
- Isotonic calibration, 419
- IT cortex. *See* Inferotemporal cortex (IT cortex)
- Iterative Refinement, 300, 348
- ItineraryPlanningAgent, 302
- James–Lange theory of emotion, 252
- Jaspers, K., 510
- Jaundice, 129–130, 154–156, 174–175, 187–192
- EVINCE-mediated, 174
- Jensen–Shannon divergence (JSD), 79, 163, 168, 199, 426, 508
- Job-Shop Scheduling Problem (JSSP), 316, 328
  - additional experimental results and analysis, 356
  - additional JSSP results and analysis, 352
  - ALAS for four representative JSSP benchmark, 354
  - ALAS’s validation replan iterations, 354
  - convergence of ALAS on five JSSP datasets and upper bound, 353
  - examples of makespan at convergence, 355–356
  - LLM prompt design, 352
  - scheduling prompt, 353
- Job shop scheduling with disruptions, 332
  - comparison of mean gap to upper bound across two benchmark datasets, 333
  - experimental results, 332–334
  - LCPR mechanism, 334–335
  - local compensation, 335
  - problem specification, 332
- John R., 6, 17
- Joy, 233–234
- Joyful affection, 277, 281
- JSD. *See* Jensen–Shannon divergence (JSD)

- JSSP. *See* Job-Shop Scheduling Problem (JSSP)
- Judge GPT-3.5, 148–150
- Judge GPT-4, 150–151
- judge text-davinci-003, 147–148
- Jung, C., 458
- Justice, 240
- Justifications of interconnections, 472–473
- Kaggle symptom-disease, 171
- Kahneman, D., xv, xxi
- Kahneman–Tversky optimization (KTO), 253
- Kaya, Z. N., 416
- Keats, J., 233
- K-nearest neighbors language model (kNN-LM), 397
- kNN-LM. *See* K-nearest neighbors language model (kNN-LM)
- Knowledge generation, 113, 243, 248
- KTO. *See* Kahneman–Tversky optimization (KTO)
- Kullback–Leibler (KL)
- decay, 169
  - divergence, 205
  - justification for KL divergence, 182
- Label validation, 199
- cross-validation between machine and human annotators, 200–201
  - cross-validation with multiple annotators, 200
- Label-error detection, 178
- Lam, M., 386
- LangGraph, 292
- Language
- language-relevant emotions, 252
- translation harnesses, 3
- Language Model Mentor (LMM), 134
- Language model(s), 51, 498–499, 501
- application to, 182
  - capabilities and implications, 28
  - distinctive capabilities, 29–30
  - exploring unknown unknowns, 36–45
  - larger always better, 35–36
  - limitations and opportunities, 28–29
  - limitations of human knowledge, 34–35
  - polydisciplinary, 30–31
  - polymodality, 32
  - post-training value alignment, 32–33
  - pre-training censorship, 33
- Large Language Model–Monte Carlo Tree Search (LLM-MCTS), 292
- Large Language Models (LLMs), xv, xxi, 5, 20, 27, 35, 49, 63, 87–88, 113–114, 120, 164, 166, 247, 269, 288, 321, 361, 391, 433, 435, 449, 495
- capability, 100
  - comparison of RAG and fine-tuning for enhancing LLMs, 436
  - computational complexity analysis, 181
  - context limitations and strategic preservation, 291–292
  - fine-tuning, 436
  - formulas, 180–181
  - limitations, 313–315

- limitations necessitating
  - SagaLLM's key requirements, 291
- LLM-Orchestrated Intelligence, 290
- metrics for LLM debate evaluation, 180
- orchestrate two equally competent LLMs in structured debate, 183
- prompt design, 352
- related work, 435
- remediation playbook for LLM enhancement, 446
- retrieval-augmented generation, 436–437
- retrospective and adaptive learning, 437–445
- self-validation gap necessitating independent validation, 291
- statelessness necessitating transactional integrity, 291
- UCCT + RAFEL, 445–449
- Latency, 60, 179
- Latent Dirichlet Allocation (LDA), 7
  - latent clusters of, 17
- Latent Semantic Analysis (LSA), 7
- Latent semantic models, 16–17
- LCPR. *See* Local Compensation Protocol (LCPR)
- LDA. *See* Latent Dirichlet Allocation (LDA)
- Le Guin, U. K., 470
- Learning, 72
  - learning-threshold scaling, 408
- LeCun, Y., 73
- Left-to-right language model, 91
- Lemma, 344–345
- Lethargy, 130, 154
- Letter-writing behaviors, 256
- LFTs. *See* Liver function tests (LFTs)
- Li, B., 11
- Li, F.-F., 19
- Limited-context LLMs, RAG with, 58–59
- Linear methods, 15
- Linear-linear progressions, 271
- Linguistic behaviors, 281
  - “To My Sister” of different linguistic behaviors, 280
  - reflects intentions and emotions, 503–504
  - rewrites depicting different linguistic behaviors, 281
- Linguistics, 6
  - perspective, 6–7
- Liu, D., 309
- Liver function tests (LFTs), 131, 192
- LLM-MCTS. *See* Large Language Model–Monte Carlo Tree Search (LLM-MCTS)
- LMM. *See* Language Model Mentor (LMM)
- Local compensation, 327
- Local Compensation Protocol (LCPR), 324
  - corollary, 345
  - key observations, 345
  - lemma, 344–345
  - mechanism, 334–335
  - specification and lemma proofs, 343
- Long-context LLMs, RAG with, 59
  - high precision and recall, 59–60
  - low latency and cost, 60
- Longest Processing Time (LPT), 332
- Longest Subsequent (LSO), 332

- LoRA. *See* Low-Rank Adaptation (LoRA)
- Love, 241  
emotion of, 277
- Love expressions, 261  
intensity and dominant emotions, 261
- Love letter experiment  
Few Shot grounding as conscious semantics, 268  
implications for, 268
- Low contentiousness, 185, 190  
Claude's convergent assessment, 190–191  
Gemini's convergent assessment, 185  
GPT-4's refined analysis, 185, 190
- Low-Rank Adaptation (LoRA), 393
- LPT. *See* Longest Processing Time (LPT)
- LSA. *See* Latent Semantic Analysis (LSA)
- LSO. *See* Longest Subsequent (LSO)
- Lyttelton, L. M. O., 279–280
- Machine and human annotators, cross-validation between, 200–201
- Machine learning (ML), 5, 18  
evolution of ML with big data, 19  
MapReduce to ML at scale, 18  
methods, 272  
models, 34, 125, 197, 200, 455
- Machine priors, 74
- Machine translation, 99
- MACI. *See* Multi-Agent Collaborative Intelligence (MACI)
- MAD. *See* Multi-agent debate (MAD)
- MAD systems. *See* Multi-Agent Dialogue systems (MAD systems)
- Maieutics, 90, 384  
examples of, 104  
generalization with, 103–105  
method, 87, 93, 98, 380–381
- Malaise, 130
- Manichaeism, 468
- Manual design, 91
- Manual itinerary planning, 300
- Many-shots example, 55–57  
template for being empathetic, 57
- MapReduce to machine learning at scale, 18
- MAS. *See* Memory aware synapses (MAS); Multi-agent systems (MAS)
- Masked LLMs, 91
- Massive Multitask Language Understanding (MMLU), 27, 118
- Master coordinator, 328
- Maximum-entropy exploration, 168–169
- Maximum likelihood (ML), 212  
default, 63  
pattern distribution, 69
- Maximum Likelihood Estimation (MLE), 419
- MBD. *See* Media Bias Detector (MBD)
- Mean Reciprocal Rank (MRR), 173
- Measurable validation, 384
- Mechanism transparency, 179
- Mechanistic interpretability, 68
- Mechanistic perspective, 395
- Media Bias Detector (MBD), 200
- Medical imaging, 2, 22

- Memory, [xvi](#)
  - logging, [328](#)
  - modules, [322](#)
- Memory aware synapses (MAS), [78](#)
- Mesopotamia, [465](#)
- Metaphorical interpretation, [457](#)
- MFQ. *See* Multi-Level Feedback Queue (MFQ)
- MI. *See* Mutual information (MI)
- Microsoft, [28, 288](#)
- Minimizing Re-optimization Costs, [328](#)
- Mitchell, T., [18](#)
- Mitigate hallucinations, refine context to, [115–116](#)
- Mixed emotions, [272–275](#)
  - excerpts from Zelda Sayre to F. Scott Fitzgerald, [274](#)
  - sample texts from gab hate corpus, [273](#)
  - statistics of sentiments distributed across letters, [275](#)
- ML. *See* Machine learning (ML); Maximum likelihood (ML)
- MLE. *See* Maximum Likelihood Estimation (MLE)
- MMLU. *See* Massive Multitask Language Understanding (MMLU)
- Model-based era, [14](#)
  - probabilistic models, [16–18](#)
  - vector space models, [14–16](#)
- Model-based methodologies, [13](#)
- Modeling language, [255–257](#)
- Moderate contentiousness, [185, 188](#)
  - Claude’s reassessment, [189–190](#)
  - Gemini’s reassessment, [185](#)
  - GPT-4’s response, [185, 189](#)
- Moderator
  - feedback, [260](#)
  - initializes agent GPT-A, [456–457](#)
  - opening prompt, [153](#)
- Moderator’s intervention, [471](#)
  - GPT-A agrees with GPT-B, [473](#)
  - GPT-B conveys new topics to GPT-A, [473](#)
  - justifications of interconnections, [472–473](#)
  - recommendations for focused topics, [471–472](#)
  - relevance to Adam And Eve narrative, [472](#)
- Moderator’s prompts, [136–138](#)
  - refinement of debate topics, [138](#)
  - topics proposed by agent A and agent B, [137](#)
- Modern models, [xv, 437](#)
- Modesty, [240](#)
- Modular agent architecture, [323](#)
- Modular architecture, [381](#)
- Monitoring agents, [322](#)
- Moral frameworks, [487](#)
- Most Process-Sequence Remaining (MPSR), [332](#)
- MPSR. *See* Most Process-Sequence Remaining (MPSR)
- MRR. *See* Mean Reciprocal Rank (MRR)
- MTGP. *See* Multi-Tree Genetic Programming (MTGP)
- Multi-agent
  - approach, [373](#)
  - reactive planning, [352](#)
  - scalability, [166](#)
- Multi-Agent Collaborative Intelligence (MACI), [362](#)
- Multi-agent debate (MAD), [227, 393](#)

- anchors, 428
- arbitration, 428
- calibrated control, 427
- case, 428–429
- cost, reproducibility, and ethics, 429–430
- design goal, 425–426
- diagnostics and notation, 426–427
- evaluation framework, 429
- with EVINCE, 397
- harmonization, 428
- hybrid stopping, 427
- illustrative example, 428
- MAD experiment plan, 429
- phase-guided debate, 427–428
- plateau detection, 427
- primary tests, 429
- promise and challenges, 397–398
- scope and intent, 425
- stance-based anchors and label harmonization, 428
- trade-offs, 429
- UCCT, 398, 425
- Multi-Agent Dialogue systems (MAD systems), 164
- Multi-agent LLM frameworks, 292
  - limited context preservation, 293
  - missing transaction semantics, 292
  - validation gaps, 292
- Multi-agent orchestration, 363
  - for collaborative intelligence, 362
- Multi-agent socrasynth, 116–117
  - evaluative stage, 122–123
  - generative stage, 117–122
- Multi-agent systems (MAS), 288
- Multi-Level Feedback Queue (MFQ), 374
- fading out of consciousness, 375–376
- interrupt and synchronization mechanisms, 375
- interrupt into and fading out of consciousness, 375
- scheduler, 374–375
- Multi-LLM Agent Collaborative Intelligence (MACI), xxi, 35, 495
  - MACI Governs LLMs, 506–507
- Multi-LLM debate framework, 227
- Multi-Tree Genetic Programming (MTGP), 332
- Multilinguality, 20
- Multimodal data integration, 21
- Multimodal LLMs
  - empirical study, 233–238
  - qualifying and quantifying emotions, 229–233
  - qualifying and quantifying ethics, 238–243
  - Romeo and Juliet, 244
- Multiple annotators, cross-validation with, 200
- Multivariate Pattern Analysis (MVPA), 12
- Mutual information (MI), 79, 165, 203, 205, 398, 502, 508
- Mutual reinforcement, 294
- MVPA. *See* Multivariate Pattern Analysis (MVPA)
- Nagel, T., 496
- Named-Entity Recognition (NER), 100
- Narrow AI, 4
- National interests, 152
- National Security Agency (NSA), 209, 218

- Natural language (NL), 101
- Natural Language Processing (NLP), 5, 34, 50, 87, 118, 207, 385, 433
- Natural Language Toolkit (NLTK), 101
- Needle-in-a-haystack, 59
- NER. *See* Named-Entity Recognition (NER)
- Network construction, 347
- Neural networks, 5, 21, 367
- Neural processes, 367
- Neuroplasticity, 469
- Neuroscience, 6, 12, 483
- Neuroscientific angle, 455
- Neuroscientific models, 506
- Neuro-symbolic AI, 21
- Neurotechnologies, 474, 488
- Newell, A., 13
- NL. *See* Natural language (NL)
- NLP. *See* Natural Language Processing (NLP)
- NLTK. *See* Natural Language Toolkit (NLTK)
- Noisy context, 447
- Non-binary identities, 469
- Non-interactive, 167
- Non-linear dimensionality reduction approach, 15
- Non-linear progressions, 271
- Nonstructural protein 1 (NS1), 178
  - antigen test for dengue, 186
- Notation convention, 69–70
- NSA. *See* National Security Agency (NSA)
- Obstructive jaundice, 187–191
- O’Dell, D. V., 88
- One Shot
  - formulation, 100
- learning, 51
- template, 99–100
- Online Virtual Assistance Lab (OVAL), 55
- OOD. *See* Out-Of-Distribution (OOD)
- OpenAI, 28, 49, 90
- Open-domain logical reasoning, 117
- Operation-Level, 297
- Operation State, 295, 303
- Operational estimators, 394
- Operational proxies, 397
  - open issues and UCCT view, 397
  - prior work, 397
- Optimization
  - and algorithm specifications, 204
  - EDT, 205
  - optimize initial condition, 204–205
  - optimize interaction dynamics, 205–206
  - optimizing convergence criteria, 206–207
  - techniques, 202
- Optimize interaction dynamics, 205
- Orchestrator, xv–xvi
- “Ostracization”, 466
- Out-Of-Distribution (OOD), 78, 393
- Output validation, 304
- OVAL. *See* Online Virtual Assistance Lab (OVAL)
- Overfitting, 6, 501
- Pacific Islander cultures, 461
- Panpsychism, 366
  - theory of, 367
- “Panspermia”, 465
- Paradox, 32

- Parameter-efficient fine-tuning techniques (PEFT techniques), 78
- Partisan annotations, value of, 208
- Pasenadi, K., 464
- Passenger Grouping Agent, 348
- Patience, 240
- Pattern access, 67
- Pattern density, 395  
estimation, 363  
manipulation, 404
- Pattern recognition, 395
- Pattern repositories, 267–268  
principle, 69, 398
- Paul, R., 88, 372
- PCA. *See* Principal Component Analysis (PCA)
- PCR, 186
- PEFT techniques. *See*  
Parameter-efficient fine-tuning techniques (PEFT techniques)
- Pensive, 233
- Perception, 369
- Persistent Context Repository, 314
- Persistent Execution Memory, 323
- Persistent memory, 289, 328
- Pilot study, 106–108  
example article, 106
- Plan Validator Agent, 351
- Planning phase, 349
- Plutchik, R., 229–230, 233, 251, 254, 266
- PO. *See* Preference optimization (PO)
- Polarized emotional responses, 264
- Polarized emotions in one article, 277–278  
letter to Sophie, 279
- original text of poem “To My Sister”, 278
- soldier’s letter during war, 278–279
- will of Laura Mary Octavia Lyttelton, 279–280
- Polydisciplinary, 30–31, 511  
capabilities, 454  
synthesis, 513
- Polydisciplinary knowledge, 30  
fusion, 21
- Polymodal, 32
- Polymodality, 32
- Polynthesis, 496  
from breadth to depth, 471  
dialogue round, 457, 462, 468  
dialogues, 486  
future directions, 492–493
- GPT-A answers GPT-B, 460–462, 467–468, 476–477, 481–482, 487–488
- GPT-A asks GPT-B, 457–458, 462–463, 468–469, 473–474, 477–478, 483
- GPT-A’s, 489
- GPT-A’s feedback to moderator, 484
- GPT-B answers GPT-A, 458–459, 464–465, 470–471, 474–476, 478–480, 484–485
- GPT-B asks GPT-A, 459–460, 465–467, 476, 480–481, 486
- GPT-B’s, 489–490
- implications for AI-assisted discovery, 492
- limitations and methodological challenges, 491
- methodological refinements, 491–492

- moderator initializes agent
  - GPT-A, 456–457
- moderator’s intervention,
  - 471–473
- warm-up breadth probing, 456
- Post-debate conciliatory remarks,
  - 152–153
- Posterior
  - distribution, 392
  - log-odds, 417
  - odds, 417
  - selection, 419
  - shorthand, 79
- Posterior density amplification, 75–76
  - procedure, 77
  - takeaway, 77
- Post-retrieval processing, 58
- Post-training value alignment, 32–33
- Practitioner gap, 76
- Pragmatic, 32
- Predictive test, 72, 76
- Predictive theory, gaps in, 68
- Pre-execution validation, 304
- Preference optimization (PO),
  - 252–253
- Prefrontal cortex, 43, 458, 464
- Pre-training censorship, 33
- Pricing agent, 351
- Pride, 269–271
- Primary critique, 53
- Principal Component Analysis (PCA),
  - 15
- Prior log-odds, 417
- Prior reference cluster, 392
- Probabilistic models, 9, 16–18
- Problem-solving, 2, 372
- Problematic content, 250
- Productive tension, 509
- Prompt engineering, 49
  - prompting methods, 50–57
  - RAG, 57–60
- Prompt template engineering, 99
  - basic One Shot template, 99–100
  - clarification with definition, 100–101
  - counterfactual reasoning, 105
  - generalization with maieutics method, 103–105
  - verification with elenchus method, 101–103
- Prompt(s), 330–331
  - design, 49, 399
  - prompt-based techniques, 67
- Prompting, 87, 90
  - access, 67
  - chain of thought, 52–53
  - Few Shots, 51–52
  - further improvement techniques, 54
  - illustrative examples, 54
  - many-shots example, 55–57
  - methods, 50
  - One Shot, 51
  - prompt example, 54–55
  - tree of thoughts, 53–54
  - Zero Shot, 50–51
- Prudence, 240
- Psychology support, 231
- Public awareness, 479
- Public discourse, 485
- Public frontier models, 171
- Public-private collaboration, 152
- Public-private partnerships, 146
- Pullman, P., 461–462
- Purity, 240
  - Qualifying and quantifying emotions, 229–231

- behavioral emotion analysis
  - model, 232–233
- emotion inclusion and exclusion criteria, 233
- observations, 231
- Plutchik's wheel of emotions, 230
- Qualifying and quantifying ethics, 238
  - ethical alignment with context, 242–243
  - ethical violation correlates to emotions, 238–239
  - twelve vices and twelve pairs of vices, 239–241
  - wheel of virtue, 241–242
- Qualitative anchoring
  - demonstrations, 72
- Few Shot semantic re-anchoring, 72–73
  - fine-tuning, 77–79
  - multi-agent, 79–80
  - retrieval, 75–77
  - vision, 73–75
- Qualitative geometry, 412
- Qualitative reasoning assessment, 206
- Qualitative shift, 398
- Quality
  - combine predictions weighted by, 183–184
  - evaluation, 122
- Quantitative anchoring control, 363
- Quantitative metrics, 206
- Quantitative predictiveness, 80
- Quantum-computing query with RAFEL and UCCT, 447–449
- Queue Reordering, 327, 334
- RAFEL. *See* Retrospective and Adaptive Framework for Enhancing LLMs (RAFEL)
- RC&B. *See* Recollection and Comprehension with Biases (RC&B)
- RC&H. *See* Recollection and Comprehension with Hallucination (RC&H)
- Reactive planning, 309
  - ALAS three-layer architecture for, 324–325
  - under disruptions, 352
  - experimental results of, 331–332
- Reactive prompt, 330
- REALM. *See* Retrieval-Augmented Language Model (REALM)
- Reasonableness
  - evaluation module, 203
  - scrutinizing with, 203–204
- Reasoning
  - capabilities, 442
  - evaluation, 113
  - and explanation, 20
  - technique, 87
- Recollection
  - and comprehension, 438, 443–444
  - method, 94
- Recollection and Comprehension with Biases (RC&B), 441
- Recollection and Comprehension with Hallucination (RC&H), 441
- Reconciliation, 428, 461
- Recovery protocol execution, 305
- Rectified Linear Unit (ReLU), 267
- Refinement, 326, 348
  - agent, 351

- Reflecting emotional depth, 272
- Reflective LLM collaboration
- AllSides, 201
  - annotation quality, 215–216
  - bias detection, 207–212
  - bias mitigation, 212–214
  - biased ground truth, 201–202
  - contribution, 202
  - debiasing procedure, 203–204
  - experiments, 207
  - justifications of biased articles, 218–223
  - label validation, 199–201
  - methodology, 202
  - optimization and algorithm specifications, 204–207
  - quality distribution of Wikipedia pages, 201
  - related work, 199
  - summary of EVINCE debate on news D1, 216–218
- Reflective LLM dialogue framework (RLDF), 198
- approach, 202
- Regime selection model, 392
- Regulating language, 255–257
- Regulatory foundation, 507
- Reinforcement learning (RL), 5, 376, 445
- algorithm, 374
- Reinforcement Learning from Human Feedback (RLHF), xxiii, 79, 134, 242, 247, 252–253, 386, 505
- Reinforcement learning with AI feedback (RLAIF), 252–253
- Relative Preference Optimization (RPO), 253
- ReLU. *See* Rectified Linear Unit (ReLU)
- Remediation strategies, 437, 444
- implementation considerations, 445
  - selecting appropriate intervention, 444–445
  - sourcing data for remediation, 445
- Renewable energy sources, 478
- Repetition, 405
- Representational adaptation, 395
- Representational distance, 395
- Representing data, 15
- Republican scandals
- bias-rating distributions, 212
  - comparison of bias assessments, 211
  - illustrates distribution of bias ratings, 211
  - results on, 210–212
- Resource Allocation Agent, 348, 350
- Resource equity, 152
- Retrieval access, 67
- Retrieval-Augmented Generation (RAG), xxii, 29, 49, 57, 63–64, 75, 393, 397, 434–437, 445, 449, 511
- ablations, 425
  - admission, 423
  - architecture and data flow, 57
  - baselines, 424
  - budget, 423
  - control, 423
  - datasets, 424
  - diversity within relevance, 424
  - encoders, 422
  - evaluation plan, 424
  - with limited-context LLMs, 58–59
  - with long-context LLMs, 59–60
  - metrics, 425

- mismatch, 423
- notation and estimators, 422
- notes and limitations, 425
- open issues and UCCT view, 397
- practical procedure, 424
- prior work, 397
- quality and cohesion, 422–423
- quality rubric, 422
- representative work, 58
- scope, 421
- selection and ordering under budget, 423
- stopping, 424
- takeaway, 425
- UCCT view of, 421
- Retrieval-Augmented Language Model (REALM), 397
- Retrieval-Enhanced Transformer (RETRO), 397
- Retrieval techniques, 58
- Retriever-generator coupling, 436
- RETRO. *See* Retrieval-Enhanced Transformer (RETRO)
- Retrospective and Adaptive Framework for Enhancing LLMs (RAFEL), 434
- quantum-computing query with, 447–449
- Retrospective learning, 437
- benchmarking, 438
- DIAG, 438–440
- PRBE, 440–444
- RAFEL with four phases, 438
- remediation strategies, 444–445
- Retrospective pilot, 518
- Revised statement, 260
- Reward-Driven Behavior, 383
- RL. *See* Reinforcement learning (RL)
- RLAIF. *See* Reinforcement learning with AI feedback (RLAIF)
- RLDF. *See* Reflective LLM dialogue framework (RLDF)
- Robotics, 2, 5
- Robust

  - AI system, 481
  - frameworks, 485

- Role-conditioned anchoring, 79
- EDT pairing, 80
- EVINCE at glance, 80
- handoff, 80
- high-level UCCT formulation, 79
- posterior shorthand, 79
- scope of section, 80
- traditional debates fail, 79
- UCCT interpretation of signals, 80
- Route Planning Agent, 347, 350, 352
- RPO. *See* Relative Preference Optimization (RPO)
- Rule-based AI model, 7–8
- Rule-based chatbots, 13
- Rule-based era, 13–14
- Rule-based systems, 14, 21
- Runtime layer, 327
  - disruption handling, 327–328
  - master coordinator and memory logging, 328
- Runtime monitor, 325
- Russell, S., 228
- Sad, 233
- Sadness, 234
- SagaCoordinatorAgent, 302
- SagaLLM, 289
  - classical sagas *vs.* SagaLLM, 293
  - compensatory analysis, 315

- context management
  - requirements, 297
- design with travel planning, 297–305
- experimental design, 306
- experiments, 306
- independent validation
  - requirements, 296
- LLM limitations necessitating
  - SagaLLM's key requirements, 291–292
- multi-agent LLM frameworks
  - and transaction limitations, 292–293
- observations, 315
- related work, 290
- remediation, 309, 313–315
- system requirements for, 293–294
- Thanksgiving dinner problem, 306–309
- transaction management
  - systems, 290–291
- transactional integrity
  - requirements, 294–296
- wedding reunion logistics problem, 310–315
- Sayre, Z., 263, 274
- Scalable intensity, 233, 254
- Scalable orchestration, 180
- Scale-Invariant Feature Transform (SIFT), 34, 514
- Scaling law problem, 35
- Schachter–Singer theory, 252
- Scheduling Agent, 348
- Scherer, K. R., 231, 254, 266
- Schrödinger, E., 369
- Schrodinger, E.
  - model, 370
- Schwarz, N., 252
- Sedol, L., 513
- SeEvo work. *See* Self-evolutionary work (SeEvo work)
- Select relevant methods, 90
- Selectivity, 76, 423
- Self-consistency, 396
- Self-evolutionary work (SeEvo work), 332
- Self-regulation, 126, 139
- Self-respect, 239
- Self-supervised behaviors, 256
- Self-supervised encoder, 74
- Self-supervised learning, 20
  - algorithm, 256
- Self-Supervised with Human Feedback (SSHF), 242
- Self-validation, 167
  - gap necessitating independent validation, 291
- Semantic anchoring, 64
  - principle, 69, 398
- Semantic distance computation, 363
- Semi-supervised learning methods, 200
- Sensory organs, 370
- Sentiment analysis, 2
- Sequence Likelihood Calibration (SLiC), 253
- Sequential planning, experimental results of, 331
- Serenity, 233
- Serology testing, 131, 186, 192
- Serotonin levels, 376
- Set similarity, 9
- SFT. *See* Supervised Fine-Tuning (SFT)
- Shakespeare, W., 234
- Shame, 272
- Shannon entropy, 198
- Shortest Processing Time (SPT), 332

- Shortest Total Processing Time (STPT), 332
- SI. *See* Synaptic intelligence (SI)
- SIFT. *See* Scale-Invariant Feature Transform (SIFT)
- Similarity, 2, 12
- cognitive psychology perspective, 9–12
  - computer science perspective, 7–9
  - linguistic perspective, 6–7
  - perspectives on, 6
  - section remarks, 12–13
- Similarity measurement, 19–21
- context-aware era (2010s–), 19–20
  - data-centric era (2000s–), 18–19
  - eras of, 13
  - impact on, 19
  - model-based era (1970s–), 14
  - rule-based era (1950s–), 13–14
  - section remarks, 21–22
- Simon, H. A., 13
- Simple linear mapping, 217
- Single-agent refinements, 167
- SLiC. *See* Sequence Likelihood Calibration (SLiC)
- Slow-moving regulation, 143
- Smith, C. A., 251
- Social cognitive theory, 372
- Social contexts, 271
- Social media
- campaigns, 470
  - platform, 257
- Societal preparation, 485
- SocraSynth, xxii, 21, 29, 36–37, 113–114, 116, 165, 227, 444, 496, 510
- agent A's fourth round, 145–146
  - agent A's opening arguments, 138–139
  - agent A's second round, 140–142
  - agent A's third round, 143–144
  - agent B's counterarguments, 139–140
  - agent B's fourth round, 146–147
  - agent B's second round, 142–143
  - agent B's third round, 144–145
  - agents and roles, 116
  - agreed upon debate topics and descriptions, 125
  - algorithm, 120–122
  - conditional statistics, 114
  - contentiousness parameter, 132–133
  - CRIT evaluations by judges, 147–151
  - debate beats Q&A in information quality, 127–128
  - debate beats Q&A in quality, 151–152
  - debate from opening to fourth round, 138
  - decomposing debate subject into topics, 124
  - empirical study, 123–124
  - evaluative stage, 127
  - fourth round, 145
  - generative stage, 124–127
  - healthcare debate, 153–157
  - moderator's prompts, 136–138
  - modulating debate with contentiousness, 115
  - multi-agent SocraSynth, 116–123
  - opening round, 138
  - platform, 31
  - policy discussion, 124

- post-debate conciliatory remarks, 152–153
- refine context to mitigate hallucinations, 115–116
- remarks on related work, 133–134
- second round, 140
- supplemental materials, 136
- symptom checking, 128–131
- tables of topic proposal and refinement, 136
- third round, 143
- Socratic method, 87–88, 92–94
  - counterfactual reasoning, 98–99
  - illustrative critical reading example, 94–95
  - method of definition, 95–96
  - method of dialectic, 97–98
  - method of elenchus, 96–97
  - method of maieutics, 98
  - remarks on CRIT, 99
- Socratic symposium, 114
- Socratic synthesis, 114
- Socratic “teaching through questioning” method, 93
- Socratic tradition, 507
- Soldier’s letter during war, 278–279
  - “To My Sister” poem rewritten to reflect negative emotion
  - “despair”, 279
- Sophie, letter to, 279
- Spatial memory enables perspectives, 517
- Spatiotemporal historian (SAGA), 518
- Specialized agent distribution, 316
- Speculative reasoning, 396
- SPT. *See* Shortest Processing Time (SPT)
- SSHF. *See* Self-Supervised with Human Feedback (SSHF)
- Stakeholder, 141–142
- State commitment, 304
- State management, 303–304
- State-space analysis, 346–348
- Statelessness necessitating transactional integrity, 291
- Statistical inference and similarity, 16
- Stimulus-response model, 365
- Stoddard, H. A., 88
- STPT. *See* Shortest Total Processing Time (STPT)
- Structure-aware variants, 436
- Structured exploration, 492
- Structured validation protocols, 316
- Subcritical regime, 71, 418–419
- Success probability model, 400
- Supercritical regime, 71, 418–419
- Superintelligent AI, 482
- Supervised Fine-Tuning (SFT), 393
- Supervised learning, 5
- Support Vector Machines (SVMs), 15–16, 33
- Supporting arguments, 217
  - final agreement and consensus, 218
  - first rebuttal and counter-rebuttal, 217
  - initial assessments, 217
  - moderate tone reassessment, 217–218
- Susceptibility, 247
- SVMs. *See* Support Vector Machines (SVMs)
- Symbol swap, 73
- Symbolic tag, 74
- Symptom checking, 128–129

- final joint recommendations to GP, 130–131
- hepatitis *vs.* jaundice, 129–130
- superiority of debate over Q&A in gaining insights, 131
- Synaptic intelligence (SI), 78
- Synchronization mechanisms, 375
- Synsets, 7
- System administrators, 257
- Tabu Search (Tabu), 334
- Taillard (TA), 332
- Task, 72
- execution agents, 301–302
  - exemplars, 79
- Teaching mode, 106
- Temperance, 240
- Template construction, 325
- Template Construction Layer, 325
- Template workflow, 350
- Temporal Constraint Agent, 348
- Temporal memory enables retrospectives, 517
- Text analysis, 2
- Textual data, 15
- Thanksgiving dinner problem, 306
- commonsense augmentation, 306
  - context narrowing, 307–309
  - reactive planning for Thanksgiving dinner after James, 308
  - SagaLLM remediation, 309
  - Thanksgiving dinner coordination problem, 307
- Thanksgiving dinner planning workflow with commonsense augmentation, 308
- Three-branch framework design, 253–254
- BEAM, 254–255
- behavior analysis, 259
- DIKE, 255–257
- emotion modulation, 259
- ERIS, 257–259
- illustrative example, 259
- revised statement, 260
- Threshold calibration, 419
- calibrating critical level  $S_c$ , 419–420
  - validation protocol, 420
- Threshold effects, 67
- Threshold validation, 393, 403
- Threshold-based attention mechanisms, 384
- Threshold-based transitions, 381
- Threshold-crossing behavior, 370
- Threshold-crossing dynamics, 71
- identification with E1 logistic calibration, 419
  - interpretation, 72
  - mathematical characterization, 72
  - posterior odds and Bernstein bound, 417
  - regimes of activation, 71
  - from selection to task success, 418
  - setup and assumptions, 416–417
  - sharp transition and width, 418
  - subcritical regime, 418
  - supercritical regime, 418
  - theorem, 416
  - transition width, 418
- Threshold-crossing principle, 69, 398
- Threshold-like flips, 72
- Tononi, G., 364

- Top-down
  - attention, 370
  - processing, 366
- ToT. *See* Tree-of-Thought (ToT)
- Traditional MFQs, 374
- Traditional optimization methods, 322–323
- Traditional quantum-end mechanism, 374
- Traditional solvers, 335
- Train Booking Compensation, 305
- TrainBookingAgent, 302
- Training
  - datasets, 74
  - documents, 228
- Training data, 35
  - potential, 178
- Transaction
  - execution, 304
  - management systems, 290–291
  - properties, 311–313
- Transaction flow, 304
  - comprehensive validation framework, 304
  - transaction execution sequence, 304
- Transaction limitations, 292
  - limited context preservation, 293
  - missing transaction semantics, 292
  - validation gaps, 292
- Transaction-specific compensations, 304–305
- Transactional consistency, 290
- Transactional integrity, 288, 293
  - dependency tracking and compensation planning, 295–296
  - requirements, 294
- transaction state management, 295
- transaction state management in SagaLLM, 295
- Transactional scenarios, 291
- Transactional state preservation, 316
- Transformer model, 7
- Transition width, 418
- Transparency, 485
- Transparent reasoning, 178
- Transportation scheduling, 329
  - comparison of ride-sharing solutions generated by ALAS and baseline LLMs, 331
  - dynamic urban ride-sharing, 330
- experimental results of reactive planning, 331–332
- experimental results of sequential planning, 331
- problem specification, 330
- prompts, 330–331
- Travel coordination agent records, 309
- Travel planning, design with, 297
  - agent architecture and code structures, 301–304
  - architecture of SagaLLM, 298
  - compensation and recovery mechanisms, 304
  - recovery protocol execution, 305
  - specifications, 297–299
  - transaction flow and validation protocol, 304
  - transaction-specific compensations, 304–305
  - travel planning problem, 297
  - two-phase workflow architecture, 299–301

- Traveling Salesman Problem (TSP), [323](#)
- Tree-of-Thought (ToT), [292](#)  
prompting, [53–54](#)
- Treisman, A., [9](#)
- Trust-disgust, [231](#)
- TSP. *See* Traveling Salesman Problem (TSP)
- Turner, M., [252, 512](#)
- Two Shot anchors, [72–73](#)
- Two-phase workflow architecture, [299](#)  
automated SagaLLM execution  
overview, [300](#)  
detailed phase 2  
implementation, [300–301](#)  
manual itinerary planning, [300](#)  
travel-planning workflow  
showing transition, [299](#)
- UCCT. *See* Unconscious-Conscious Complementarity Thesis (UCCT)
- Unconscious patterns, [64](#)  
repository, [64, 68](#)
- Unconscious processes, [366](#)
- Unconscious reflection, [516](#)
- Unconscious-Conscious Complementarity Thesis (UCCT), [251, 263, 266](#)  
failure modes, [268–269](#)  
implications for love letter experiment, [268](#)  
interpretation of signals, [80](#)  
LLMs, [269](#)  
mathematical foundations, [267–268](#)  
nature of unconscious processing, [267](#)  
threshold crossing, [267](#)
- Unification, gaps in, [68](#)
- Unified Cognitive Consciousness Theory (UCCT), [xvi, xxii, 35, 63–64, 414, 435, 496](#)  
accessing  $P_{\text{prior}}$  via Zero Shot behavior, [71](#)
- aim, [393](#)
- algorithm, [401](#)
- anchoring instantiations, [70, 400](#)
- anchoring strength, [70](#)
- assumptions, [401](#)
- background, limitations, and experimental rationale, [394](#)
- calibrating critical level  $S_c$ , [402](#)
- chain-of-thought, [396](#)
- cognitive foundations and consciousness theory, [66–67](#)
- comparing anchoring methods, [402](#)
- contrast and UCCT's novelty, [68](#)
- contributions, [394](#)
- core theoretical architecture, [69, 398–399](#)
- cross-base accuracy change after fine-tuning, [408](#)
- design implications for prompt engineering, [402](#)
- diversity within relevance, [420–421](#)
- effective evidence strength, [401](#)
- emergent behavior and threshold effects, [67](#)
- established, [414–415](#)
- experiment-specific metrics, [403–412](#)
- experiments, [403](#)
- explanation, [396](#)
- Few Shot accuracy *vs.* shots for bases, [407](#)

- Few Shot learning statistics from sigmoid fits, 405
- Few Shot success model, 71 framework, 68–69, 398 gaps in unification and predictive theory, 68 geometric trajectory analysis, 396–397 implication for subsequent experiments, 412–413 in-context learning, 394–395 in-context learning as Bayesian inference, 67 insight, 75 key contributions, 66 limitations and future work, 413–416 MAD experiment plan, 429–430 main theorem, 401–402 mathematical foundations, 69, 399 maximum likelihood *vs.* semantic anchoring in, 65 measuring pattern activation experimentally, 400–401 mechanistic interpretability and internal circuits, 68 Meta-LLaMA-3.1-8B-instruct, 411 modeling capability shift, 399 for multi-agent debate, 425–429 multi-agent debate with EVINCE, 397–398 notation bridge, 394 notation convention, 69–70 open issues and UCCT view, 397 outlook, 403, 415 Phi-4, 412 pre-debug test, 76 predictions for ICL effectiveness, 395 prompting, retrieval, and pattern access, 67 qualitative anchoring demonstrations, 72–80 quantifying anchoring strength, 400 Qwen3-14B, 413 for RAG, 421–425 RAG and operational proxies, 397 related work, 66 roadmap, 394 scope, 66 scope generalization after SFT, 408 setup, 421 shift occurs, 401 success probability model, 400 threshold calibration, 419–420 threshold crossing and phase transitions, 401 threshold-crossing dynamics, 71–72, 416–419 UCCT+RAFEL, 445 UCCT-guided diagnosis, planning, and control, 77–79 unifies major adaptation methods, 70 validating, 403 view, 76, 398 view of RAG, 421 Unknown knowns, 490, 493, 517 Unknown unknowns, 30–31, 454, 457 deep-dive phase, 43–44 expansive conversation, 41–45 exploratory phase, 42–43 exploring, 36–37 first iteration, 38

- observations, 40–41, 45
- scientific debate, 37–41
- second iteration, 38–40
- synthesis of key points in consciousness debate, 42
- Unlabeled pre-training, 74
- Unsupervised learning, 5
- Unsupervised sensory, 74
- Urban Ride Sharing (URS), 323, 345–346
  - agent placement in urban ride sharing network, 350
  - agent specifications and protocols, 349
  - generating planner W\* walkthrough, 346
  - output, 348
  - reactive planning under disruptions, 352
  - simplified URS problem statement, 346
  - state-space analysis, 346–348
  - from workflow template to execution workflow, 348–352
- Urine test for zika, 186
- URS. *See* Urban Ride Sharing (URS)
- “Uthenticity”, 477
- Validation, 326, 348, 498
  - agents, 322
  - function, 122
  - gaps, 292
  - response protocols, 296
- Validation protocol, 304, 420
  - comprehensive validation framework, 304
  - transaction execution sequence, 304
- Vapnik, V., 18
- Vector space models, 8–9, 14–16
- Vehicle Specialization Agent, 351
- Vices, twelve vices and twelve pairs of, 239–241
- Virtual Assistant, 377
- Virtual realities, 469
- Virtue, wheel of, 241–242
- Virtuous cycle, 508
- Vision Few Shot Learning, 399
- Vision-only encoder, 74
- Visual cortex, 267
- Vulnerability, 247
- wa Thiong, N., 468
- Wang, M., 386
- Wasserstein distance (WD), 80, 168, 189, 398
- Watson, J., 102
- WD. *See* Wasserstein distance (WD)
- Wedding reunion logistics problem, 310
  - Claude 3.7 ignored traffic delay, 312
  - context narrowing, 310–311
  - DeepSeek’s failed reactive schedule after traffic alert, and GPT-4o made similar errors, 313
  - LLM limitations and SagaLLM remediation, 313–315
  - LLMs *vs.* SagaLLM on context management, 315
  - transaction properties, 311–313
  - wedding reunion logistics schedule, 311
  - wedding reunion reactive schedule by GPT-o1 indicates, 314

- Weighted Minkowski distance function, 8–9
- Weighted-Minkowski function, 11
- Whack-a-Mole game, 248  
limitations, 255
- White Box Analysis, 334–335
- WHO. *See* World Health Organization (WHO)
- Wikipedia, 1, 201
- WIP. *See* Work-In-Progress (WIP)
- Word embeddings, 20
- Wordsworth, W., 262, 280
- Work-In-Progress (WIP), 327
- Workflow
  - complete meta-planner algorithm for workflow generation, 337
  - template to execution workflow, 348–352
  - workflow-level, 297
- Workflow blueprinting layer, 325
- agent role specification, 325–326  
validation and refinement, 326
- workflow template construction, 325
- World Health Organization (WHO), 379
- World State, 464
- Xing J., 376
- YAWL. *See* Yet Another Workflow Language (YAWL)
- Yellowish skin and eyes, 130
- Yet Another Workflow Language (YAWL), 290
- Zero Shot
  - accessing  $P_{\text{prior}}$  via Zero Shot behavior, 71
  - baseline, 72
  - learning, 50–51
- Zika, urine test for, 186
- Zoroastrianism, 468

# Multi-LLM Agent Collaborative Intelligence

*The Path to Artificial General Intelligence*

**Edward Y. Chang**

Today's large language models excel at pattern recall yet falter on long-range planning, self-critique, context loss, and the tendency of maximum-likelihood training to reward popularity over quality. MACI offers a promising route to AGI by orchestrating specialized LLM agents through explicit protocols rather than enlarging a single model. Several modules remedy complementary weaknesses: adversarial-collaborative debate surfaces hidden assumptions; critical-reading rubrics filter incoherent arguments; information-theoretic signals steer dialogue quantitatively; transactional memory enables reliable long-horizon execution; and a dual-agent ethical court adjudicates outputs. Crucially, MACI also modulates linguistic behavior, tuning each agent's contentiousness and emotional tone, so the collective explores ideas from contrasting, affect-aware perspectives before converging.

Fourteen aphorisms distill the framework's philosophy, including:

- Intelligence emerges from regulated collaboration, not isolated brilliance
- Exploration must remain in tension with exploitation

Across healthcare diagnosis, investment support, scheduling, supply-chain management, and news-bias mitigation, MACI ensembles deliver significant improvements in reasoning depth, planning horizon, and reliability compared with similar-sized single models. By uniting structured debate, information-theoretic coordination, persistent memory, affect-aware discourse, and deliberative ethics, MACI demonstrates that rigorously validated multi-agent collaboration provides a practical, interpretable path toward robust general intelligence.

## ABOUT ACM BOOKS



ACM Books is a series of high-quality books published by ACM for the computer science community. ACM Books publications are widely distributed in print and digital formats by major booksellers and are available to libraries and library consortia. Individual ACM members may access ACM Books publications via separate annual subscription.

ISBN 979-8-4007-3178-5



9 798400 731785