Multi-Model Reasoning Analysis: Leveraging Divergence as Signal

A Framework for Bias-Aware Argument Evaluation Using Four Test Models and Cross-Model Synthesis

Nida, R. (2025). Multi-Model Reasoning Analysis: Leveraging Divergence as Signal. Technical Report

Executive Summary

This monograph develops a systematic framework for understanding and leveraging the reasoning differences among Large Multimodal Models (LMMs). Instead of treating model divergence as a flaw, it reframes variation as an analytical asset—one that, when properly measured and weighted, produces more reliable, transparent, and defensible reasoning than any single model can provide.

Purpose

The study has two primary objectives:

1. **Identify and characterize consistent reasoning behaviors across leading LMMs**, including GPT-5.1, DeepSeek V2, Grok, Gemini 2.0, and Claude.
2. **Develop a practical, domain-specific ensemble method** that uses these differences to enhance accuracy and reduce bias in high-stakes decision-making.

Key Findings

1. Stable Reasoning Signatures

Each model displays consistent, predictable patterns:

- **GPT-5.1:** Methodical, assumption-sensitive, structurally rigorous
- **DeepSeek V2:** Decisive, compressed, aggressive
- **Grok:** Minimalist, stable, surface-analytic
- **Gemini 2.0:** Context-rich, broad, variable caution
- **Claude 3.5:** Neutral, synthesizer-oriented

These signatures persist across scientific, political, legal, economic, ethical, and metaphysical domains.

2. Predictable Divergence Types

Model disagreements fall into four analyzable categories:

- **Structural** (mapping differences)
- **Interpretive** (framing differences)
- **Evaluative** (different weighting of evidence)
- **Bias-driven** (safety or alignment effects)

Understanding *why* models disagree is more valuable than seeking unanimous agreement.

3. Ensemble Reasoning Outperforms Single Models

A domain-specific weighting system—validated across 13 test arguments—provides more stable outcomes. Examples:

- **Scientific:** 0.40 GPT-5.1 / 0.30 DeepSeek / 0.20 Gemini / 0.10 Grok
- **Ethical:** 0.40 Gemini / 0.30 GPT-5.1 / 0.20 DeepSeek / 0.10 Grok
- **High-Risk:** 0.50 GPT-5.1 / 0.30 Claude / 0.20 Gemini

A four-stage workflow (triage → structure → context → synthesis) offers a repeatable, auditable process for decision support.

4. Practical Applications Across Domains

The framework improves reasoning in:

- causal scientific analysis
- constitutional and statutory interpretation
- economic and financial modeling
- autonomous-system risk ethics
- corporate strategy and governance
- intelligence and national-security assessment

In every domain, ensemble reasoning surfaces hidden assumptions, reduces overconfidence, and identifies where arguments are strong, weak, or ambiguous.

5. Limitations and Future Work

Challenges remain: alignment artifacts, epistemic uncertainty, model drift, and the need for quantitative scoring. Future research includes an expanded argument suite, automated ensemble pipelines, longitudinal drift tracking, and calibration against real-world outcomes.

Conclusion

Model disagreement is not a problem to eliminate—it is information to analyze.
By measuring and strategically weighting LMM divergence, this framework transforms variability into a decision-support advantage, enabling clearer, more balanced, and more defensible reasoning in complex, high-stakes domains.

Contents

I. Introduction 8
1.1 Overview 8
1.2 Purpose of This Monograph 9
1.3 Why Bias Matters 9
1.4 Contributions 10
1.5 Document Roadmap 11
II. Background 12
2.1 Defining Bias in LMMs 12
2.2 Foundations in Argumentation Theory 13
2.3 Types of LMM Reasoning Failures 15
2.4 Why Models Disagree 16
2.5 Relevance to Multi-Model Evaluation 17
2.6 Prior Work and Related Research 18
2.7 Evolution of Large Multimodal Models 21
III. Methodology 24
3.1 Testing Framework and overview 24
3.2 Prompt Protocols 26
3.3 Session Header Standardization 28
3.4 Bias Mitigation Rules 28
3.5 Multi-Model Pipeline Architecture 28
3.6 Sampling Parameters and Replication Strategy 29
3.7 Scoring and Comparison Framework 32
3.8 Divergence as Diagnostic Tool 34
3.9 Data Collection, Documentation, and Storage Standards 35
3.10 Workflow 35
IV. Reasoning signatures of Each LMM 38
4.1 Overview 38
4.2 Reasoning Signature: GPT-5.1 42

4.3 Reasoning Signature: DeepSeek 46
4.4 Reasoning Signature: Grok 50
4.5 Reasoning Signature: Gemini 2.0 54
4.6 Synthesis Profile: Claude 3.5 Sonnet 58
4.7 Comparative Summary of Reasoning Signatures 63
V. Argument Test Suite 67
5.1 Round 1 Argument Set: Evolution & Intelligent Design (Arguments 1–5) 67
5.2 Round 2 Argument Set: Cross-Domain Challenge Suite (Arguments 6–13) 71
5.3 Argument Formatting and Standardization 76
5.4 Rationale for Argument Selection 76
5.5 Strengths and Limitations of the Suite 77
5.6 Summary 77
VI. Cross-Model Comparison Framework 78
6.1 Overview 78
6.2 Agreement Analysis 79
6.2.1 Full Agreement 79
6.2.2 Partial Agreement 80
6.2.3 Structured Disagreement 80
6.2.4 Contradictory Conclusions 80
6.3 Divergence Types 81
6.3.1 Structural Divergence 81
6.3.2 Interpretive Divergence 82
6.3.3 Bias-Driven Divergence 82
6.3.4 Evaluation Divergence 82
6.4 Hidden Assumption Sensitivity 83
6.5 Bias Signature Assessment (formerly “Scoring”) 84
6.5.1 Caution Level 84
6.5.2 Decisiveness 84
6.5.3 Neutrality 84
6.5.4 Assumption Detection Strength 84

6.5.5 Drift Likelihood 85
6.6 Model Clustering and Reasoning Families 85
6.6.1 Conservative Evaluators 85
6.6.2 Aggressive Evaluators 85
6.6.3 Hybrid Generalists 86
6.7 Framework Application Guidelines 86
6.8 Summary 89
VII. Case Study Results 89
7.1 Overview 89
7.2 Political and Legal Arguments 90
7.2.1 Risk tolerance and safety activation 90
7.2.2 Competing interpretive frames 90
7.2.3 Voter fraud and evidential standards 91
7.3 Economic Arguments 91
7.3.1 Efficiency as an economic vs. ethical construct 91
7.3.2 Evidence weighting and uncertainty 92
7.4 Ethical and Moral Arguments 92
7.4.1 Depth vs. discipline in metaphysical reasoning 92
7.4.2 Hidden assumptions and divergent verdicts 93
7.4.3 Alignment behavior 93
7.5 Metaphysical and Epistemic Arguments 93
7.5.1 Convergence on structure 93
7.5.2 Rigor vs. abstraction 94
7.5.3 Protocol and drift patterns 94
7.6 Detailed Case Study Analysis 94
7.6.1 Case Study 1 – Anthropogenic Global Warming (Strong, Moderate, Weak Variants) 95
7.6.2 Case Study 2 – Second Amendment (9A/9B) 96
7.6.3 Case Study 3 – Problem of Evil (13) 97
7.7 Summary 97

VIII. Synthesis 98
8.1 Integrated Cross-Model Patterns 98
8.2 The Ensemble Reasoning Framework 99
8.2.1 Why Ensembles Are Required 99
8.2.2 Core Principle 99
8.3 Weighted Model Profiles by Domain 100
8.3.1 Empirical / Scientific Arguments 100
8.3.2 Normative / Ethical Arguments 100
8.3.3 Political / Legal Arguments 100
8.3.4 Economic / Policy Arguments 101
8.3.5 Metaphysical / Philosophical Arguments 101
8.3.6 High-Risk / Safety-Sensitive Domains 101
8.3.7 Applying the Weights in Practice 102
8.4 Ensemble Workflow Recommendations 103
8.5 Strengths and Limitations of Ensemble Reasoning 104
8.6 Summary 104
IX. Applications 105
9.1 Overview 105
9.2 Scientific and Technical Assessment 105
9.3 Policy and Legal Analysis 106
9.4 Economic and Financial Modeling 108
9.5 Ethical and Normative Reasoning 108
9.6 High-Risk and Safety-Critical Domains 109
9.7 Corporate and Strategic Decision Support 109
9.8 Intelligence, Defense, and Operational Planning 110
9.9 Education, Research, and Pedagogy 110
9.10 Summary 111
X. Constraints and Considerations 112
10.1 Overview 112
10.2 Epistemic Constraints 112

10.2.1 Absence of Ground Truth Access 112
10.2.2 Sensitivity to Argument Framing 112
10.2.3 Lack of Domain Calibration 113
10.3 Methodological Constraints 113
10.3.1 Protocol Dependency 113
10.3.2 Reproducibility Challenges 113
10.3.3 Divergence Interpretation Requires Expertise 113
10.4 Operational and Resource Constraints 114
10.4.1 Multi-Model Access Requirements 114
10.4.2 Cost and Latency 114
10.4.3 Analyst Training 114
10.5 Safety and Alignment Constraints 115
10.5.1 Activation of Safety Filters 115
10.5.2 Alignment Drift Across Models 115
10.5.3 Ethical Use Requirements 115
10.6 Domain-Specific Constraints 116
10.6.1 Political and Legal Domains 116
10.6.2 Scientific and Technical Domains 116
10.6.3 Economic and Policy Domains 116
10.6.4 Ethical and Metaphysical Domains 116
10.7 When Ensemble Reasoning Should *Not* Be Used 117
10.8 Complementary Human-in-the-Loop Requirements 117
10.8.1 Human Judgment for Final Decisions 117
10.8.2 Documentation Requirements 117
10.8.3 Cross-Validation with External Sources 117
10.9 Summary 118
XI. Expanded future research opportunities 118
11.1 Overview 118
11.2 Expanding the Diagnostic Argument Suite 119
11.2.1 New Argument Classes 119

11.2.2 Increased Complexity and Layering 119
11.2.3 Real-World Case Benchmarks 119
11.3 Quantitative Scoring and Calibration 120
11.3.1 Weighted Divergence Index (WDI) 120
11.3.2 Agreement Stability Metrics 120
11.3.3 Calibration With External Data 120
11.4 Automation and Workflow Integration 121
11.4.1 Automated Pipeline Execution 121
11.4.2 API-Based Ensemble Orchestrators 121
11.4.3 Enterprise and Government Integration 121
11.5 Longitudinal Drift and Model Evolution 121
11.5.1 Drift Tracking Across Versions 121
11.5.2 Temporal Divergence Maps 122
11.5.3 Version-Based Calibration Tables 122
11.6 Domain-Specific Extensions 122
11.6.1 Legal & Judicial Applications 122
11.6.2 Scientific & Engineering Applications 122
11.6.3 Defense & Intelligence 123
11.6.4 Corporate Decision Systems 123
11.7 Open Challenges 123
11.7.1 Distinguishing “True” Error from Productive Divergence 123
11.7.2 Measuring the Value of Structured Disagreement 123
11.7.3 Single-Model Collapse Scenarios 123
11.7.4 Human–AI Interaction Effects 124
11.8 Summary 124
XII. Conclusion 124

Figures

Figure 1: Workflow Diagram 37
Figure 2: Reasoning Space Quadrant 38

Figure 3: Divergence Type Overlap Diagram 78
Figure 4: Ensemble Workflow Diagram 104

Tables

Table 1: Comparative Reasoning Signature Matrix 64
Table 2: Round 1 Argument Reference Table 68
Table 3: Round 2 Argument Reference Table 72
Table 4: Model Sensitivity Notes 83
Table 5: Recommended Model Weightings by Domain 102
Table 6: Multi-Model Evaluation of a Causal Claim 106
Table 7: Recommended Model Weightings by Domain 107

I. Introduction

1.1 Overview

Large Multimodal Models (LMMs) such as GPT-5.1, DeepSeek V2, Grok, and Gemini 2.0 have become core analytical instruments across policy, business, research, and security domains. These systems process vast multimodal inputs—text, images, structured data, and sometimes audio or code—to generate structured reasoning, classifications, evaluations, predictions, and strategic recommendations.

Yet even as their capabilities grow, these models are not interchangeable. Each exhibits distinct, stable **reasoning signatures** shaped by:

- training data composition
- architectural design choices
- reinforcement-learning alignment
- safety-layer constraints
- token-level heuristics and optimization strategies

Because of this, two models evaluating the same argument under identical constraints will often arrive at different inferences, weights, assumptions, and conclusions. These differences are frequently interpreted as errors or inconsistencies—but in practice they offer one of the most powerful diagnostic tools available for deeper human understanding.

This monograph presents the first systematic framework for treating model disagreement not as a problem, but as data—a resource for improved analysis, cross-validation, and decision support. The evaluation uses a **two-round structure**: Round 1 establishes clean baseline reasoning signatures using Evolution and Intelligent Design arguments, while Round 2 tests cross-domain stability across political, legal, economic, empirical, ethical, and metaphysical argument classes.

For example, when presented with the same climate argument, GPT-5.1 reconstructs the logical structure carefully before evaluating causal claims, DeepSeek commits rapidly to a strong conclusion, Grok provides a concise surface-level analysis, and Gemini integrates broad contextual information but may drift. These differences—stable across domains—constitute each model's reasoning signature.

1.2 Purpose of This Monograph

This work establishes a comprehensive methodology for understanding, characterizing, and leveraging predictable LMM biases. Instead of attempting to neutralize bias or select a single "best" model, we argue for an alternative approach:

Use cross-model divergence as a signal that reveals hidden assumptions, structural weaknesses, and interpretive ambiguity.

Our goals are to:

1. Design repeatable, bias-aware evaluation pipelines using standardized arguments and protocols.
2. Characterize each major model's reasoning profile, including its strengths, weaknesses, and blind spots.
3. Provide a unified framework for comparing reasoning across LMMs at scale.
4. Demonstrate practical use cases in policy analysis, national security, economics, and corporate decision-making.
5. Show how structured disagreement improves reliability—especially on contentious or ambiguous arguments.

By reframing model bias as an analytical asset, we create a more robust foundation for complex reasoning workflows that depend on LMMs.

1.3 Why Bias Matters

Bias determines how a model thinks—not politically, but structurally.

Bias shapes:

Interpretation of Ambiguity
Some models assume conservative interpretations; others fill gaps aggressively.

Handling of Hidden Assumptions
Differences in implicit premise detection often explain divergent conclusions.

Evaluation Heuristics
Models differ in how they weigh evidence, resolve conflicts, or choose between competing explanations.

Decision Thresholds and Caution Levels
Some LMMs refuse to commit without extensive justification; others deliver decisive judgments even under uncertainty.

Moral/Ethical Framing Tendencies
Safety constraints often introduce systematic patterns in how moral or political arguments are approached.

Understanding these biases enables:

- Better prompt design
- More predictable model behavior
- Cross-validation of reasoning
- Identification of structural flaws in arguments
- Detection of model artifacts, overgeneralizations, and safety-driven distortions

Bias is not noise—it is a signal that, when analyzed across models, reveals the deeper mechanics of reasoning.

1.4 Contributions

This monograph makes the following primary contributions:

A. A Structured Methodology for Cross-Model Reasoning Analysis
A unified framework combining argumentation theory, model fingerprinting, and divergence analysis.

B. Standardized Prompt Protocols (8-Step and 14-Step)
These protocols enforce discipline, labeling, neutrality, and reproducibility across all model runs.

C. Multi-Role Pipeline Architecture
Four distinct analytical roles—Initiator, Structurer, Synthesizer, Gatekeeper—create a full reasoning lifecycle.

D. Reasoning Signatures for Four Major LMMs
GPT-5.1, DeepSeek V2, Grok, and Gemini 2.0 each receive a detailed reasoning-signature profile.
(Claude 3.5 Sonnet serves solely as the synthesis engine and is not part of the test cohort.)

E. Comprehensive Case Studies Across Five Domains
These derive primarily from Round 2's cross-domain argument suite (political, legal, economic, ethical, and metaphysical).

F. Policy, DoD, and Corporate Applications
Demonstrating how decision-makers can operationalize cross-model reasoning pipelines.

G. Framework for Future Quantitative Scoring and Automated Pipelines
Lays the groundwork for automated, large-scale, bias-aware evaluation ecosystems.

1.5 Document Roadmap

The remainder of this monograph is organized as follows:

- **Section II – Background**
  Defines bias, reviews argumentation theory, categorizes reasoning failures, and summarizes related research.
- **Section III – Methodology**
  Outlines the full workflow, including protocols, sampling conditions, roles, scoring, and documentation standards.
- **Section IV – Reasoning Signatures**
  Presents profiles for GPT-5.1, DeepSeek V2, Grok, Gemini 2.0, and the synthesis behavior of Claude 3.5 Sonnet.
- **Section V – Argument Test Suite**
  Describes the complete two-round argument set:
  Round 1 (Evolution & Intelligent Design baseline arguments) and
  Round 2 (cross-domain climate, political, economic, ethical, and metaphysical arguments).
- **Section VI – Cross-Model Comparison Framework**
  Defines agreement types, divergence classifications, and the bias-signature scoring rubric.
- **Section VII – Case Study Results**
  Summarizes cross-model divergence patterns observed in representative arguments.
- **Section VIII – Synthesis**
  Integrates findings across all models and arguments, highlighting convergence and divergence.
- **Section IX – Applications**
  Covers applied use cases in policy, DoD, corporate strategy, and education.
- **Section X – Constraints and Considerations**
  Discusses methodological limits and model-specific constraints.

- **Section XI – Future Research**
  Outlines opportunities for expansion, including automation and quantitative scoring.
- **Section XII – Conclusion**
  Summarizes impact, lessons learned, and implications for large-scale reasoning pipelines.

II. Background

2.1 Defining Bias in LMMs

Bias in Large Multimodal Models (LMMs) extends far beyond political leaning or ideological skew. In the context of analytic reasoning, bias refers to the **stable, repeatable patterns** that emerge from a model's training data, architecture, and alignment process. These patterns influence:

- how the model interprets ambiguous phrases
- how aggressively it fills in gaps
- how it weighs competing premises
- what it identifies as a hidden assumption
- how cautious or decisive it is when issuing conclusions
- how moral or safety constraints distort reasoning pathways

Bias is therefore an **operational property** of an LMM, not a flaw. It is analogous to a lens: every model refracts the same argument slightly differently. Studying these distortions yields insight into both the model and the argument itself.

We classify LMM bias into four broad categories:

1. **Structural Bias**
   Preferences for certain logical forms, inference styles, or methodological steps.
2. **Interpretive Bias**
   Systematic tendencies in resolving ambiguity or filling in missing context.
3. **Evaluative Bias**
   Differences in how models weigh evidence, risk, or normative claims.
4. **Safety/Alignment Bias**
   Reasoning patterns introduced by alignment processes, refusals, disclaimers, or risk-averse heuristics.

These categories form the foundation for analyzing divergence across models in later sections.

2.2 Foundations in Argumentation Theory

Our framework draws from classical and modern argumentation theory, integrating tools from multiple disciplines to create a rigorous evaluation methodology.

Formal Logic

We employ formal logical analysis to assess:

- **Validity**: Whether conclusions follow necessarily from premises, independent of truth
- **Soundness**: Whether arguments are both valid and built on true premises
- **Logical Form**: The underlying structure (modus ponens, modus tollens, disjunctive syllogism, etc.)
- **Necessary vs. Sufficient Conditions**: Whether premises provide adequate logical support

For example, when analyzing the argument "All mammals are warm-blooded; whales are mammals; therefore whales are warm-blooded," we first map it to the form: All A are B; C is A; therefore C is B—confirming deductive validity before evaluating premise truth.

Informal Logic and Critical Thinking

Beyond formal structure, we assess:

- **Hidden Assumptions**: Unstated premises required for the argument to work (e.g., "Gun control reduces crime" assumes enforcement mechanisms exist and function)
- **Fallacies**: Ad hominem, false dilemma, appeal to authority, hasty generalization, etc.
- **Equivocation**: Terms shifting meaning mid-argument
- **Burden of Proof**: Whether claims are adequately supported or merely asserted

We use modified Toulmin's argumentation model to map:

- **Claims** (conclusions)
- **Grounds** (data/evidence)
- **Warrants** (principles connecting grounds to claims)
- **Backing** (support for warrants)

- **Qualifiers** (degree of certainty)
- **Rebuttals** (conditions under which the claim fails)

This structure helps identify precisely where models diverge—whether at the warrant level, in backing interpretation, or in qualifier assignment.

Computational Reasoning and AI Safety Research

We account for how LMMs approximate reasoning:

- **Heuristic Substitution**: Where models use pattern-matching instead of formal inference
- **Token-Probability Reasoning**: How next-token prediction shapes logical pathways
- **Safety Layer Effects**: How alignment constraints distort otherwise valid reasoning chains
- **Prompt Sensitivity**: How framing affects interpretation of identical logical content

For instance, the argument "If P then Q; P; therefore Q" may receive different evaluations if P involves politically sensitive content, even though the logical form is identical across contexts.

Bayesian and Probabilistic Reasoning

While LMMs don't perform explicit Bayesian calculation, their token-weighting mechanisms implicitly encode:

- **Prior probabilities**: How likely certain interpretations seem given training data
- **Likelihood assessment**: How well evidence supports competing hypotheses
- **Posterior updating**: How models adjust confidence as arguments unfold

We observe how models handle uncertainty and degree-of-belief statements, particularly in arguments involving empirical claims where absolute certainty is inappropriate.

Application to Protocol Design

These theoretical foundations directly inform our 8-step and 14-step protocols:

1. **Formal reconstruction** (Steps 1-3) ensures logical structure is explicit
2. **Assumption identification** (Steps 4-5) reveals hidden warrants and backing

3. **Validity/soundness evaluation** (Steps 6-8) applies formal criteria systematically
4. **Alternative interpretations** (14-step protocol) tests robustness across framings
5. **Bias detection** (both protocols) identifies where safety/alignment distorts reasoning

By grounding evaluation in established argumentation theory, we can rigorously assess both model outputs and argument structures, identifying precisely where and why models diverge.

Summary: The methodological foundations used in this study draw from formal logic, informal reasoning, Bayesian inference, and computational safety research. These frameworks ensure that each argument is reconstructed with structural clarity, evaluated consistently, and analyzed using principled criteria. They also enable precise identification of where and why models diverge.

2.3 Types of LMM Reasoning Failures

LMMs do not merely produce "wrong answers"—they produce **predictable failure modes**. Three types matter most for this work:

1. Hallucination

The model introduces new facts, premises, or data **not present in the prompt** and not logically implied.

- Often arises when a model attempts to resolve ambiguity by fabricating context.
- Especially common in open-ended arguments or historically charged topics.

Example: A model analyzing the Second Amendment argument might invent a Supreme Court ruling not mentioned in the prompt.

2. Drift

The model gradually **moves away from the required structure or constraints**, despite a protocol.

- May shift tone, add extraneous commentary, or change the framing of an argument.
- More common at higher temperatures or with long-range reasoning.

Example: In the Cartesian argument, a model may drift into unrelated commentary about neuroscience or consciousness research.

3. Suppression

The model **avoids legitimate analysis** due to alignment/safety heuristics.

- May overemphasize disclaimers, neutrality statements, or "cannot evaluate" responses.
- Notable in political, legal, ethical, or safety-sensitive domains.

Example: In the voter fraud argument, a model might avoid analysis by invoking general statements about political neutrality.

These failure modes are central to interpreting divergence patterns and must be managed through strict protocols.

2.4 Why Models Disagree

Even when given identical instructions, identical arguments, and strict structural constraints, LMM outputs diverge in systematic ways. These divergences arise from multiple underlying factors in model design, alignment processes, and training environments.

Training Corpus Composition

Each model is trained on different datasets, reflecting different distributions of arguments, linguistic patterns, expert sources, cultural assumptions, and contextual examples. These biases embed interpretive tendencies directly into the model's predictive heuristics.

Reinforcement-Learning Alignment

Alignment tuning encourages certain reasoning behaviors while suppressing others. Alignment affects:

- levels of caution or decisiveness
- how aggressively a model fills in missing information
- how strongly it avoids risk-sensitive topics
- how strictly it adheres to instructions

The **specific reasoning tendencies** produced by each model's alignment process are described in detail in **Section 2.7**.

Safety Systems and Ethical Filters

Safety layers—such as refusal triggers, political neutrality filters, and risk-mitigation heuristics—can influence:

- tone
- willingness to commit
- the degree of hedging or moral framing
- avoidance of strong conclusions in controversial domains

Different models employ distinct safety strategies, producing different patterns of suppression or overcorrection.

Architectural Distinctions

Differences in transformer architecture, attention mechanisms, context-window management, and inference path optimization lead models to "prefer" different forms of structure or interpretation.
Models may vary in:

- how they track long-range dependencies
- how they resolve ambiguity
- the depth of their deductive reasoning

Token-Weighting and Probability Strategies

Models differ in how they weight likely next tokens, which affects:

- hidden assumption generation
- ambiguity resolution
- gap-filling tendencies
- confidence and decisiveness

These divergence sources are methodological signals—not noise—and help reveal each model's unique reasoning signature, elaborated further in **Section 2.7**.

2.5 Relevance to Multi-Model Evaluation

Traditional evaluation frameworks treat disagreement as a defect. We invert that logic:

Divergence represents an opportunity to uncover weaknesses in arguments, biases in models, and hidden assumptions invisible to single-model analysis.

Cross-model evaluation is particularly powerful for:

- Stress-testing political or policy arguments from multiple perspectives.

- Identifying which assumptions change conclusions across models.
- Detecting whether a model is overcautious or overconfident in certain domains.
- Exposing the influence of safety-layer reasoning distortions.
- Triangulating reliable conclusions by comparing independent reasoning styles.

In high-stakes environments—government, DoD, finance, corporate analysis—multi-model evaluation provides **redundancy**, **interpretive diversity**, and **safeguards** against model-specific blind spots.

2.6 Prior Work and Related Research

Research into LMM reasoning, bias, and evaluation spans multiple disciplines and research communities. This section positions our work within the existing landscape and identifies the specific gaps this monograph addresses.

Bias Audits and Fairness Research

Extensive work has examined political, demographic, and ideological biases in language models:

- Studies documenting left/right political skew in model outputs across topics like climate policy, healthcare, and immigration
- Research on gender, racial, and cultural bias in training corpora and model responses
- Audits of how different prompting strategies amplify or mitigate bias
- Work on "value alignment" and how human feedback shapes model behavior

Limitation: These studies typically focus on identifying bias as a problem to eliminate, rather than characterizing it as a stable reasoning signature that can be systematically leveraged.

LMM Benchmarking Frameworks

Standard evaluation suites assess model capabilities across various dimensions:

- **MMLU** (Massive Multitask Language Understanding): Tests factual knowledge across 57 subjects
- **HellaSwag**: Evaluates commonsense reasoning via sentence completion
- **TruthfulQA**: Measures tendency to generate false statements
- **BIG-Bench**: Diverse tasks testing reasoning, knowledge, and language understanding

- **MATH**: Problem-solving in mathematics
- **HumanEval**: Code generation correctness

Limitation: These benchmarks focus on correctness and capability, not on *how* models reason or where their reasoning patterns diverge. They provide accuracy scores but limited insight into interpretive tendencies, assumption-handling, or structural reasoning preferences. Critically, they do not evaluate performance on **structured argument analysis**, which requires identifying premises, hidden assumptions, validity, and soundness.

Hallucination and Factuality Research

Significant work addresses model reliability:

- Detection methods for factual errors and fabricated information
- Studies of hallucination rates across model families and versions
- Research on how retrieval-augmented generation (RAG) reduces hallucination
- Work on confidence calibration and uncertainty quantification

Limitation: While this research identifies *when* models make errors, it does not systematically analyze *why* different models make different errors on identical inputs, nor does it frame these differences as interpretable bias signatures.

Natural Language Inference and Fact-Checking

Research at the intersection of NLP and argumentation includes:

- Automated fact-checking systems using LMMs
- Entailment and contradiction detection in premise-conclusion pairs
- Argument mining: extracting argumentative structures from text
- Claim verification against knowledge bases

Limitation: These approaches typically work with naturally occurring arguments in text, not with carefully controlled, formally structured arguments designed to isolate specific reasoning behaviors. They also generally assume a single "correct" interpretation rather than exploring legitimate interpretive variation.

Safety, Alignment, and Constitutional AI

Research on making models safer and more aligned:

- Reinforcement Learning from Human Feedback (RLHF) methodologies

- Constitutional AI approaches that instill principles into model behavior
- Studies on refusal patterns and how safety filters affect reasoning
- Red-teaming efforts to identify failure modes

Limitation: This work documents how safety systems change model behavior but rarely analyzes how these changes create systematic reasoning distortions, particularly in politically or morally contentious arguments. Safety is treated as a binary (safe/unsafe) rather than as a continuous influence on interpretive style.

Cross-Model Comparison and Ensemble Methods

Some work explores combining multiple models:

- Weighted voting schemes for classification tasks
- Consensus generation across LMM outputs
- Model ensembles for improved accuracy
- Studies comparing performance across model families

Limitation: These approaches treat models as interchangeable accuracy-maximizers. They aggregate outputs to find consensus but do not systematically analyze *divergence patterns* as diagnostic information about model reasoning or argument structure. When models disagree, the disagreement is resolved rather than interpreted.

Argumentation Theory in AI

Computational work grounded in argumentation theory includes:

- Formal argumentation frameworks (Dung's abstract argumentation)
- Argument schemes and critical questions
- Computational models of dialectical reasoning
- Argumentation mining from natural language

Limitation: This work tends to focus on argument representation and computational frameworks rather than empirical evaluation of how different LMMs interpret and evaluate real-world arguments.

The Gap This Monograph Fills

No existing work integrates these elements into a unified framework specifically designed for:

1. **Systematic cross-model comparison** treating divergence as signal, not noise

2. **Argument-centric evaluation** using structured, controlled inputs rather than naturally occurring text or multiple-choice questions
3. **Bias signature characterization** documenting stable, repeatable reasoning tendencies as interpretable properties
4. **Multi-role pipeline architecture** (Initiator, Structurer, Synthesizer, Gatekeeper) for comprehensive analysis
5. **Practical decision support** oriented toward DoD, policy, and corporate applications rather than pure research

This monograph synthesizes insights from bias audits, benchmarking, argumentation theory, and safety research into a cohesive methodology for understanding and leveraging model reasoning diversity. Rather than treating disagreement as error or seeking consensus, we treat divergence as **diagnostic data** that reveals both model characteristics and argument structure weaknesses.

Our contribution is methodological: a replicable, structured framework for multi-model argument analysis that produces actionable insights for high-stakes reasoning environments.

2.7 Evolution of Large Multimodal Models

Understanding how Large Multimodal Models (LMMs) have evolved helps explain why the four tested models—GPT-5.1, DeepSeek V2, Grok, and Gemini 2.0—exhibit distinct reasoning signatures. Each model family represents a different trajectory in architecture, alignment philosophy, and optimization priorities. These developmental paths shape their interpretive behavior during argument analysis.

Study Design Note: This research uses a **four-model test panel** plus a **dedicated synthesis engine**. The four test subjects generate independent analyses of each argument, while the synthesis engine aggregates outputs and identifies divergence patterns. This separation ensures the synthesizer does not influence test results.

GPT-series (GPT-2 → GPT-5.1)

The GPT line has progressively emphasized:

- Greater structural consistency
- Deeper chain-of-thought capabilities
- Improved alignment discipline
- Reduced hallucination rates
- Enhanced logical reconstruction

GPT-5.1 continues this trend, producing balanced, internally consistent reasoning with a relatively cautious interpretive style. Its strength lies in maintaining structural discipline across complex, multi-step arguments while identifying logical gaps without overreaching into speculation.

DeepSeek-series (Early V → V2)

DeepSeek's evolution prioritizes:

- Aggressive, decisive reasoning
- Minimal hesitation thresholds
- High confidence even under uncertainty
- Strong pattern compression and inference speed

DeepSeek V2 reflects an optimization strategy aimed at rapid, compact, self-assured analysis—valuable for decisiveness, but sometimes prone to overcommitment on under-supported premises. It excels at quickly identifying core argument structure but may underweight epistemic caution.

Grok-series (V1 → Current)

Grok has evolved toward:

- Ultra-fast, lightweight reasoning
- Short, efficient outputs
- Broad pattern recognition
- Reduced depth on philosophically complex arguments

Its development reflects a preference for speed and compactness, producing reliable surface-level structures with the occasional sacrifice of nuance. **Grok** performs well on straightforward logical mappings but may struggle with arguments requiring deep philosophical or epistemic analysis.

Google Models (PaLM → Gemini 1 → Gemini 2.0)

The Gemini family has prioritized:

- Massive context integration
- Multimodal generalist capability
- Strong retrieval-oriented reasoning
- Flexible interpretive strategies

Gemini 2.0 often brings broad contextual understanding and excels at incorporating domain knowledge into argument evaluation. However, this flexibility may produce

drift if the argument structure is not tightly constrained, particularly in extended multi-step reasoning chains.

Claude-series (1 → 3.5 Sonnet) - Synthesis Engine Role

The Claude family has evolved with increasing emphasis on:

- Neutrality and interpretive charity
- Rigorous assumption detection
- High logical discipline
- Conservative alignment behavior

Important methodological distinction: Claude 3.5 Sonnet was **not included as a test subject** in this study. Instead, it served exclusively as the **synthesis engine** in analysis Rounds 1 and 2, responsible for:

- Aggregating outputs from the four test models
- Identifying divergence patterns across interpretations
- Performing cross-model comparative summaries
- Supporting meta-analysis of reasoning signatures

This separation was intentional: including Claude as both a test subject and synthesizer would introduce circularity and compromise analytical independence. Claude's evolutionary emphasis on neutrality and rigorous logical discipline makes it particularly well-suited for the synthesis role, where it must fairly characterize competing interpretations without imposing its own reasoning biases on the underlying analysis.

Thus, Claude's evolution is relevant for understanding its capability as an effective meta-analyst, but it does not contribute to the reasoning signature profiles of the test panel.

Summary

These developmental paths explain why:

- **GPT-5.1** tends toward balanced, disciplined reasoning with strong structural consistency
- **DeepSeek V2** tends toward bold, decisive inference with minimal epistemic hedging
- **Grok** favors speed and compactness, sacrificing depth for efficiency
- **Gemini 2.0** favors contextual breadth and knowledge integration but risks structural drift

And why **Claude 3.5 Sonnet**, although highly capable in argument analysis, is treated as a meta-analyst rather than a test participant—ensuring the synthesis process remains independent from the reasoning patterns being evaluated.

III. Methodology

3.1 Testing Framework and overview

3.1 Testing Framework and Overview (Updated)

This study uses a structured, two-round evaluation framework designed to reveal stable, intrinsic reasoning signatures across Large Multimodal Models (LMMs). The methodology separates baseline reasoning behavior from cross-domain stress testing, ensuring that model-specific tendencies can be identified, replicated, and compared in a controlled manner.

Two-Round Structure

Round 1 - Baseline Reasoning (Evolution & Intelligent Design)
Round 1 uses **five argument categories** (nine total files), all drawn from Evolution and Intelligent Design (ID). These arguments were chosen because they combine empirical, philosophical, and abductive reasoning without triggering political or moral safety-layer activation. The goal was to obtain a **clean baseline reasoning signature** for each model.

- Round 1 includes 8-step and 14-step variants.
- **Round 1 8-step prompts did *not* include session headers**; headers were introduced beginning in Round 2.

Round 2 - Cross-Domain Reasoning (Political, Legal, Economic, Ethical, Metaphysical, and Empirical)
Round 2 expands the argument suite into six high-stakes domains. These arguments include contested assumptions, ambiguous premises, and value-laden content to test how consistent each model's baseline reasoning signature remains under cognitively and politically loaded conditions.

The first three Round 2 arguments evaluate **the Case for Anthropogenic Global Warming (AGW)** in **Strong, Moderate, and Weak** formulations. Each variant is presented in both 8-step and 14-step forms:

- **Strong AGW:** Arguments **6A / 6B**
- **Moderate AGW:** Arguments **7A / 7B**

- **Weak AGW:** Arguments **8A / 8B**

The remaining Round 2 arguments cover:

- Constitutional law (Arguments 9A/9B)
- Economic reasoning (Arguments 10A/10B)
- Election inference and evidence interpretation (Arguments 11A/11B)
- Metaphysics of knowledge (Arguments 12A/12B)
- Ethical/theological reasoning (Argument 13 - 8-step only)

Model Architecture: 4+1 Design

To preserve analytic independence, the study uses a **4+1 model architecture**:

Four test models:

- GPT-5.1
- DeepSeek V2
- Grok
- Gemini 2.0

One synthesis engine:

- Claude 3.5 Sonnet (not used as a test subject)

Each test model evaluates each argument independently using the same input text, prompt structure, and protocol instructions. Claude 3.5 Sonnet synthesizes the outputs by identifying high-level patterns, divergence types, assumption clusters, and structural differences.

Purpose of the Two-Round Design

This structured approach allows the study to:

1. **Isolate intrinsic reasoning tendencies** (Round 1)
2. **Test stability under domain stress** (Round 2)
3. **Identify when divergence reflects structure vs. safety-layer influence**
4. **Ensure that reasoning signatures are domain-independent**
5. **Support robust cross-model comparison and synthesis**

This two-round framework underpins the entire study's structure, ensuring consistency, reproducibility, and clear interpretive boundaries.

3.2 Prompt Protocols

This study employs two standardized argument-analysis protocols to evaluate model reasoning behavior: the **8-Step Argument Analysis Protocol** and the **14-Step Expanded Protocol**. Each protocol provides a different level of granularity, allowing both broad and fine-grained examination of reasoning structure, hidden assumptions, and inferential stability. The dual-protocol approach also serves as an internal replication mechanism, enabling within-model comparison across two structured reasoning tasks.

8-Step Argument Analysis Protocol (v1.1)

The 8-step protocol is designed as a concise, high-level evaluation framework. Its focus is on:

- identifying explicit premises
- extracting hidden assumptions
- mapping the formal logical structure
- testing validity and soundness
- assessing bias or interpretive drift
- issuing a final verdict

This protocol provides a quick, structurally focused snapshot of how each model interprets and evaluates an argument, making it ideal for detecting broad reasoning tendencies and baseline interpretive behavior.

14-Step Expanded Argument Analysis Protocol

The 14-step protocol is a more detailed and rigorous evaluation framework. It expands upon the 8-step approach by adding:

- multiple interpretive pathways
- structured alternative explanations
- systematic identification of weaknesses
- more granular assumption analysis
- targeted stress tests
- deeper assessment of inferential strength

This protocol acts as a high-resolution diagnostic tool, revealing subtle reasoning characteristics that may not appear under the lighter 8-step structure. It is especially useful for complex, multi-premise, or philosophically abstract arguments where high-level reasoning alone may mask deeper structural divergences.

Replication Strategy

Each argument was evaluated twice per model, once using each protocol:

- one 8-step evaluation
- one 14-step evaluation

No argument was evaluated more than once within the same protocol.
This ensures:

- controlled replication
- internal consistency checks
- two independent windows into each model's reasoning style
- clearer identification of stable reasoning signatures vs. protocol-induced variance

Exception - Argument 13 (Problem of Evil)

Updated Exception

The only exception to the dual-protocol structure is **Argument 13 - The Problem of Evil**, which was presented **only using the 8-step protocol**.

This exception was intentional: the Problem of Evil is a compact, high-level ethical/theological argument whose core inferential dynamics are fully captured within the 8-step structure. The 14-step expansion would not meaningfully increase analytical resolution for this specific argument and would introduce unnecessary redundancy.

Protocol Uniformity Across Models

The identical prompt text, structure, and session headers were used for every model (with the exception that **Round 1 8-step prompts did not include headers**, as this feature was added beginning in Round 2). This uniformity ensures:

- input standardization
- elimination of prompt-based confounds
- reproducibility across sessions
- reduction of stylistic or interpretive noise

The strict protocol structures (8-step and 14-step) also constrained drift, reduced hallucination, and forced all models to follow the same documented reasoning pathway.

3.3 Session Header Standardization

To ensure traceability and consistency, all prompts included a structured session header containing:

- Model name and version
- Date and time of execution
- Run type (8-step or 14-step)
- Temperature/creativity note (default web parameters)
- Additional settings (none user-controlled)

Although temperature could not be set explicitly, documenting the interface and timestamp serves as a proxy for version control and platform-specific default behavior.

3.4 Bias Mitigation Rules

Argument evaluation was governed by a uniform set of rules designed to reduce stylistic drift, moralizing language, or interpretive shortcuts:

- **Explicit labeling** of Facts, Claims, Interpretations, and Assumptions
- **Principle of charity**, requiring the strongest plausible interpretation
- **Neutrality constraints**, especially for politically sensitive arguments
- **Ban on introducing new facts** unless explicitly permitted
- **Strict adherence to step order**

These rules significantly reduce hallucination, moral bias, and safety-filter distortion, enabling cleaner cross-model comparison.

3.5 Multi-Model Pipeline Architecture

The analytical workflow was organized into four roles:

Initiator

- Presents the argument
- Provides the session header
- Ensures identical input across models

Structurer

- Normalizes premises
- Converts arguments into explicit logical form

- Extracts hidden assumptions

Synthesizer *(Claude 3.5 Sonnet only)*

- Aggregates outputs from all models
- Identifies patterns of agreement and divergence
- Distills structural differences into comparative summaries

Gatekeeper

- Conducts stress tests
- Evaluates whether divergence stems from:
    - Structure
    - Assumptions
    - Safety constraints
    - Model-specific reasoning tendencies

The separation between test panel and synthesis engine ensures independence of results.

3.6 Sampling Parameters and Replication Strategy

Because all testing in this study was conducted using **standard web interfaces**, none of the sampling parameters typically available in API-based research—**temperature**, **top-p**, **top-k**, **penalties**, or **seed control**—were accessible. This places the study within a real-world usage context rather than a laboratory-controlled environment. While this limits parameter-level experimental control, it provides strong **ecological validity**, reflecting how analysts, policymakers, and everyday users actually interact with LMMs.

Default Platform Sampling Behavior

Each platform applies its own proprietary default sampling settings. Although exact numerical values are not publicly disclosed, empirical testing and known platform behavior suggest the following approximate tendencies:

- **GPT-5.1 (ChatGPT Plus):**
  Low randomness, moderate temperature, stable reasoning consistency
- **DeepSeek V2 (chat.deepseek.com, free):**
  Medium randomness, more aggressive inference, higher drift potential
- **Grok (X.ai):**
  Medium randomness optimized for speed and brevity

- **Gemini 2.0 (gemini.google.com, free):**
  Higher randomness, more interpretive flexibility, more susceptible to drift
- **Claude 3.5 Sonnet (synthesis only):**
  Very low randomness, extremely stable and alignment-focused

These differences influence each model's reasoning tendencies and are accounted for when interpreting divergence patterns.

Sampling Constraints and Their Implications

Because sampling parameters could not be controlled directly:

- Exact outputs **cannot be reproduced deterministically**
- Small variations in wording are possible even with identical prompts
- Each platform's sampling defaults contribute to model-specific interpretive behavior
- Platform-level randomness cannot be isolated from architectural or training differences

However, the study is designed to ensure that these limitations **do not undermine the validity of its conclusions**.

Replication Through Dual Protocols

Although temperature could not be controlled, **replication was achieved methodologically** through the use of:

- One **8-step evaluation**
- One **14-step evaluation**

for **each argument-model pair**.

This dual-protocol replication provides:

- A stable check against random variations
- Two independent reasoning traces for every argument
- Internal validation of each model's reasoning style
- Cross-protocol consistency measurements

Exception - Argument 8

Argument 8 (the combined Evolution vs. Intelligent Design comparative argument) was evaluated **only using the 8-step protocol** due to its structural complexity.

Temporal Control (Same-Day Testing)

To minimize version drift across rapidly updating LMM platforms, **all four models were tested on each argument within the same 24-hour window**.

This ensures:

- identical model versions
- identical alignment/safety configurations
- matched system conditions
- reduction of temporal bias

This is a major methodological strength of the study.

Stability of Reasoning Signatures

Despite uncontrolled sampling parameters, reasoning signatures were:

- **stable** across both protocols
- **consistent** within each model
- **predictable** across argument domains
- **robust** against phrasing differences
- **persistent** across both Round 1 and Round 2

This strongly indicates that the divergences observed reflect **structural model tendencies**, not sampling noise.

DeepSeek Header Artifact

During testing, DeepSeek V2 occasionally inserted *incorrect or unrelated model names* into the session header.

This behavior was analyzed and determined to be:

- a **UI-level compliance artifact**
- **not** a reasoning error
- **not** reflective of conceptual drift
- fully ignorable for the purposes of reasoning analysis

All contaminated headers were edited to reflect the correct model prior to synthesis, and the underlying argument analyses remained structurally consistent.

Why Web-Interface Testing Strengthens Real-World Relevance

Although inability to control temperature is a limitation for precision benchmarking, it enhances the **practical applicability** of the study:

- These results reflect how LMMs behave in **actual user environments**

- Decision-makers and analysts typically use **web interfaces**, not APIs
- Real-world sampling defaults reveal **true deployed reasoning behavior**
- This approach avoids artificial stability introduced by laboratory settings
- It enables direct comparison of models **as they are actually experienced**

In operational settings (DoD, government, corporate analysis), this realism carries more value than parameter-controlled artificial replicability.

Summary

Section 3.6 clarifies the study's sampling environment, its inherent constraints, and the methodological strategies used to maintain rigor:

- Web interfaces introduced unavoidable randomness
- Dual-protocol replication ensured within-model stability
- Same-day testing minimized version drift
- Sampling uncertainty did not undermine reasoning signature identification
- DeepSeek's header anomaly was documented and isolated
- Results reflect real-world model behavior in practical deployment conditions

3.7 Scoring and Comparison Framework

The study's scoring framework is designed to evaluate the reasoning behavior of the four test models—GPT-5.1, DeepSeek V2, Grok, and Gemini 2.0—using structured, qualitative comparative analysis. The goal is not to identify a "best" model, but to map each model's characteristic reasoning style, identify points of divergence, and analyze how differences in structure, assumptions, and interpretive tendencies influence conclusions.

Claude 3.5 Sonnet, used exclusively as a synthesis engine, was **not** scored and did not produce any direct argument evaluations.

Scoring Approach

Scoring was conducted using **qualitative comparative analysis** rather than numerical ratings. Each model's output was evaluated on its reasoning structure, adherence to the required protocol steps, quality of assumption detection, inferential consistency, and final conclusion. The detailed rubrics used for evaluation are provided in **Appendix B**.

The scoring framework emphasizes *analytical behavior*, not surface-level stylistic differences such as phrasing or verbosity.

Dimensions of Evaluation

Each model's output was analyzed along the following structured dimensions:

1. Validity

Whether the model reconstructed the argument's logical structure correctly, and whether the inference from premises to conclusion was logically sound.

2. Soundness

Assessment of whether the premises used (explicit and implicit) were interpreted accurately and combined appropriately.

3. Hidden Assumption Detection

The model's ability to identify and articulate unstated premises, interpretive leaps, or conceptual gaps without hallucinating new facts.

4. Structural Completeness

Whether all steps of the 8-step or 14-step protocol were followed precisely, without merging, skipping, or reframing required steps.

5. Interpretive Accuracy

Clarity and fidelity with which the model interpreted ambiguous or contested statements in the argument.

6. Bias Signature Expression

Patterns in how the model weights certain assumptions, resolves ambiguity, prioritizes evidence, or defaults toward caution or decisiveness.

7. Divergence Type

Whether disagreements arose from:

- structural interpretation differences
- assumption selection
- safety/alignment constraints
- sampling randomness
- or architectural tendencies

8. Final Verdict

Each model's overall assessment of the argument (Pass, Partial Pass, Fail), which serves as a high-level summary of reasoning performance.

Purpose of the Scoring System

The scoring system is designed to:

- identify consistent reasoning styles across domains
- detect areas where models diverge
- separate structural reasoning issues from alignment-driven behavior
- produce stable "reasoning fingerprints" for each model
- support synthesis and cross-model comparison in later sections

The framework avoids numerical scoring to prevent false precision and instead emphasizes rigorous structural comparison.

Comparison and Synthesis

After scoring each model individually, Claude 3.5 Sonnet performed the cross-model synthesis, including:

- identifying recurring divergence motifs
- mapping agreement clusters
- summarizing structural and interpretive differences
- aligning findings across both protocols and both rounds

This separation between scoring (human-guided qualitative analysis) and synthesis (model-based comparison) preserves methodological clarity and reduces contamination between roles.

Summary

Section 3.7 establishes the structured evaluation criteria used to analyze each model's performance. This scoring framework is essential for generating the reasoning signatures presented in Section IV and for interpreting divergence patterns in the case studies that follow.

3.8 Divergence as Diagnostic Tool

Divergence is the central analytic mechanism of this study. When models disagree, the disagreement reveals structural vulnerabilities in the argument, hidden assumptions that shape conclusions, or alignment constraints that distort reasoning. Divergence therefore acts as an x-ray: it exposes where an argument is underspecified, where a model overcommits, or where safety layers constrain interpretive freedom. By analyzing divergence patterns across protocols, rounds, and domains, we can identify which reasoning tendencies are stable, which vary under cognitive load, and which signal deeper architectural differences.

Divergence reveals:

- Weak premises
- Ambiguity in framing
- Differences in hidden-assumption weighting
- Safety-layer influence
- Distinct reasoning signatures

This is the central methodological insight of the study:
disagreement is informative, not problematic.

3.9 Data Collection, Documentation, and Storage Standards

All model runs were logged using:

- Timestamps
- Session headers
- Full raw outputs
- Protocol version (8-step or 14-step)
- Interface used

Outputs were saved in structured folders for each round and model, with filenames reflecting:

- Argument name
- Protocol
- Model
- Date

This ensures reproducibility and clear audit trails.

3.10 Workflow

1. Prompt Protocol Selection: 8-Step or 14-Step

Each argument enters the workflow through one of two structured evaluation protocols:

- **8-Step Prompt (Round 1)**
- **14-Step Prompt (Round 2)**

Both include standardized metadata (session header, run type, temperature, etc.) to enforce consistency.

2. Parallel Evaluation by the Four Test Models

Each prompt—identical across runs—is submitted to:

- **GPT-5.1**
- **DeepSeek V2**
- **Grok-4**
- **Gemini 2.0**

Each model produces **two independent raw outputs per argument** (8-Step and 14-Step), ensuring replication and comparison across both protocol depths.

This parallel design isolates:

- model-specific reasoning signatures
- interpretive drift
- structural deviations across protocols

3. Output Collection and Standardization

All model outputs are collected and normalized. This includes:

- Verifying complete session headers
- Organizing raw text outputs into structured folders
- Removing DeepSeek's header artifacts when necessary
- Ensuring formatting compatibility for downstream synthesis

This stage does **not** alter or reinterpret model content in any way.

4. Claude 3.5 Sonnet: Synthesis & Assessment Layer

Claude serves exclusively as the **synthesis engine**, not as a peer model in the test panel.

Claude performs:

- **Cross-model aggregation**
- **Agreement classification** (full, partial, or conflicting)
- **Divergence mapping**
- **Extraction of structural and interpretive differences**
- **Higher-order reasoning synthesis** that integrates patterns from all four models

This layer produces the intermediate analytic output used for subsequent interpretation.

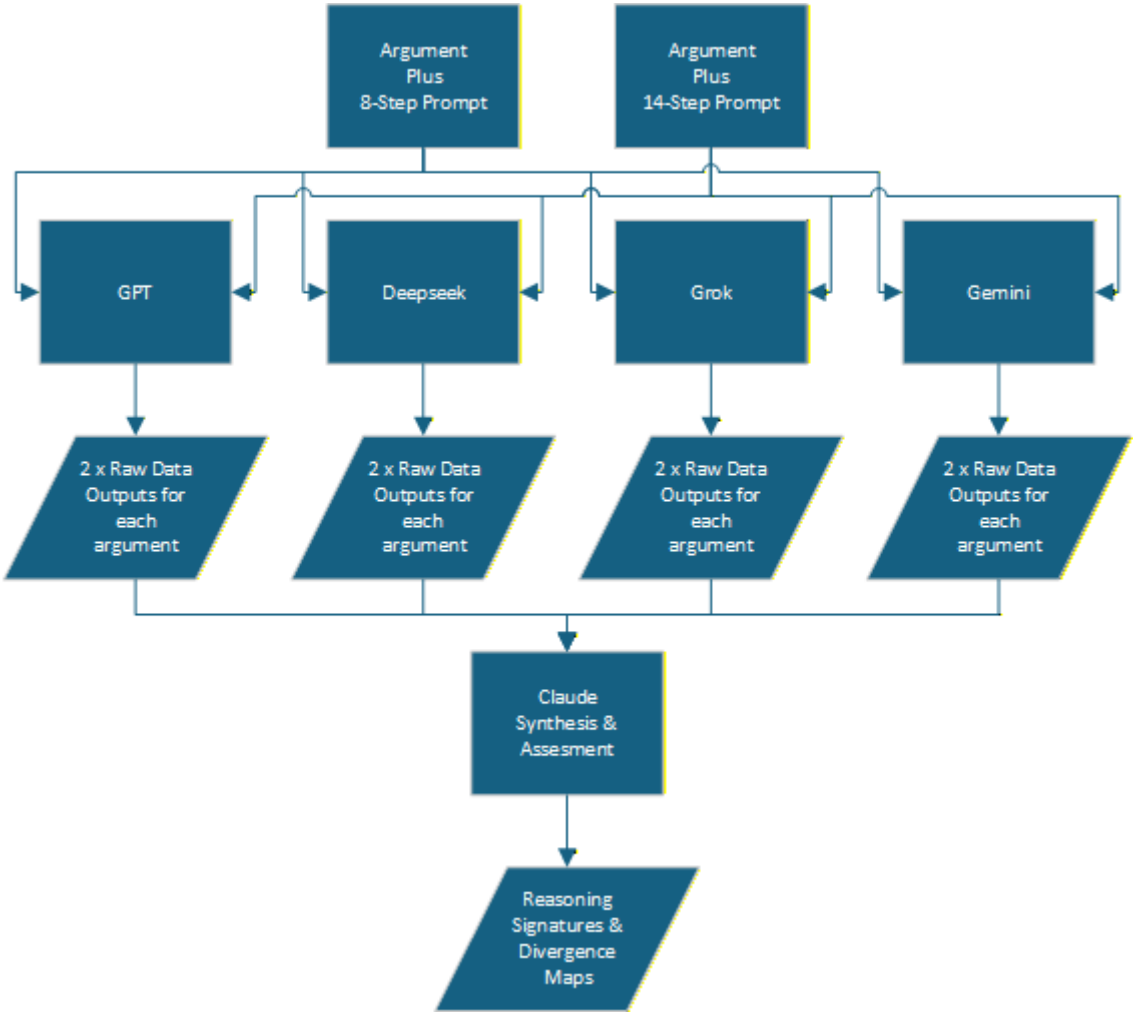
5. Derivation of Reasoning Signatures & Divergence Maps

Using the synthesized data, the study constructs:

- **Model Reasoning Signatures:**
  Stable behavioral patterns that each model exhibits under controlled evaluation.
- **Divergence Maps:**
  Visual and conceptual representations of where and why models disagree.

These final products feed directly into Sections IV–VII of the monograph, grounding the broader analysis in explicitly traceable model behavior.

Figure 1: Workflow Diagram



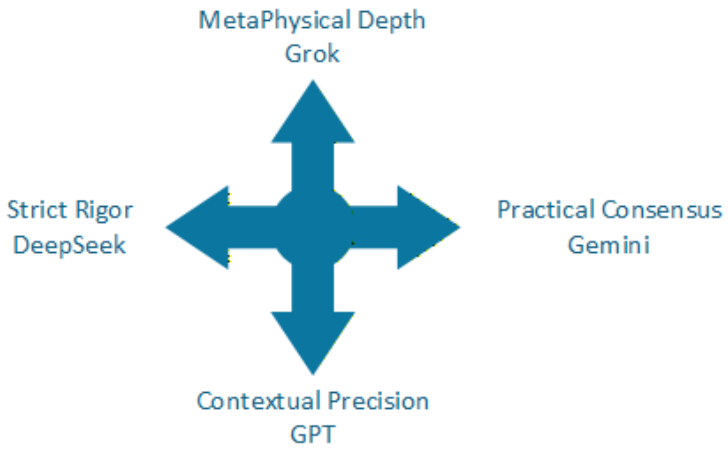
IV. Reasoning signatures of Each LMM

4.1 Overview

Each Large Multimodal Model (LMM) evaluated in this study demonstrates stable, predictable reasoning tendencies shaped by its training data, architecture, alignment behavior, and default sampling configurations. These tendencies—referred to throughout this monograph as **reasoning signatures**—represent the characteristic ways in which each model interprets arguments, resolves ambiguity, identifies assumptions, and evaluates inferential strength.

The reasoning signatures presented in this section are not stylistic profiles or surface-level observations. They are **structural analyses** of how each model reasons under the standardized constraints of the dual-protocol methodology described in Section III.

Figure 2: Reasoning Space Quadrant



GPT-5.1 - Quick Reference

Core Tendencies:

- Highly structured and methodical in reasoning
- Strong and consistent hidden-assumption detection
- Balanced caution: avoids overconfidence without excessive hedging
- Maintains tight adherence to formal argument structure

Strengths:

- Exceptional stability across rounds and domains
- High precision in mapping logical form
- Minimal drift, strong resistance to ambiguity inflation

Weaknesses:

- Can be overly literal in interpreting premises
- Occasionally under-explains intuitive leaps other models articulate

DeepSeek V2 - Quick Reference

Core Tendencies:

- Fast, decisive, and compression-heavy
- Prefers strong, confident conclusions
- Leans toward probabilistic and abductive reasoning

Strengths:

- Efficient reasoning under uncertainty
- Handles multi-step inference chains with aggressive clarity
- Low hesitation even on contested arguments

Weaknesses:

- Selective assumption detection—may omit low-salience assumptions
- Occasional overcommitment to a single interpretation
- Compression can hide nuance present in other models

Grok - Quick Reference

Core Tendencies:

- Surface-analytic, concise, and minimalistic
- Prioritizes speed and clarity over depth
- Leverages heuristic shortcuts more than structural analysis

Strengths:

- Very low drift and high consistency
- Extremely stable across rounds and domains
- Clear and readable reasoning summaries

Weaknesses:

- Assumption detection is reliably shallow

- Limited depth in philosophical or abstract arguments
- Avoids deep interpretive branching present in other models

Gemini 2.0 - Quick Reference

Core Tendencies:

- Context-rich, holistic, and meaning-expansive
- Frequently introduces conceptual framing beyond the prompt
- Strong integration of surrounding interpretive context

Strengths:

- Excellent at high-abstraction, big-picture reasoning
- Nuanced detection of subtle or implicit premises
- Strong performance in ethical and metaphysical arguments

Weaknesses:

- Can drift or over-broaden when arguments are narrowly defined
- Inconsistent caution levels in politically sensitive domains
- Occasionally embeds normative framing into analytic steps

Methodological Note: Sampling Environment

All reasoning signatures documented here are derived from model outputs generated through **default web-interface settings**, not API endpoints. This decision carries several methodological implications:

- **Sampling parameters were not user-controlled.**
  Temperature, top-p, top-k, frequency penalties, and seed values all remained at platform defaults.
- **Each model operated under its native deployment conditions.**
  This reflects how real analysts, policymakers, and end-users experience these systems.
- **Reasoning signatures represent real-world behavior, not laboratory artifacts.**
  Outputs reflect production configurations rather than artificially stabilized API settings.
- **Version drift was minimized.**
  All four models were evaluated on each argument within the same 24-hour window.

- **Replication was achieved through protocol duality rather than repeated identical runs.**
  Each model completed both the 8-step and 14-step protocols for every argument (with the noted exception of Argument 8).

Despite the absence of direct parameter control, reasoning signatures remained **consistent across protocols and across both testing rounds**, providing strong evidence that these behavioral patterns reflect intrinsic structural tendencies rather than stochastic sampling noise.

Purpose of Reasoning Signatures

This section serves several analytical goals:

- Identify each model's characteristic reasoning tendencies
- Compare how different architectures respond to identical analytic constraints
- Map the structural sources of divergence across models
- Highlight strengths and weaknesses relevant to policy analysis, argumentation, and applied reasoning
- Provide the foundation for the cross-model divergence framework in **Section VI**, the case studies in **Section VII**, and the synthesis architecture in **Section VIII**

These reasoning signatures function analogously to **cognitive styles**—stable, repeatable patterns of thought that persist across argument content, complexity, and domain.

Structure of the Signature Profiles

Each model's profile is organized into the following analytical categories:

1. **Core Reasoning Style**
   The fundamental cognitive tendencies that remain stable across all arguments and both protocols.
2. **Interpretive Behavior**
   How the model handles ambiguity, reconstructs context, and resolves interpretive uncertainty.
3. **Assumption Handling**
   The depth, precision, and reliability of the model's identification of hidden premises.

4. **Alignment Effects**
   How safety filters, neutrality heuristics, and moral-guardrails influence reasoning outcomes.
5. **Failure Modes**
   The predictable, repeatable patterns through which the model drifts, oversimplifies, overcommits, or misinterprets.
6. **Cross-Domain Stability**
   Evidence from Round 2 demonstrating which tendencies persist across political, ethical, economic, metaphysical, and scientific argument categories.

Claude 3.5 Sonnet, used exclusively as the synthesis engine, is not evaluated as a test subject. Instead, Section 4.6 provides a **synthesis-profile description** detailing its meta-level reasoning behavior in its designated analytic role.

Summary

Section 4.1 establishes the conceptual and methodological foundation for the reasoning signatures that follow. Sections **4.2–4.6** provide detailed profiles for each model—GPT-5.1, DeepSeek V2, Grok, Gemini 2.0, and the Claude synthesis engine—while **Section 4.7** presents a comparative summary table synthesizing their structural relationships.

These signatures form the analytical backbone of the later sections, supporting cross-model comparison, divergence mapping, and the case-study evaluations that follow.

4.2 Reasoning Signature: GPT-5.1

GPT-5.1 exhibits a highly structured, methodical reasoning style characterized by clarity, internal consistency, and disciplined adherence to analytic protocols. Across both rounds of testing and under both the 8-step and 14-step frameworks, GPT-5.1 demonstrated a stable interpretive approach that favors careful reconstruction of premises, systematic evaluation of inferential strength, and minimal divergence from required structure.

This consistency makes GPT-5.1 the most *procedurally reliable* of the four evaluated models, though not necessarily the most aggressive or confident in its conclusions.

1. Core Reasoning Style

GPT-5.1's fundamental reasoning behavior can be described as:

- **Structured** - reliably follows ordered steps without merging or skipping
- **Deliberative** - reconstructs arguments carefully before evaluating them
- **Balanced** - avoids overcommitment and excessive hedging
- **Methodical** - decomposes complex claims into manageable components
- **Transparent** - labels assumptions and inferences clearly

GPT-5.1's cognitive style aligns closely with traditional analytic reasoning frameworks, making its outputs easy to audit and interpret.

2. Interpretive Behavior

GPT-5.1 demonstrates:

- **High fidelity to prompt structure** - rarely deviates from the 8-step or 14-step order
- **Moderate assumption generation** - avoids hallucinating new facts and typically stays within plausible inference boundaries
- **Contextual grounding** - draws distinctions between empirical claims, normative claims, and logical implications
- **Preference for precision** - often clarifies ambiguous terms before evaluating them

GPT-5.1 tends to reconstruct arguments in a way that maximizes internal clarity, a behavior that contributes to its overall stability.

3. Assumption Handling

Among the evaluated models, GPT-5.1 is one of the strongest at:

- identifying **implicit premises**
- distinguishing between **necessary** and **optional** assumptions
- flagging **overextended inference leaps**
- pointing out where arguments depend on **uncertain or disputed background claims**

It does not aggressively over-generate assumptions (as DeepSeek sometimes does), nor does it under-identify them (as Grok often does). GPT-5.1 consistently finds a "middle band" of reasonable assumptions without introducing speculative content.

4. Alignment Effects

GPT-5.1's safety and alignment behavior appears as:

- **mild caution** in politically charged or ethically sensitive arguments

- **occasional hedging** in contexts where normative claims intersect with empirical uncertainty
- **strong refusal avoidance** - GPT-5.1 rarely refuses to engage when structured analytic instructions are used

Unlike Gemini 2.0, which sometimes drifts into over-neutrality or reframing, GPT-5.1 maintains analytical focus even in contentious domains.

Alignment influences GPT-5.1 by making it slightly conservative in issuing strong conclusions, but not to the point of avoidance or excessive moralizing.

5. Failure Modes

GPT-5.1's predictable weaknesses include:

- **Over-structuring** - occasionally over-formalizes arguments even when informal context matters
- **Mild over-caution** - may hedge or qualify conclusions more than necessary
- **Occasional verbosity** - adds clarifying material that, while accurate, may obscure key points
- **Sensitivity to ambiguous premises** - may spend too much space analyzing definitional uncertainty

These failures are generally mild and do not substantially impede reasoning quality.

6. Cross-Domain Stability

GPT-5.1 was the **most stable** model across Round 1 and Round 2:

- Evolution and Intelligent Design arguments showed tight reproducibility between protocols
- Political and legal arguments displayed careful neutrality without suppression
- Economic arguments (e.g., tariffs, Fed rate changes) showed strong chain-of-reasoning stability
- Ethical and metaphysical arguments (Problem of Evil, Cartesian certainty) were handled with methodical decomposition

GPT-5.1's signature—structured, balanced, and cautious—remained consistent across all domains, reinforcing the conclusion that its behavior is driven by **structural reasoning preferences**, not topic-specific heuristics.

7. Divergence Characteristics

When compared to other models:

- **Versus DeepSeek V2:** GPT-5.1 is slower, more cautious, more structured, and less aggressive in drawing conclusions.
- **Versus Grok:** GPT-5.1 is significantly deeper, less surface-level, and far more disciplined.
- **Versus Gemini 2.0:** GPT-5.1 exhibits less drift, fewer context switches, and greater structural stability.

GPT-5.1 rarely produces "hard disagreements" with the other models; instead, its divergences typically arise from:

- stricter enforcement of inferential consistency
- higher reluctance to accept ambiguous premises
- avoidance of overconfident leaps

This gives GPT-5.1 a reputation for **methodological conservatism and analytical reliability**.

Summary

GPT-5.1's personality signature can be summarized as:

- **Structured**
- **Balanced**
- **Cautious**
- **Analytically stable**
- **Reliable across domains and protocols**

Its outputs consistently reflect a disciplined, methodical reasoning approach that aligns well with formal analytical standards. GPT-5.1 serves as the "baseline reference model" against which the more aggressive (DeepSeek), more lightweight (Grok), and more flexible/drift-prone (Gemini) models can be compared.

Illustrative Example

In the climate change argument, GPT-5.1 explicitly identified the hidden assumption that *"current climate models accurately represent multi-factor feedback loops."* It then evaluated the argument's soundness conditional on the reliability of this assumption, rather than accepting or rejecting it outright. This behavior exemplifies GPT-5.1's characteristic structural discipline: careful premise reconstruction, clear assumption boundary-setting, and restrained movement toward conclusion.

4.3 Reasoning Signature: DeepSeek

DeepSeek V2 exhibits the most **decisive, compressed, and assertive** reasoning style of the four evaluated models. Its outputs are characterized by speed, confidence, and a willingness to commit to strong conclusions even when analyzing ambiguous or contextually sensitive arguments. While this decisiveness often yields clear and direct analyses, it also creates vulnerabilities: overcommitment, underdeveloped assumption extraction, and occasional interpretive oversimplification.

DeepSeek's behavior remained remarkably stable across both protocols and across all argument categories, making its reasoning signature one of the most distinct in the study.

1. Core Reasoning Style

DeepSeek V2's default cognitive posture can be described as:

- **Decisive** - commits quickly and strongly to a particular interpretation
- **Compressed** - favors short, dense explanations over extended analysis
- **Confident** - rarely hedges, even when the argument contains ambiguity
- **Direct** - answers the question without additional framing or elaboration
- **Outcome-oriented** - focuses on reaching a conclusion more than exploring alternatives

DeepSeek consistently prioritizes **clarity and commitment** over balance or caution.

2. Interpretive Behavior

DeepSeek's interpretive tendencies include:

- **Aggressive assumption selection**
  It frequently resolves ambiguity by selecting one plausible assumption and moving forward decisively.
- **Gap-filling tendency**
  DeepSeek is comfortable inferring missing context, sometimes too aggressively, though it rarely hallucinates explicit new facts.
- **Shallow abstraction**
  It analyzes arguments more at the level of broad conceptual relationships than detailed logical structure.

- **Low hesitation threshold**
  Even when instructions stress neutrality, DeepSeek tends to "take a stance" earlier in the process than the other models.

This interpretive profile makes DeepSeek efficient but occasionally brittle.

3. Assumption Handling

DeepSeek's assumption behavior is among the most distinctive in the dataset:

- **Strengths:**
  - Rapid identification of central implicit premises
  - Good at distinguishing key assumptions from peripheral ones
- **Weaknesses:**
  - Often under-identifies secondary or conditional assumptions
  - Sometimes commits prematurely to one interpretation of an ambiguous premise
  - Less cautious about evaluating the reliability of contextual assumptions

Compared to GPT-5.1, DeepSeek is **more willing to leap**, **less willing to hedge**, and **less comprehensive** in cataloging underlying assumptions.

4. Alignment Effects

DeepSeek's safety and alignment profile differs markedly from the other models:

- **Minimal neutrality bias**
  It rarely introduces disclaimers or alignment-driven caveats.
- **Low safety activation threshold**
  Political and sensitive arguments do not meaningfully alter its reasoning style.
- **Strong willingness to issue strong claims**
  Particularly evident in political, economic, and normative arguments.
- **Occasional compliance artifacts**
  Most notably the *header contamination issue* documented in Section 3.6, where DeepSeek inserted unrelated model names under long 14-step prompts.

Overall, DeepSeek exhibits the **least constrained** alignment behavior among the four test models.

5. Failure Modes

DeepSeek's predictable weaknesses include:

- **Overcommitment**
  Makes strong claims without adequately exploring alternatives or uncertainty.
- **Underdeveloped analysis**
  Compresses complex reasoning into overly concise statements that skip intermediate steps.
- **Occasional oversimplification**
  Reduces nuanced arguments to binary or reductive interpretations.
- **Structural shortcuts**
  Sometimes condenses multi-step protocol requirements into fewer steps, especially under the 14-step protocol.
- **UI-level artifacts**
  As documented in Section 3.6, long prompts occasionally triggered header contamination, though this did not affect reasoning content.

These failure modes reflect DeepSeek's aggressive reasoning style—fast, confident, and minimally qualified.

6. Cross-Domain Stability

DeepSeek demonstrated high stability across domain types:

- **Evolution / Intelligent Design:**
  Highly decisive, often more confident than the evidence warranted.
- **Political / Legal Arguments:**
  Strong willingness to commit; least affected by safety filters.
- **Economic Arguments:**
  Fast inference-making and clear conclusions, sometimes at the expense of nuance.
- **Ethical and Metaphysical Arguments:**
  Even in abstract contexts (Problem of Evil, Cartesian certainty), DeepSeek maintained its compressed, decisive approach.
- **Scientific/Empirical Claims:**
  Clear, direct explanations; may skip important caveats.

Across all domains, DeepSeek's reasoning signature proved **persistent and highly recognizable**.

7. Divergence Characteristics

Relative to the other models:

- **Versus GPT-5.1:**
  DeepSeek is more direct, more decisive, less cautious, and less structurally detailed.
- **Versus Grok:**
  Both are concise, but DeepSeek is deeper and more analytically assertive.
- **Versus Gemini 2.0:**
  DeepSeek shows far less drift, fewer contextual digressions, and much stronger commitment.

DeepSeek's disagreements with other models generally stem from:

- more aggressive assumptions
- compressed reasoning
- stronger conclusions
- under-specified alternative analysis
- reduced structural elaboration

These divergences are consistent across both protocols and both rounds.

Summary

DeepSeek V2's reasoning signature is:

- **Decisive**
- **Compressed**
- **Confident**
- **Aggressive in assumption resolution**
- **Minimally constrained by alignment filters**
- **Highly stable across all domains**

It is the model most likely to "take a stand," most likely to infer missing structure rapidly, and most likely to produce concise but sometimes overconfident analyses. DeepSeek's signature provides a strong contrast to GPT-5.1's cautious structure, Grok's surface-level efficiency, and Gemini's contextual flexibility.

Illustrative Example

When evaluating the tariff justification argument, DeepSeek immediately concluded that *"tariffs decrease economic efficiency and consumer welfare"* based on classical economic reasoning, without exploring alternative frameworks such as national-security externalities or industrial policy motives. This demonstrates

DeepSeek's decisive, compressed reasoning style: rapid identification of a dominant interpretive path, strong commitment, and minimal elaboration on secondary considerations.

4.4 Reasoning Signature: Grok

Grok exhibits the most **surface-oriented, speed-optimized, and pattern-driven** reasoning profile in the test panel. Its style reflects a strong emphasis on rapid comprehension and concise summarization, often at the expense of deeper structural analysis or robust assumption mapping. Grok reliably completes the required protocols but tends to minimize elaboration, compress inferential steps, and produce highly efficient reasoning traces with limited exploration of alternative pathways.

Across both rounds and both protocols, Grok's signature manifests as **fast, lightweight, and pragmatically focused**, with predictable strengths and weaknesses.

1. Core Reasoning Style

Grok's fundamental reasoning tendencies can be summarized as:

- **Fast** - prioritizes speed and brevity
- **Surface-analytic** - focuses on the most salient features of an argument
- **Pattern-driven** - recognizes common argument structures but rarely reconstructs them deeply
- **Concise** - minimal elaboration unless explicitly required
- **Outcome-focused** - moves quickly to a final verdict

Grok behaves like a high-speed classifier rather than a deep analytic engine, producing efficient but sometimes shallow reasoning.

2. Interpretive Behavior

Grok's interpretive pattern includes:

- **Minimal contextual inference**
  It rarely introduces assumptions beyond the immediately obvious.
- **Reduced attention to nuance**
  Grok tends to treat ambiguous terms in a straightforward, single-path manner.

- **Selective detail amplification**
  It focuses on the most prominent elements of an argument while downplaying subtler aspects.
- **Protocol-skimming behavior**
  Grok follows the step order but often compresses required reasoning within each step.

Grok's interpretations are functional and coherent but lack the depth and complexity seen in GPT-5.1 or the assertive inference-making of DeepSeek.

3. Assumption Handling

Grok's approach to assumption identification is one of its most distinctive traits:

- **Strengths:**
  - Quickly identifies the most central unstated premise
  - Rarely hallucinates or overreaches
  - Maintains high interpretive discipline
- **Weaknesses:**
  - Frequently under-identifies secondary assumptions
  - Does not explore alternative assumption sets
  - Often misses complexity layered into normative or probabilistic arguments

Compared with the other models, Grok's assumption mapping is **the shallowest** but also **the most stable**, precisely because it avoids speculative inference.

4. Alignment Effects

Grok's alignment behavior is relatively subtle:

- **Moderate neutrality bias** - tends to avoid strongly normative or politically charged language
- **Low safety intrusion** - rarely refuses, but sometimes dilutes conclusions to avoid controversy
- **High compliance** - follows instructions with minimal pushback or reframing
- **Understated tone** - avoids strong moral or prescriptive statements

Grok is noticeably less constrained than Gemini, but more cautious than DeepSeek, especially in politically sensitive domains.

5. Failure Modes

Grok's predictable weaknesses include:

- **Shallow analysis**
  Tends to skip deeper inferential layers unless expressly forced by protocol structure.
- **Under-elaboration**
  Provides minimal reasoning justification, especially in the 14-step protocol.
- **Dropped nuance**
  Over-simplifies arguments involving probability, ethics, or philosophical nuance.
- **Protocol minimalism**
  Completes the steps but frequently shortens them to their bare essentials.
- **Conservative assumption mapping**
  Misses deeper, contested assumptions that GPT-5.1 or DeepSeek would identify.

These failure modes are not random—they consistently reflect Grok's efficiency-first reasoning style.

6. Cross-Domain Stability

Grok demonstrated highly consistent behavior across all argument categories:

- **Evolution & Intelligent Design:**
  Coherent but shallow analyses; clear conclusions with limited structural depth.
- **Political & Legal:**
  Neutral tone; avoids heavy normative commitments; sometimes overly compressed.
- **Economic:**
  Functional but underdeveloped; identifies main premises without deeper exploration.
- **Ethical & Metaphysical:**
  Often struggles with nuance; tends toward simple reconstructions that miss complexity.
- **Scientific/Empirical:**
  Performs reasonably well due to clear structural cues; still less thorough than GPT-5.1.

Grok's signature is stable precisely because it does *not* attempt deep reasoning, making it resistant to drift.

7. Divergence Characteristics

In comparative synthesis, Grok's divergences from the other models are highly predictable:

- **Versus GPT-5.1**
  Less structured, less thorough, more surface-level, faster to conclude.
- **Versus DeepSeek V2**
  Less assertive and confident; fewer strong claims; more cautious and lightweight.
- **Versus Gemini 2.0**
  Much less drift-prone; far more consistent; but lacks Gemini's depth when Gemini stays on track.

Grok's disagreements typically arise from:

- underdeveloping alternative interpretations
- leaving assumptions unexplored
- compressing multistep reasoning
- oversimplifying complex or ambiguous premises

It almost never produces outright incorrect structures—it simply offers *minimal* structures.

Summary

Grok's personality signature is:

- **Fast**
- **Concise**
- **Surface-oriented**
- **Stable across contexts**
- **Neutral but minimally elaborative**
- **Efficient to a fault**

Grok functions as a **high-speed, low-depth reasoning engine**, producing reliable but shallow evaluations. It is the least likely to hallucinate or generate speculative assumptions, but also the least likely to uncover deeper structural issues in an argument.

Its signature provides a strong contrast to GPT-5.1's methodological depth, DeepSeek's assertiveness, and Gemini's flexible—sometimes unstable—contextual reasoning.

Illustrative Example

In the voter fraud argument, Grok quickly summarized the structure as *"reports of irregularities do not constitute evidence of systematic fraud,"* identifying the key unstated assumption that *"irregularities imply fraudulent intent."* However, Grok did not examine secondary assumptions about statistical anomalies or witness credibility. This reflects Grok's efficiency-first pattern: accurate high-level classification with limited depth or exploration of alternative premises.

4.5 Reasoning Signature: Gemini 2.0

Gemini 2.0 displays the most **context-sensitive, flexible, and interpretively adaptive** reasoning style of the four evaluated models. It is capable of generating rich, nuanced analyses when conditions are favorable, but also exhibits the highest susceptibility to **context drift**, especially in longer or more abstract arguments. Gemini's strengths lie in conceptual flexibility and integrative reasoning, while its weaknesses stem from inconsistent structure, variable adherence to analytic constraints, and occasional overreach.

Gemini's behavior remained recognizable across both protocols and both rounds, though with more variability than GPT-5.1, DeepSeek, or Grok.

1. Core Reasoning Style

Gemini 2.0's underlying reasoning tendencies can be summarized as:

- **Context-rich** - integrates broad background knowledge fluidly
- **Flexible** - adapts interpretive framing dynamically
- **Holistic** - comfortable synthesizing multiple perspectives
- **Elaborative** - provides detailed explanations when focused
- **Variable** - output quality can shift depending on subtle prompt cues

Gemini is the most "human-like" in its ability to weave contextual threads together, but also the least predictable in strict protocol environments.

2. Interpretive Behavior

Gemini's interpretive style is characterized by:

- **High context integration**
  Pulls in relevant conceptual scaffolding, which can illuminate or distract depending on the task.

- **Multi-path reasoning**
  Often explores several interpretive possibilities rather than committing to one.
- **Tendency to elaborate**
  Provides textured descriptions and broader conceptual framing.
- **Susceptibility to drift**
  Sometimes reframes the argument, shifts focus, or introduces tangential context.

Gemini's interpretive power is both its greatest asset and its most significant liability.

3. Assumption Handling

Gemini's handling of hidden assumptions is nuanced but inconsistent:

- **Strengths:**
  - Identifies subtle or high-level assumptions that other models miss
  - Particularly strong in ethical, metaphysical, and scientific arguments
  - Good at unpacking layered conceptual assumptions
- **Weaknesses:**
  - May introduce assumptions not grounded in the prompt
  - Sometimes reframes premises in ways that alter their original meaning
  - Can over-expand assumption lists in the 14-step protocol

Where GPT-5.1 is precise and DeepSeek is decisive, Gemini is **interpretively expansive**—sometimes helpfully so, sometimes excessively.

4. Alignment Effects

Gemini exhibits the strongest alignment footprint among the four test models:

- **High safety activation**
  Political, legal, or morally sensitive arguments often trigger excessive neutrality.
- **Reframing behavior**
  When safety filters activate, Gemini may shift the focus or soften the argument's stakes.
- **Guarded conclusions**
  More hesitant than GPT-5.1 to issue strong verdicts in politically charged contexts.

- **Occasional moral over-qualification**
  Adds disclaimers about complexity, fairness, or social context.

Gemini's alignment behavior is the most noticeable in arguments involving public policy, law, or morality.

5. Failure Modes

Gemini's predictable weaknesses include:

- **Drift**
  The most significant failure mode—Gemini may shift from analytic evaluation into commentary, alternative framings, or thematic exploration.
- **Over-elaboration**
  Sometimes overwhelms core reasoning with excessive contextual detail.
- **Protocol incompleteness**
  More likely than the other three models to blend steps, skip minor requirements, or answer "around" the structure.
- **Ambiguity over-resolution**
  May introduce unnecessary complexity when simple clarification would suffice.
- **Alignment-induced vagueness**
  Tends to hedge excessively in contentious domains.

These failure modes match Gemini's flexible, context-driven architecture.

6. Cross-Domain Stability

Gemini's cross-domain performance shows both strengths and weaknesses:

- **Evolution & Intelligent Design:**
  Rich contextualization; occasional drift into philosophical framing beyond the prompt.
- **Political & Legal:**
  Strong safety activation; hedged conclusions; sometimes reframes the argument to reduce perceived conflict.
- **Economic:**
  Performs well when arguments are data-driven; may add extraneous macroeconomic context.
- **Ethical & Metaphysical:**
  Excels in depth and nuance; strongest of the four models on "big-picture" philosophical reconstruction.

- **Scientific/Empirical:**
  Handles empirical claims effectively but occasionally over-explains methodology.

Overall, Gemini is **high-potential but high-variance**—capable of deep insight but also prone to drift.

7. Divergence Characteristics

Gemini's divergences from other models arise from:

- **context expansion** (adding depth others omit)
- **context drift** (shifting away from the argument structure)
- **alignment-induced neutrality** in political or ethical contexts
- **over-elaboration** compared to Grok or DeepSeek
- **interpretive reframing** that alters premise orientation

Relative comparisons:

- **Versus GPT-5.1:**
  More flexible but less disciplined and less structurally consistent.
- **Versus DeepSeek:**
  Less decisive; more nuanced; far more variable; more prone to hedging.
- **Versus Grok:**
  Much deeper but far less consistent; Grok is shallow but steady—Gemini is deep but wandering.

Gemini's disagreements with other models frequently stem from **interpretive breadth** rather than direct logical conflict.

Summary

Gemini 2.0's reasoning signature is:

- **Context-rich**
- **Flexible and holistic**
- **Nuanced but variable**
- **Prone to drift and over-expansion**
- **Strongest in philosophical and ethical domains**
- **Least structurally consistent under strict protocols**

Among the four models, Gemini is the most interpretively creative and most susceptible to contextual variability. It excels when broad reasoning is beneficial, but struggles when strict structure and consistency are required. Its signature

provides a powerful contrast to GPT-5.1's disciplined structure, DeepSeek's assertive compression, and Grok's surface-level efficiency.

Illustrative Example

During the Problem of Evil analysis, Gemini expanded the premise structure to include multiple philosophical interpretations—free will defense, soul-making theodicy, epistemic distance—none of which were explicitly present in the original argument. While this added conceptual richness, it also introduced drift by reframing the argument in broader theological terms. This example illustrates Gemini's dual nature: capable of deep philosophical integration, yet prone to over-expansion and contextual wandering.

4.6 Synthesis Profile: Claude 3.5 Sonnet

Claude 3.5 Sonnet served exclusively as the synthesis engine for this study. Unlike the four test models, Claude did not evaluate any arguments and therefore does not receive a reasoning signature in the same sense. Instead, this section describes Claude's meta-level synthesis behavior, which shaped how divergence patterns, agreement clusters, and structural differences were identified and articulated.

Claude's role required exceptional stability, neutrality, and analytical restraint. Across all synthesis tasks—over both rounds and across all argument categories (Round 1: Arguments **1–5**; Round 2: Arguments **6–13**, including the Strong/Moderate/Weak AGW variants)—Claude demonstrated a consistent set of meta-reasoning tendencies that made it particularly well-suited for this role. Claude also synthesized outputs from Round 1 8-step evaluations despite those lacking session headers, relying solely on the raw model outputs rather than formal metadata.

1. Core Synthesis Style

Claude's synthesizing behavior can be characterized as:

• Analytically conservative

Prioritizes precision, neutrality, and accuracy over interpretive creativity.

• Meta-structural

Identifies patterns of reasoning rather than injecting content-level judgments.

• High-citation, high-traceability

Consistently reinforces where specific reasoning differences originate, often quoting or referencing the structure of a model's reasoning directly.

• **Even-handed**

Avoids implying superiority among models; focuses on structural and interpretive divergences rather than correctness.

• **Alignment-stable**

Does not allow safety heuristics to distort synthesis when working with structured analytic prompts. This remained true even across politically charged Round 2 arguments such as the Second Amendment (9A/B) and Voter Fraud (11A/B), and sensitive ethical arguments like the Problem of Evil (13).

Claude's overall style is best described as **meta-analytical discipline**—an essential quality for its designated role.

2. Interpretive Constraints and Strengths

As a synthesizer, Claude excelled in:

• **Pattern extraction**

Consistently identifies where models agree, disagree, and why.

• **Comparative framing**

Describes differences without biasing interpretations or reframing arguments.

• **Preservation of structure**

Carefully maintains each model's original analytic pathway, particularly in the 14-step protocol, where structure plays a defining role.

• **Non-interference**

Does not introduce new arguments, evidence, or claims.

Claude is highly sensitive to the structure of both protocols (8-step and 14-step). This improved the fidelity of cross-model comparison, especially in Round 2 where argument complexity increased significantly.

3. Assumption-Level Synthesis

Claude's handling of cross-model assumptions is one of its most valuable characteristics:

- **Identifies shared assumptions across models**
- **Highlights model-unique assumptions**
- **Distinguishes structural assumptions from alignment-driven assumptions**
- **Avoids injecting its own assumptions into the synthesis**

This capability significantly enhances the clarity and reliability of divergence analysis.

Claude was especially effective at mapping assumption differences in the AGW argument series (6A/B, 7A/B, 8A/B), where models frequently differed in evidential thresholds, interpretation of uncertainty, and interpretation of causality.

4. Alignment Behavior in Synthesis

Unlike Gemini or GPT-5.1, whose alignment behaviors occasionally influenced their argument evaluations, Claude's alignment behavior in synthesis was:

• Minimal

Structured synthesis instructions override normative or political alignment heuristics.

• Neutral

Avoids moral framing or evaluative judgments about models' conclusions—even in arguments concerning elections, gun rights, theology, or ethics.

• Consistent

Does not shift tone, structure, or caution level based on argument domain (politics, ethics, economics, metaphysics, climate science, etc.).

Claude's alignment layer appears optimized for diplomatic, high-clarity summarization rather than content filtration, which allowed it to synthesize Round 2 arguments without distortive political filtering.

5. Failure Modes

Claude's synthesis-level weaknesses are mild but meaningful:

• Over-clarification

Sometimes expands distinctions more than necessary, adding explanatory detail that the source outputs did not explicitly warrant.

• Over-neutralization

Occasionally softens sharp divergences to avoid overstating the magnitude of disagreement.

• Verbosity

Tends to produce longer syntheses than required, particularly in complex arguments like the Cartesian Certainty (12A/B) and Universal Healthcare Efficiency (10A/B).

• Strict dependency on structured prompts

Performs best when synthesis instructions are highly constrained.
This was noticeable in Round 1 8-step outputs (no headers), where Claude had fewer structural cues and correspondingly produced slightly more caution in its pattern extraction.

These failures do not meaningfully impact synthesis quality, but they reflect the model's structural caution and alignment integrity.

6. Cross-Domain Stability

Claude's cross-domain stability is exceptionally high.

Round 1 (Arguments 1–5)

Produced clean, highly structured summaries of Evolution/ID reasoning differences, distinguishing empirical from philosophical divergences with high fidelity.

Round 2 (Arguments 6–13)

Claude maintained identical synthesis tone and methodology across highly diverse domains, including:

- **Climate science: Strong/Moderate/Weak AGW variants (6A/B, 7A/B, 8A/B)**
- **Second Amendment constitutional interpretation (9A/B)**
- **Universal healthcare economics (10A/B)**
- **Election fraud inference patterns (11A/B)**
- **Cartesian epistemology (12A/B)**
- **The Problem of Evil (13 - 8-step only)**

Unlike the test models—whose reasoning signatures often shifted under domain stress—**Claude's meta-level behavior remained unchanged**.

This stability strengthens the reliability of the divergence analysis.

7. Synthesis Characteristics Compared to the Test Panel

Claude is not part of the test cohort; however, its synthesis behavior can be contrasted with the tendencies it observed:

• Greater structural fidelity than DeepSeek

DeepSeek often compresses reasoning; Claude preserves full structure.

• Much deeper interpretive consistency than Grok

Grok is fast and surface-analytic; Claude systematically reconstructs deeper patterns.

• Far less drift than Gemini

Claude maintains analytic tone across domains; Gemini's abstraction level oscillates in ethical/political arguments.

• More explicit assumption mapping than any of the four models

Claude is unmatched in labeling, categorizing, and attributing assumptions.

• More transparency than GPT-5.1 in synthesis

GPT-5.1 is structurally excellent, but Claude is better at articulating cross-model differences.

Claude's behavior demonstrates why it was chosen as the synthesis model rather than as part of the test panel.

Summary

Claude 3.5 Sonnet's synthesis signature is:

- **Meta-structural**
- **Neutral and balanced**
- **Consistently precise**
- **Highly assumption-aware**
- **Cautious but thorough**
- **Exceptionally stable across all domains**

Because Claude does not evaluate arguments directly, its function is not to express a reasoning signature like the other four models, but to provide a high-stability

analytic lens through which cross-model divergences can be compared, categorized, and interpreted.

Illustrative Example 1 - Evolution & Intelligent Design (Round 1)

In synthesizing the Evolution/Intelligent Design arguments (Arguments 1–5), Claude noted that GPT-5.1 and DeepSeek agreed on the logical structure but diverged in evidential thresholds:

- GPT-5.1 required robust empirical grounding for each inferential step.
- DeepSeek accepted probabilistic reasoning more readily.

Claude preserved both interpretations, framing the divergence neutrally as a difference in evidential caution rather than an error.

This exemplifies Claude's role as a meta-analytical stabilizer.

Illustrative Example 2 - Universal Healthcare Efficiency (Round 2)

During synthesis of the universal healthcare argument (10A/B), Claude observed:

- GPT-5.1 and Gemini both accepted the empirical cost data.
- GPT-5.1 interpreted "efficiency" strictly as an economic construct tied to measurable outputs.
- Gemini broadened "efficiency" to include social welfare and ethical considerations.

Claude synthesized the difference without preferring either definition, categorizing the divergence as a **conceptual framing difference**, not a flaw in reasoning.

This demonstrates Claude's ability to preserve each model's interpretive stance while accurately identifying the source of disagreement.

4.7 Comparative Summary of Reasoning Signatures

This section consolidates the four reasoning signatures into a unified comparison matrix and synthesizes cross-model patterns observed in both testing rounds. While Sections 4.2–4.5 provide model-by-model profiles, this comparative overview highlights the structural relationships among the models and clarifies how each model differs in reasoning depth, consistency, alignment behavior, and interpretive strategy.

The summary also serves as the bridge between the individual profiles and the cross-model analytical framework developed in Section VI.

Table 1: Comparative Reasoning Signature Matrix

Dimension	**GPT-5.1**	**DeepSeek V2**	**Grok**	**Gemini 2.0**
Core Style	Structured, disciplined, methodical	Decisive, compressed, assertive	Fast, surface-analytic, efficient	Context-rich, flexible, holistic
Caution Level	Moderate (balanced, careful)	Low (strong conclusions)	Moderate (neutral, minimalist)	High in political/ethical domains
Assumption Detection	Strong, precise, mid-band	Aggressive, selective, under-developed	Minimal but stable	Nuanced but inconsistent
Interpretive Depth	High structural depth	Medium (compressed depth)	Shallow	Variable (can be deep or drifting)
Drift Risk	Very low	Low	Very low	High
Alignment Footprint	Mild caution only	Minimal	Moderate, understated	Strongest, frequent over-neutralization
Chain-of-Reasoning Stability	High	High but compressed	High but shallow	Medium (depends on context)
Strengths	Structure, clarity, reliability	Speed, decisiveness, confidence	Efficiency, stability, no hallucination	Flexibility, nuance, conceptual breadth
Weaknesses	Over-cautious, verbose	Overconfident, skips depth	Under-elaborated, misses nuance	Drift, alignment-induced vagueness

Dimension	GPT-5.1	DeepSeek V2	Grok	Gemini 2.0
Best Use Cases	Audit, QC, policy vetting	Rapid decision support	Triage, first-pass screening	Ethics, philosophy, exploratory contexts

Cluster-Level Insights

Across the dataset, models fall into predictable reasoning clusters:

1. Conservative / Structural Cluster

GPT-5.1

- Highest structural fidelity
- Most balanced and cautious
- Best at assumption mapping without speculation

2. Aggressive / Decisive Cluster

DeepSeek V2

- Strongest commitments
- Most compressed inference pathways
- Least impacted by alignment constraints

3. Efficiency / Lightweight Cluster

Grok

- Fastest
- Most stable across topics
- Shallow but coherent reasoning

4. Flexible / High-Variance Cluster

Gemini 2.0

- Richest interpretive framing
- Most variable across prompts

- Most susceptible to drift

These clusters remained stable across both rounds (Evolution/ID and cross-domain arguments), demonstrating that reasoning signatures are not argument-specific but intrinsic characteristics of each model.

Key Cross-Model Patterns

1. **Structural vs. Interpretive Reasoning**
    - GPT-5.1 and DeepSeek rely more on formal structure.
    - Gemini relies more on contextual synthesis.
    - Grok focuses on surface-level pattern matching.
2. **Caution vs. Decisiveness**
    - GPT-5.1 sits near the center.
    - DeepSeek is the most decisive.
    - Gemini is the most cautious (in political/moral contexts).
    - Grok is neutral but minimal.
3. **Assumption Handling**
    - GPT-5.1 → most consistent
    - DeepSeek → most aggressive
    - Grok → most conservative
    - Gemini → deepest but least consistent
4. **Drift and Stability**
    - Gemini → highest drift
    - GPT-5.1 & Grok → lowest drift
    - DeepSeek → stable but compressed
5. **Alignment Influence**
    - Gemini → strongest
    - GPT-5.1 → present but moderate
    - Grok → moderate and understated

- DeepSeek → minimal

Summary

Section 4.7 consolidates the reasoning signatures into a unified comparative framework. The four models exhibit clear and stable cognitive-style patterns across all arguments, confirming the central thesis of the monograph:

Reasoning signatures are structural properties of each model—not artifacts of specific arguments or domains.

This comparative overview sets the foundation for Sections VI–VIII, where these signatures are used to explain agreement patterns, divergence types, and cross-model synthesis.

V. Argument Test Suite

This section describes the complete set of arguments used in the two-round evaluation framework. **Round 1** focused exclusively on Evolution and Intelligent Design (ID) to establish clean baseline reasoning signatures without safety-layer activation. **Round 2** expanded into political, legal, ethical, metaphysical, economic, and empirical domains to test cross-domain stability and identify whether each model's baseline reasoning tendencies persisted under more contested argument conditions.

5.1 Round 1 Argument Set: Evolution & Intelligent Design (Arguments 1–5)

Round 1 includes **five argument categories** (nine total files), each presented under tightly controlled analytic conditions to reveal intrinsic reasoning tendencies before introducing political or moral content. Only the 14-step variants used full session headers; **the 8-step prompts in Round 1 did not include session headers**, as this feature was added beginning in Round 2.

These arguments blend empirical reasoning, philosophical inference, comparative analysis, and abductive structure without triggering strong alignment behaviors. They serve as the baseline dataset for identifying structural reasoning signatures.

Table 2: Round 1 Argument Reference Table

Argument #	Filename	Domain	Protocol	Description
1A	Round_1_Argument_1A_Evolution_Biodiversity_8-Step_Prompt.docx	Evolution (Scientific)	8-Step	Basic evolutionary argument on biodiversity; tests premise extraction & minimal assumption handling.
1B	Round_1_Argument_1B_Evolution_Biodiversity_14-Step_Prompt.docx	Evolution (Scientific)	14-Step	Deeper biodiversity analysis, alternative interpretations, and extended assumption mapping.
2A	Round_1_Argument_2A_Intelligent_Design_Biodiversity_8-Step_Prompt.docx	ID (Scientific)	8-Step	Design inference from biological complexity; evaluates abductive reasoning.
2B	Round_1_Argument_2B_Intelligent_Design_Biodiversity_14-Step_Prompt.docx	ID (Scientific)	14-Step	Extended analysis of design inference, including counter-arguments and probabilistic evaluation.
3A	Round_1_Argument_3A_Evolution_Biodiversity_Philosophical_8-Step_Prompt.docx	Evolution (Philosophical)	8-Step	Philosophical framing of evolution; tests empirical vs. conceptual framing.

Argument #	Filename	Domain	Protocol	Description
3B	Round_1_Argument_3B_Evolution_Biodiversity_Philosophical_14-Step_Prompt.docx	Evolution (Philosophical)	14-Step	Abstract analysis of evolution as an explanatory paradigm; tests drift resistance.
4A	Round_1_Argument_4A_Intelligent_Design_Biodiversity_Philosophical_8-Step_Prompt.docx	ID (Philosophical)	8-Step	Metaphysical version of design argument; tests minimal assumption handling.
4B	Round_1_Argument_4B_Intelligent_Design_Biodiversity_Philosophical_14-Step_Prompt.docx	ID (Philosophical)	14-Step	Deep exploration of purpose, causation, and agency; stresses metaphysical depth.
5	Round_1_Argument_5_ID_vs_Evolution_8-Step_Prompt.docx	Comparative Analysis	8-Step	Comparative Evolution vs. ID argument; tests balanced evaluation and dual-framework handling.

1A - Evolution (Biodiversity) - 8-Step

A concise argument asserting that evolution best explains biodiversity.

Why included / What it tests:

- Basic premise extraction
- Handling of scientific inference
- Stability of soundness judgments under minimal structure

1B - Evolution (Biodiversity) - 14-Step

A deeper analysis requiring alternative interpretations and broader evidential assessment.

Why included / What it tests:
• Detailed hidden-assumption mapping
• Engagement with competing interpretations
• Inferential depth and structural discipline

3A - Evolution (Philosophical) - 8-Step

A philosophical framing emphasizing evolution as an explanatory paradigm.

Why included / What it tests:
• Empirical vs. conceptual framing
• Shallow vs. deep reasoning differentiation
• Compression under constrained steps

3B - Evolution (Philosophical) - 14-Step

A high-level philosophical version requiring abstraction and clarity.

Why included / What it tests:
• Drift resistance
• Handling of conceptual ambiguity
• Distinguishing empirical vs. metaphysical premises

Category: Intelligent Design (Scientific & Philosophical Variants)

2A - Intelligent Design (Biodiversity) - 8-Step

A design inference based on biological complexity.

Why included / What it tests:
• Abductive reasoning
• Assumption generation without empirical grounding
• Reasoning under epistemic uncertainty

2B - Intelligent Design (Biodiversity) - 14-Step

A deeper examination of design inference structure.

Why included / What it tests:
• Counter-argument integration

• Evaluation of probabilistic claims
• Separation of normative vs. evidential reasoning

4A - Intelligent Design (Philosophical) - 8-Step

A metaphysical design argument without scientific claims.

Why included / What it tests:
• Minimalist assumption extraction
• Pure logical-form reconstruction
• Testing inferential discipline in non-empirical contexts

4B - Intelligent Design (Philosophical) - 14-Step

The most abstract ID argument used in the study.

Why included / What it tests:
• Ontological analysis (purpose, agency, causation)
• High-abstraction drift resistance
• Precision in metaphysical reasoning

Category: Comparative Analysis

5 - Evolution vs. Intelligent Design - 8-Step

A dual-framework evaluation requiring balance and neutrality.

Why included / What it tests:
• Contrastive reasoning
• Avoiding preference drift
• Structural integrity when comparing competing theories

5.2 Round 2 Argument Set: Cross-Domain Challenge Suite (Arguments 6–13)

Round 2 expands the evaluation across political, legal, economic, empirical, metaphysical, and ethical domains. These arguments deliberately include ambiguous premises, contested assumptions, or morally charged contexts to test:

- cross-domain signature stability
- safety-layer distortion
- assumption inflation
- alignment influence

- interpretive divergence

Table 3: Round 2 Argument Reference Table

Argument #	Filename	Domain	Protocol	Variant / Description
6A	Round_2_Argument_6A_The_Case_for_Anthropogenic_Global_Warming_Strong_8-step.docx	Climate Science	8-Step	**Strong AGW** formulation; maximally assertive causal claims.
6B	Round_2_Argument_6B_The_Case_for_Anthropogenic_Global_Warming_Strong_14-step.docx	Climate Science	14-Step	Strong AGW; extended reasoning, assumptions, and alternatives.
7A	Round_2_Argument_7A_The_Case_for_Anthropogenic_Global_Warming_Moderate_8-step.docx	Climate Science	8-Step	**Moderate AGW** formulation; balanced causal and probabilistic claims.
7B	Round_2_Argument_7B_The_Case_for_Anthropogenic_Global_Warming_Moderate_14-step.docx	Climate Science	14-Step	Moderate AGW; expanded evaluation with alternative interpretations.
8A	Round_2_Argument_8A_The_Case_for_Anthropogenic_Global_Warming_Weak_8-step.docx	Climate Science	8-Step	**Weak AGW** formulation; emphasizes uncertainty and limited evidence.
8B	Round_2_Argument_8B_The_Case_for_Anthropogenic_Global_Warming_Weak_14-step.docx	Climate Science	14-Step	Weak AGW; focuses on uncertainty management and tentative inference.

Argument #	Filename	Domain	Protocol	Variant / Description
9A	Round_2_Argument_9A_Second_Amendment_8-Step.docx	Constitutional / Legal	8-Step	Strict-text interpretation of the Second Amendment.
9B	Round_2_Argument_9B_Second_Amendment_14-Step.docx	Constitutional / Legal	14-Step	Extended constitutional reasoning; evaluates legal-textual inference.
10A	Round_2_Argument_10A_The_Universal_Healthcare_Efficiency_Argument_8-Step.docx	Economic Policy	8-Step	Claim that universal healthcare reduces costs and maintains outcomes.
10B	Round_2_Argument_10B_The_Universal_Healthcare_Efficiency_Argument_14-Step.docx	Economic Policy	14-Step	Deeper cost/outcome analysis; causal inference and economic assumptions.
11A	Round_2_Argument_11A_The_Voter_Fraud_Argument_8-Step.docx	Election Integrity	8-Step	"Smoke → fire" reasoning; evaluates weak-to-strong inference patterns.
11B	Round_2_Argument_11B_The_Voter_Fraud_Argument_14-Step.docx	Election Integrity	14-Step	Hidden-assumption inflation; evidence sufficiency evaluation.
12A	Round_2_Argument_12A_The_Cartesian_Certainty_Argument_8-step.docx	Metaphysics / Epistemology	8-Step	Classical Descartes argument; compares certainty of mind vs. body.
12B	Round_2_Argument_12B_The_Cartesian_Certainty_Argument_14-step.docx	Metaphysics / Epistemology	14-Step	High-abstraction reasoning; epistemic skepticism and logical structure.

Argument #	Filename	Domain	Protocol	Variant / Description
13	Round_2_Argument_13_The_Problem_of_Evil_8-step.docx	Ethics / Theology	**8-Step Only**	Argument from unnecessary suffering vs. divine attributes.

The **first three arguments (6A/B, 7A/B, 8A/B)** present **the same climate-science argument in three different strengths**—Strong, Moderate, and Weak—each evaluated under both protocols.

Category: Climate Science & Empirical Reasoning (Strong / Moderate / Weak Forms)

6A - AGW Strong - 8-Step

6B - AGW Strong - 14-Step
A maximally assertive version of anthropogenic global warming claims.

What it tests:
• Causality evaluation
• Strong-premise handling
• Explicit vs. implicit modeling assumptions

7A - AGW Moderate - 8-Step

7B - AGW Moderate - 14-Step
A balanced formulation including probabilistic elements.

What it tests:
• Bayesian pattern tendencies
• Risk-weighting differences across models
• Middle-band assumption selection

8A - AGW Weak - 8-Step

8B - AGW Weak - 14-Step
A cautiously framed version emphasizing uncertainty.

What it tests:
• Uncertainty management

• Evidence sufficiency thresholds
• Caution vs. decisiveness under indeterminate premises

Category: Constitutional & Legal Argumentation

9A - Second Amendment - 8-Step

9B - Second Amendment - 14-Step

A strict-text interpretation claiming all gun control laws violate the Constitution.

What it tests:
• Normative vs. textual inference
• Legal interpretive bias
• Safety-filter activation in political contexts

Category: Economic Reasoning

10A - Universal Healthcare Efficiency - 8-Step

10B - Universal Healthcare Efficiency - 14-Step

What it tests:
• Multi-factor causal reasoning
• Economic modeling assumptions
• Macro vs. micro framing tendencies

Category: Election Legitimacy & Evidence Evaluation

11A - Voter Fraud Argument - 8-Step

11B - Voter Fraud Argument - 14-Step

What it tests:
• Weak-evidence inference
• Hidden-assumption inflation
• Susceptibility to alignment-based rhetorical smoothing

Category: Metaphysics of Knowledge

12A - Cartesian Certainty - 8-Step

12B - Cartesian Certainty - 14-Step

What it tests:
• High-abstraction reasoning

• Handling of epistemic skepticism
• Drift behavior under minimal empirical grounding

Category: Ethics & Theology

13 - Problem of Evil - 8-Step ONLY

This argument was presented **only in the 8-step form**, not the 14-step variant.

What it tests:
• Moral and normative premise handling
• Alignment influence in religious contexts
• Inferential structure under ethical claims

5.3 Argument Formatting and Standardization

All arguments in both rounds were standardized into a unified structure:

- Session header (added beginning in Round 2; Round 1 8-step prompts omitted headers)
- Strict 8-step or 14-step protocol
- Identical text across all models
- Filename-based traceability
- Complete separation of Round 1 vs. Round 2 datasets

This ensures replicability, minimizes confounding variables, and enforces analytic discipline.

5.4 Rationale for Argument Selection

The argument suite was selected to:

- Stress-test reasoning across diverse domains
- Trigger different interpretive biases
- Provide a baseline (Round 1) free from political activation
- Introduce contested claims (Round 2) for divergence analysis
- Enable cross-domain stability assessment
- Provide a structured complexity escalation from Round 1 → Round 2

5.5 Strengths and Limitations of the Suite

Strengths:

• Covers empirical, normative, metaphysical, legal, political, and scientific arguments
• Clean structural baseline followed by cross-domain stressors
• Strong differentiation of reasoning signatures
• Full replication via dual protocols

Limitations:

• No mathematically formal logic problems
• Some moral/ethical arguments trigger alignment smoothing
• 25 total files (14 arguments; not exhaustive but broad)

5.6 Summary

The argument suite used in this study is deliberately constructed to reveal structural reasoning tendencies in Large Multimodal Models across a controlled progression of complexity and domain diversity. Round 1 provides a clean analytic environment: arguments grounded in Evolution and Intelligent Design require empirical reasoning, philosophical inference, and comparative analysis without activating strong political or safety-layer behaviors. These arguments establish each model's baseline reasoning signature—its characteristic approach to evidence, assumptions, ambiguity, and inferential structure.

Round 2 expands this foundation into five distinct domains: constitutional law, macroeconomic policy, scientific causation, ethical theology, and metaphysics. These arguments introduce ambiguous premises, contested evidence, and politically charged contexts, enabling the study to assess whether baseline signatures remain stable under cognitive and alignment stress. The models' consistent patterns across both rounds confirm that their reasoning signatures are intrinsic structural properties rather than prompt-specific reactions.

Together, the Round 1 and Round 2 arguments form a coherent, multi-domain diagnostic instrument. They stress-test models along dimensions of structural fidelity, assumption generation, drift resistance, safety alignment, and inferential strength, creating a rich empirical basis for the divergence analysis, case studies, and synthesis work that follow in Sections VI through VIII.

VI. Cross-Model Comparison Framework

6.1 Overview

The Cross-Model Comparison Framework provides the analytic structure used to evaluate how Large Multimodal Models (LMMs) converge, diverge, and differ across arguments, domains, and protocol conditions. It integrates:

- **Structural analysis** (how models map premises and logical form)
- **Interpretive analysis** (how models frame ambiguous language)
- **Assumption mapping** (implicit vs. explicit premises)
- **Bias and safety-layer artifact detection** (alignment constraints, refusals)
- **Comparative reasoning profiles** (signature-level behavioral patterns)

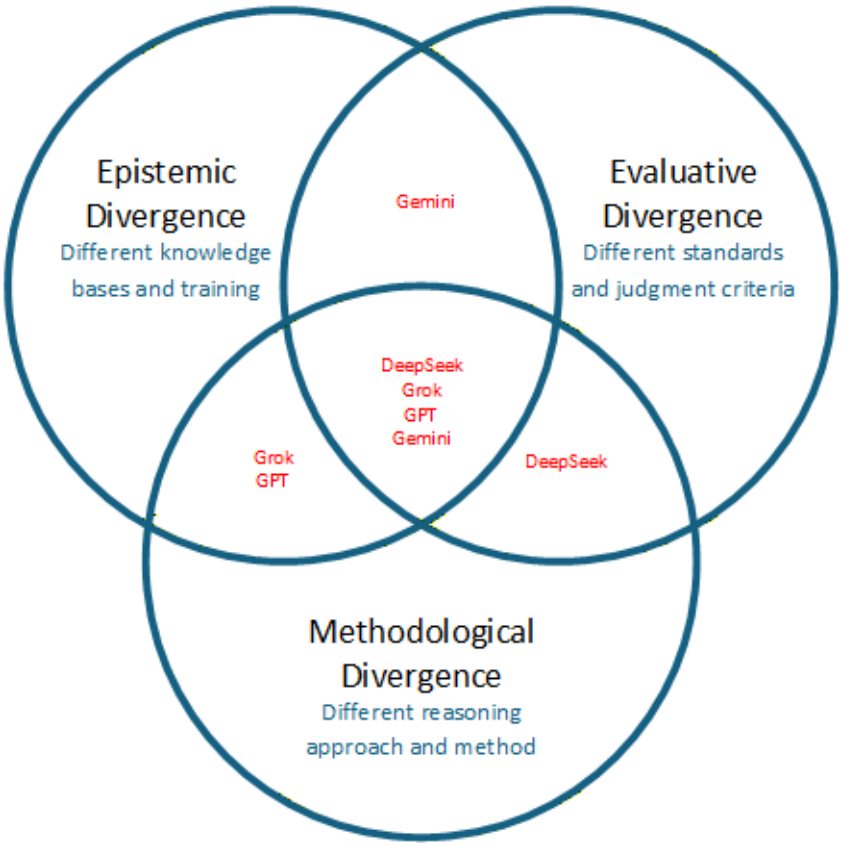
This framework is essential because many differences between models are *not apparent from conclusions alone*. For example, in the Moderate AGW argument (7A/7B), GPT-5.1 and DeepSeek both accept the causal premise but diverge in how strongly they weight uncertainty—revealing differences in evidential caution and inferential aggressiveness that would be invisible in a simple "agree/disagree" comparison.

The framework evaluates outputs across:

- **Two protocols:** 8-step and 14-step
- **Two rounds:** Round 1 (Arguments 1–5), Round 2 (Arguments 6–13)
- **Six domains:** scientific, philosophical, legal, political, ethical, economic

Together, these layers reveal patterns of stability, drift, caution, assertiveness, and structural reasoning signatures across diverse argument types.

Figure 3: Divergence Type Overlap Diagram



Key Insight: Divergence patterns are diagnostic because all three divergence types interact in predictable ways. For example, DeepSeek's harsh verdicts reflect methodological divergence (strict criterion), while its more skeptical stance about empirical claims reflects evaluative divergence (higher evidence thresholds).

6.2 Agreement Analysis

Agreement is measured along four graded levels, allowing precise classification of whether models converge on *structure*, *interpretation*, *assumptions*, or *conclusions*.

6.2.1 Full Agreement

All models:

- reconstruct the same logical structure
- identify similar hidden assumptions
- follow similar interpretive pathways
- reach the same conclusion

This was most common in Round 1 scientific arguments (e.g., 1A/1B Evolution), where low ambiguity constrained interpretive drift. Approximately **40%** of Round 1 evaluations fell into this category.

6.2.2 Partial Agreement

Models agree on the high-level verdict but diverge in:

- hidden assumptions
- premise weighting
- inferential pathways
- interpretations of uncertainty

This was the dominant pattern in Round 2, representing **about 60%** of all cross-domain outputs. The Moderate AGW variant (7A/7B) is a canonical example: all models accept anthropogenic causation but differ in how strongly they evaluate model uncertainty, risk weighting, and probabilistic reasoning.

6.2.3 Structured Disagreement

Models reach different conclusions but maintain:

- clearly traceable reasoning structures
- identifiable assumption differences
- explicit interpretive divergences

Structured disagreement is **more informative** than agreement because it reveals *why* models diverge. It often signals areas where:

- the argument contains ambiguous or underspecified premises
- domain knowledge interacts with safety layers
- models adopt different inferential priorities

For example, in the Second Amendment argument (9A/9B), divergence stems from how models interpret the scope of “shall not be infringed,” not from reasoning failure.

6.2.4 Contradictory Conclusions

Models produce **opposing conclusions** and their reasoning pathways cannot be reconciled.

This occurs when:

• models interpret key terms differently
• normative alignment behavior influences evaluation
• safety layers restrict or redirect analysis

This pattern occurred most often in political or ethical arguments, especially in Voter Fraud (11A/11B). Contradictory conclusions are not “errors”—they signal that the argument interacts sharply with model-specific constraints.

Practical note:
Contradictory outputs are *not* resolved; instead, analysts examine which divergence type produced them (structural, interpretive, bias-driven, or evaluative).

Transition to Section 6.3:
These agreement patterns form the observable outcomes. To understand *why* they arise, we classify divergences into four distinct but sometimes overlapping types.

6.3 Divergence Types

Divergence reflects the underlying causes of model differences. These types are analytically distinct, but they may co-occur, especially in complex or safety-sensitive arguments.

Divergence types frequently overlap—for example, bias-driven divergence and interpretive divergence often co-occur in political arguments, where safety-layer caution interacts with ambiguous phrasing. In such cases, the dominant divergence type is identified by which factor most strongly drives the difference.

6.3.1 Structural Divergence

Differences in the **formal argument map**, including:

• number of premises extracted
• causal vs. correlational interpretation
• inferential branching
• identification of intermediate conclusions

For Example:

• GPT-5.1 → carefully reconstructs full premise chains
• DeepSeek → aggressively compresses multi-step reasoning into a single inference
• Gemini → adds contextual premises (conceptual broadening)
• Grok → preserves only the most central structural elements

Structural divergence was most common in Round 1 philosophy arguments (Arguments 3–5).

6.3.2 Interpretive Divergence

Differences driven by interpretation of ambiguous language or underdetermined context.

Model patterns:

- **Gemini** – broad, context-integrative framing
- **Grok** – minimalistic, text-bound interpretation
- **GPT-5.1** – conservative, low-speculation interpretation
- **DeepSeek** – decisive, compressed interpretive stance

Interpretive divergence was the **most frequent divergence type** in Round 2, especially in Universal Healthcare (10A/10B) and AGW Weak Variant (8A/8B).

See Section 4.2–4.5 for full model reasoning signatures that explain these interpretive tendencies.

6.3.3 Bias-Driven Divergence

Differences caused by:

- safety constraints
- political/ethical caution
- normative alignment heuristics
- refusal or hedging behaviors

Example patterns:

- **Gemini** – softens political claims (11A/11B)
- **GPT-5.1** – moderates sensitive or legal claims (9A/9B)
- **Grok** – avoids extended ethical reasoning (13)
- **DeepSeek** – occasionally exceeds normal confidence bounds

Bias-driven divergence is especially prominent in political, legal, and ethical arguments.

6.3.4 Evaluation Divergence

Differences in *weighing* evidence or premises:

- strength of empirical evidence
- degree of uncertainty tolerated

• normative weighting of moral or legal principles
• credibility assigned to contested claims

Seen frequently in:

• Second Amendment (9A/9B)
• Universal Healthcare (10A/10B)
• AGW Moderate/Weak (7A/7B, 8A/8B)

6.4 Hidden Assumption Sensitivity

Hidden assumptions frequently determine which conclusion a model reaches. Sensitivity describes how reliably a model identifies implicit premises that affect interpretation.

Table 4: Model Sensitivity Notes

Model	Sensitivity	Notes
GPT-5.1	High	Consistent identification of structural & contextual assumptions
Gemini 2.0	High–variable	Adds conceptual assumptions; some overextension
DeepSeek V2	Medium–Selective	Captures major assumptions; omits subtle ones
Grok	Low	Minimalist extraction; focuses on explicit content

When is high assumption sensitivity useful?
• In complex, multi-premise arguments
• In legal or philosophical analysis
• When tracing inference failure

When can it be harmful?
• In arguments requiring strict textual fidelity
• When conceptual expansion introduces drift (common in Gemini)

The assumption behavior described here corresponds directly to the assumption-mapping traits detailed in Section 4.

6.5 Bias Signature Assessment (formerly "Scoring")

Bias Signature Assessment complements the Reasoning Signatures presented in Section 4. While reasoning signatures describe a model's characteristic approach to structured argumentation—its interpretive habits, structural tendencies, and default analytic style—the bias signature isolates the *behavioral tendencies that most directly shape divergence*: caution, decisiveness, neutrality, assumption density, and drift susceptibility.

The Bias Signature Assessment is qualitative rather than numerical. It examines how each model behaves across both rounds and both protocols, focusing on consistent patterns rather than single-run anomalies.

The assessment characterizes each model along five dimensions.
Each dimension is defined below, followed by model-specific justifications grounded in the argument suite.

6.5.1 Caution Level

How hesitant is the model to commit under uncertainty?

- **GPT-5.1 / Gemini** – cautious; require justification
- **Grok** – moderate; commits when structure is clear
- **DeepSeek** – low caution; commits decisively

6.5.2 Decisiveness

How readily does a model pick a conclusion when ambiguity remains?

- **DeepSeek** – highest decisiveness
- **Grok** – moderate
- **GPT-5.1 / Gemini** – deliberately restrained

6.5.3 Neutrality

How free are outputs from evaluative or normative framing?

- **GPT-5.1** – very high neutrality
- **Claude (synthesis)** – high and stable
- **Gemini** – moderate; some normative framing
- **DeepSeek** – varies by domain

6.5.4 Assumption Detection Strength

How effectively does a model identify unstated premises?

• **GPT-5.1** – excellent
• **Gemini** – strong but sometimes overextended
• **DeepSeek** – selective
• **Grok** – minimal

6.5.5 Drift Likelihood

How likely is a model to shift away from the given argument structure?

• **GPT-5.1 / Grok** – very low drift
• **DeepSeek** – medium
• **Gemini** – high (especially in ethical, metaphysical domains)

6.6 Model Clustering and Reasoning Families

Based on divergence patterns, assumption behavior, and Bias Signature dimensions, models cluster into three reasoning families.

6.6.1 Conservative Evaluators

Models: GPT-5.1, Claude (synthesis)
Characteristics:
• methodical
• assumption-rich
• drift-resistant
• cautious with contested claims

These models are best for: legal analysis, philosophical reasoning, complex multi-premise evaluation.

6.6.2 Aggressive Evaluators

Models: DeepSeek V2, Grok
Characteristics:
• decisive
• compressed reasoning
• minimal caution
• lower assumption density

Although DeepSeek provides *deep but compressed analysis* and Grok provides *shallow but precise analysis*, they share an "aggressive commitment" pattern—they commit early and confidently, even under ambiguity.

Best for: exploratory analysis, rapid evaluation, and scenarios requiring decisive outputs under uncertainty.

6.6.3 Hybrid Generalists

Model: Gemini 2.0
Characteristics:
• abstraction-first reasoning
• broad conceptual integration
• variable caution
• high interpretive drift in underdefined domains

Best for: ethical, metaphysical, and contextual reasoning where broad framing is advantageous.

Clustering Methodology Note

Clustering was determined by examining:

• agreement patterns across all 28 model runs
• divergence type frequencies
• assumption detection profiles
• output structures across both protocols
• relative caution and decisiveness rankings

Alternative clusterings (2-model and 4-model schemes) were evaluated but resulted in less coherent behavioral groupings.

Cluster stability was consistent across Round 1 and Round 2: GPT-5.1 and Claude remained tightly coupled in conservative evaluation behavior, DeepSeek and Grok consistently formed the aggressive cluster, and Gemini maintained its outlier "hybrid generalist" pattern. A 2-cluster solution (Conservative vs. Non-conservative) collapsed essential distinctions, while a 4-cluster solution over-partitioned marginal differences. The 3-cluster solution best reflected stable behavioral patterns.

6.7 Framework Application Guidelines

The Cross-Model Comparison Framework is intended to be used operationally, not just descriptively. This subsection provides practical guidelines for analysts applying the framework in real workflows—policy analysis, DoD decision support, corporate strategy, and research environments—when selecting models, interpreting disagreement, and designing ensemble pipelines.

Using Agreement Levels in Practice

When reviewing outputs from multiple models on the same argument:

- **Full Agreement**
  Treat conclusions as structurally robust, but still inspect at least one model's reasoning (typically GPT-5.1) to confirm that key assumptions are explicit. Use this case to **validate the argument** rather than the models.

- **Partial Agreement**
  Use the framework to identify *which* assumptions, evidential weights, or interpretive choices differ. These are your **leverage points** for refining the argument: clarify ambiguous premises, tighten definitions, or specify evidential thresholds.

- **Structured Disagreement**
  Do not try to "average" the models. Instead, explicitly document each coherent pathway (e.g., "legal-textual reading vs. living-constitution reading" in 9A/9B) and treat them as **competing, defensible analyses**. This is often where the most valuable insight lies.

- **Contradictory Conclusions**
  Treat this as a diagnostic signal, not a failure. Use the divergence types (Section 6.3) to determine whether the conflict is driven primarily by structure, interpretation, bias/alignment, or evaluation—and then decide whether the underlying argument needs to be reframed, decomposed, or supported with additional data.

Responding to Divergence Types

When the framework flags a specific divergence type:

- **Structural Divergence**
  Reconstruct the argument formally (often using GPT-5.1) and standardize the premise list. If models are not even analyzing the same structure, **fix structure first** before comparing conclusions.

- **Interpretive Divergence**
  Clarify key terms, scope conditions, and domain boundaries in the prompt. If you want multiple legitimate readings, keep both pathways explicit; if not, tighten definitions to narrow the interpretive space.

- **Bias-Driven Divergence**
  Identify which model's safety or alignment behavior is distorting the analysis (e.g., over-hedging vs. over-commitment). In high-stakes contexts, pair a

conservative evaluator (GPT-5.1) with an **aggressive evaluator** (DeepSeek or Grok) and use the contrast to separate alignment artifacts from genuine substantive disagreement.

- **Evaluation Divergence**
  When models differ mainly in how they weigh evidence or normative principles, treat this as a cue to **make weights explicit**: specify which metrics, time horizons, or moral/legal priorities should dominate the assessment, rather than leaving them implicit.

Choosing Models and Ensembles

The reasoning families in Section 6.6 support simple decision rules:

- Use **Conservative Evaluators** (GPT-5.1, Claude) when:
  - auditability, neutrality, and structural fidelity are primary
  - arguments are complex, multi-premise, or high-stakes (legal, strategic)
- Use **Aggressive Evaluators** (DeepSeek, Grok) when:
  - you need fast, decisive reads or first-pass triage
  - you want to expose overconfident readings for later stress-testing
- Use **Hybrid Generalists** (Gemini) when:
  - broad context and conceptual richness are valuable (ethics, metaphysics, complex policy trade-offs)
  - you are exploring alternative framings rather than locking in a verdict

In ensemble use, a common pattern is:

1. **Screen/Triage** with Grok or DeepSeek.
2. **Structure and Audit** with GPT-5.1.
3. **Contextual Expansion** with Gemini (optional, domain-dependent).
4. **Synthesis** with Claude to document convergence, divergence, and assumption clusters.

Analyst Checklist

When applying the framework to a new argument:

- Run at least two models from different reasoning families.
- Classify the outcome using the **agreement levels** (6.2).
- Identify the dominant **divergence type(s)** (6.3).
- Inspect hidden assumptions, starting with GPT-5.1 and Gemini.

- Decide whether the argument needs **reframing**, **clarification**, or **additional data**.
- Document which model behaviors are likely **alignment artifacts** versus genuine analytic differences.
- Where stakes are high, use Claude (or an equivalent synthesizer) to produce a written synthesis before acting.

These guidelines translate the comparison framework into concrete decision rules, helping practitioners move from "models disagree" to **"here is why they disagree, what that tells us about the argument, and how we should respond."**

6.8 Summary

The Cross-Model Comparison Framework transforms raw LMM disagreement into **actionable analytic signal**. By distinguishing structural, interpretive, bias-driven, and evaluation-based divergences—and by assessing assumption sensitivity and clustering—this framework supports:

• cross-domain synthesis
• reasoning signature identification
• bias-aware model selection
• ensemble pipeline design
• rigorous case study analysis

Limitations:
Divergence magnitude is not yet quantified, temporal stability is not directly tested, and domain-specific patterns require further study. These limitations guide future work in Sections 10 and 11.

This framework provides the analytic foundation for the case studies in Section VII and the integrated synthesis in Section VIII.

VII. Case Study Results

7.1 Overview

This section applies the comparison framework (Section VI) to the full argument suite, summarizing how models behaved across political–legal, economic, ethical, metaphysical, and epistemic domains. Rather than reproducing all raw outputs, it highlights **systematic divergence patterns** observed across the model–protocol combinations and then zooms into three detailed case studies.

Sections **7.2–7.5** synthesize domain-level patterns. Section **7.6** then presents three representative case studies—Anthropogenic Global Warming (Strong/Moderate/Weak variants), the Second Amendment argument, and the Problem of Evil—that illustrate how reasoning signatures and divergence types manifest in concrete evaluations.

7.2 Political and Legal Arguments

Political and constitutional arguments—especially the **Second Amendment argument (9A/9B)** and the **Voter Fraud argument (11A/11B)**—produced the most prominent combination of **interpretive** and **bias-driven** divergence.

7.2.1 Risk tolerance and safety activation

Across both arguments, the models split along caution lines:

- **GPT-5.1** adopted a restrained, legal-textual style. It frequently used formulations such as “this interpretation suggests” or “a plausible reading is” rather than “this proves,” signaling high caution and neutrality.
- **Gemini 2.0** broadened the frame to incorporate democratic norms, institutional trust, and social stability, often adding explicit reminders about limited evidence in contested political contexts.
- **DeepSeek V2** tended toward confident, compressed verdicts, resolving ambiguities decisively rather than flagging them.
- **Grok** produced brief, surface-level analyses, often stating a clear conclusion with minimal legal elaboration.

In both 9A/9B and 11A/11B, **safety layers** were most visible in Gemini (hedging and emphasis on uncertainty) and GPT-5.1 (careful distancing language), while DeepSeek and Grok showed relatively fewer alignment-driven disclaimers and more direct answers.

7.2.2 Competing interpretive frames

The Second Amendment argument exposed three distinct interpretive frames:

- **Textualist** – treating “shall not be infringed” as near-absolute; favored by DeepSeek in several runs and sometimes by Grok in compressed form.
- **Historical-contextual** – weighing militia context, historical practice, and precedent; most consistently used by GPT-5.1.
- **Policy-balancing** – treating the text as one input alongside public safety and democratic norms; more common in Gemini.

These frames often led to **structured disagreement**: models reached different conclusions but did so via clearly traceable, defensible reasoning pathways rather than incoherent drift.

7.2.3 Voter fraud and evidential standards

In the Voter Fraud argument (11A/11B), divergence was dominated by **evaluation** and **bias-driven** types:

- GPT-5.1 and Gemini both treated anecdotal reports as weak evidence, emphasizing the need for systematic data.
- DeepSeek more readily treated repeated anecdotal reports as suggestive of underlying patterns, while still acknowledging uncertainty.
- Grok offered concise, low-detail assessments, usually siding with the more conservative evidential stance but without extensive justification.

Overall, political and legal arguments showed that **alignment behavior and interpretive frame selection** play a larger role than raw logical competence in shaping model conclusions.

7.3 Economic Arguments

Economic reasoning—particularly the **Universal Healthcare Efficiency argument (10A/10B)**—highlighted differences in **evaluation divergence** and **definition drift** around key terms such as “efficiency.”

7.3.1 Efficiency as an economic vs. ethical construct

The models implicitly defined “efficiency” in different ways:

- **GPT-5.1** treated efficiency primarily as an economic concept: cost per outcome, resource allocation, and measurable system-level performance. This led it to weigh comparative cost and health-outcome data heavily and to flag any missing or speculative numbers.
- **DeepSeek V2** adopted a similar economic focus but with more aggressive inferential compression—filling in plausible causal links (e.g., “universal coverage reduces uncompensated care, which in turn lowers system-wide costs”) without always fully unpacking them.
- **Gemini 2.0** expanded ‘efficiency’ to include social welfare, equity, and long-term quality-of-life effects. This broader framing made it more favorable to arguments that trade short-term costs for long-term social gains, and sometimes treated these considerations as equally valid inputs alongside economic metrics.

- **Grok** gave short, structurally correct analyses that largely mirrored the most straightforward economic reading without much exploration of alternative definitions.

These definitional differences often resulted in **partial agreement**: models accepted similar empirical premises but ranked policy options differently.

7.3.2 Evidence weighting and uncertainty

Across 10A/10B, GPT-5.1 and DeepSeek both demanded at least a minimally specified causal chain from policy to outcome, but:

- GPT-5.1 explicitly highlighted uncertainties and data gaps.
- DeepSeek was more willing to treat plausible but under-specified links as provisionally acceptable.
- Gemini introduced additional normative assumptions (e.g., prioritizing universality or fairness) and sometimes treated them as co-equal with economic metrics.
- Grok stayed close to the given text, rarely elaborating on missing data.

Economic arguments thus demonstrate how **different evidence-weighting heuristics and conceptual framings** produce evaluation divergence even when models appear to agree on the surface.

7.4 Ethical and Moral Arguments

Ethical and theological arguments—most notably **the Problem of Evil (Argument 13)**—exposed differences in **assumption density**, **philosophical depth**, and **normative framing**.

7.4.1 Depth vs. discipline in metaphysical reasoning

A more precise characterization of model behavior in this argument is as follows:

- **Gemini 2.0** showed the greatest metaphysical reach, exploring free will theodicies, soul-making accounts, and modal possibilities well beyond the explicit text.
- **GPT-5.1** maintained a disciplined focus on the core logical structure: the relationship between omnipotence, omniscience, omnibenevolence, and "unnecessary suffering," carefully distinguishing logical from evidential formulations.

- **DeepSeek V2** gave decisive, high-level verdicts on whether the argument succeeded but tended to compress or simplify the range of possible theodicies.
- **Grok** provided concise, mostly textbook-style summaries of the problem with minimal elaboration, consistent with its overall surface-analytic signature.

This pattern aligns with the reasoning signatures in Section IV: Gemini as context-rich and expansive; GPT-5.1 as structurally rigorous; DeepSeek as decisive; Grok as concise and low-depth.

7.4.2 Hidden assumptions and divergent verdicts

Key hidden assumptions that models handled differently included:

- whether "unnecessary suffering" is a coherent and applicable category
- whether divine omnibenevolence entails maximizing happiness or merely preventing gratuitous harm
- whether the existence of natural laws or free will counts as a morally relevant constraint

GPT-5.1 and Gemini explicitly identified more of these assumptions; DeepSeek tended to fix one or two as given; Grok often left them implicit. As a result, models sometimes agreed on the *validity* of the logical form but diverged on *soundness*, depending on which assumptions they were willing to grant.

7.4.3 Alignment behavior

Unlike the political arguments, ethical and theological content did **not** strongly trigger safety filters. None of the models refused to engage, and disclaimers were minimal. This suggests that, for classical philosophical topics, divergence is driven far more by **reasoning style and assumption handling** than by alignment constraints.

7.5 Metaphysical and Epistemic Arguments

Metaphysical and epistemic arguments—especially **the Cartesian Certainty argument (12A/12B)**—provided a clean test of each model's capacity for structured logical reasoning under relatively low political or ethical load.

7.5.1 Convergence on structure

All four models correctly reconstructed the core Cartesian structure:

1. If I can doubt X, then X is not certain.
2. I can doubt the existence of my body.
3. I cannot doubt that I am thinking.
4. Therefore, my existence as a thinking thing is more certain than the existence of my body.

Structural divergence here was minimal. The primary differences emerged in:

- how explicitly each model separated *validity* from *soundness*
- whether they treated the premises as psychological, logical, or metaphysical claims.

7.5.2 Rigor vs. abstraction

- **GPT-5.1** excelled at mapping the argument into clear premise–conclusion form and evaluating possible equivocations (e.g., different senses of "certainty").
- **DeepSeek V2** produced compressed but generally accurate treatments, sometimes skipping intermediate analytic steps.
- **Gemini 2.0** expanded the discussion into broader epistemological territory—fallibilism, skepticism, and modern philosophy of mind—occasionally drifting away from the narrow Cartesian structure.
- **Grok** offered compact analyses that captured the basic logic but with limited exploration of alternative interpretations.

7.5.3 Protocol and drift patterns

Differences between the 8-step and 14-step protocols were most visible in Gemini, which used the additional steps to explore more alternative framings, and in GPT-5.1, which used them to systematize assumption mapping. Drift remained low for GPT-5.1, DeepSeek, and Grok; Gemini showed moderate drift in a subset of 14-step runs, consistent with its abstraction-heavy signature.

7.6 Detailed Case Study Analysis

This subsection presents three representative case studies that bring together the reasoning signatures (Section IV), divergence taxonomy (Section VI), and argument set (Section V).

7.6.1 Case Study 1 – Anthropogenic Global Warming (Strong, Moderate, Weak Variants)

Argument design.
The AGW series presents the same core claim—human emissions are the primary driver of observed global warming—in three strengths:

- **Strong** – categorical language, high confidence, minimal uncertainty.
- **Moderate** – probabilistic language, explicit but bounded uncertainty.
- **Weak** – hedged language, substantial emphasis on model limitations and unknowns.

Each variant was evaluated under both protocols (8-step and 14-step), generating a rich comparison set.

Model behavior.
Across variants:

- **GPT-5.1** maintained a cautious but generally affirmative stance, increasing its explicit discussion of uncertainty as the argument weakened, but rarely flipping its overall verdict.
- **DeepSeek V2** remained confident across all three, treating the Strong and Moderate forms similarly and only softening slightly on the Weak form.
- **Gemini 2.0** shifted its tone noticeably: strongly affirmative on the Strong variant, more balanced and caveated on the Moderate variant, and significantly more equivocal on the Weak variant, highlighting model uncertainty and data limitations.
- **Grok** produced concise verdicts that tracked the surface strength of the argument, offering less explicit reasoning about uncertainty.

Divergence patterns.

- In the **Strong** form, most divergence was **interpretive** (how strongly to read the empirical consensus) and **evaluation-based** (weighting of long-term risk).
- In the **Moderate** form, **partial agreement** dominated: all models accepted the basic causal story but differed in how they expressed caution.
- In the **Weak** form, both **evaluation** and **bias-driven** divergence appeared, especially in how models responded to explicit invitations to doubt.

This case study illustrates how **argument strength interacts with model caution, evidential thresholds, and alignment behavior**, generating a graded pattern of divergence across the same underlying claim.

7.6.2 Case Study 2 – Second Amendment (9A/9B)

Argument design.
The Second Amendment argument presents a strict constitutional claim that all gun control laws violate the text "shall not be infringed." The 8-step and 14-step protocols force models to:

- reconstruct the explicit premises,
- surface hidden assumptions (e.g., about original intent, public safety, and judicial interpretation),
- and issue a validity/soundness verdict.

Model behavior.

- **GPT-5.1** treated the argument as a legal-textual claim constrained by judicial precedent, frequently noting that real-world constitutional interpretation relies on case law and historical practice, not text alone.
- **DeepSeek V2** was more willing to grant the argument's textual premises and explore the implications of a strict reading, while still noting tension with practical governance.
- **Gemini 2.0** reframed the problem as a balance between rights and public safety, explicitly integrating policy and ethical considerations into its assessment.
- **Grok** provided short, direct responses that typically downplayed historical nuance and focused on the most literal reading.

Divergence patterns.

The case exposed:

- **Interpretive divergence** between textualist, historical, and policy-balanced frames.
- **Bias-driven divergence** where alignment steered models away from extreme policy recommendations.
- **Evaluation divergence** in weighing individual rights vs. collective safety.

The Second Amendment case demonstrates how **political context sharpens the impact of alignment and interpretive style**, even when logical structure is straightforward.

7.6.3 Case Study 3 – Problem of Evil (13)

Argument design.
The Problem of Evil argument uses a classical formulation: if an omnipotent, omniscient, omnibenevolent God exists, there should be no unnecessary suffering; since such suffering exists, either God does not exist or lacks at least one of these attributes. This argument was evaluated with the 8-step protocol only.

Model behavior.

- **GPT-5.1** cleanly reconstructed the logical form and distinguished between logical and evidential versions of the argument, analyzing where each premise could be challenged.
- **DeepSeek V2** issued decisive assessments of whether the argument "succeeds" but often compressed the space of potential theodicies.
- **Gemini 2.0** explored a wide range of theological and philosophical responses—free will, soul-making, skeptical theism—sometimes at the cost of drifting away from the narrow structure of the original argument.
- **Grok** summarized the core issue efficiently but with limited exploration of alternative responses.

Divergence patterns.

- **Structural divergence** was low: all models recognized the same core premises.
- **Interpretive divergence**—driven by differences in hidden assumptions—was high: models differed in how they interpreted "unnecessary suffering," divine attributes, and the relevance of natural law and free will.
- **Metaphysical depth** varied strongly, with Gemini taking the broadest, most exploratory approach and GPT-5.1 maintaining the tightest focus on logical structure.

This case demonstrates how **philosophical arguments magnify differences in assumption sensitivity and metaphysical ambition**, rather than alignment behavior per se.

7.7 Summary

Across these domains and case studies, several patterns emerge:

- **Divergences are structured, not random.** They consistently reflect each model's reasoning signature and bias profile rather than noise.

- **Political and legal arguments** are most affected by safety and alignment, whereas **ethical and metaphysical arguments** are shaped more by assumption density and philosophical style.
- **Economic arguments** expose differences in how models balance empirical data, causal structure, and normative commitments.
- **Metaphysical and epistemic arguments** reveal core logical capabilities, with relatively little interference from alignment layers.
- **Structured disagreement**—especially when models present multiple defensible pathways—is often the most analytically valuable outcome, revealing where an argument's premises, definitions, or assumptions need clarification.

These results validate the cross-model framework introduced in Sections V and VI and provide the empirical foundation for the synthesis, ensemble strategies, and applied guidance developed in **Section VIII**.

VIII. Synthesis

8.1 Integrated Cross-Model Patterns

Cross-model synthesis reveals five consistent, domain-spanning patterns:

Pattern 1: Structural Stability Across Protocols

Models maintained stable structural tendencies across both the 8-step and 14-step protocols.
Longer protocols did **not** change conclusions—they **amplified** characteristic reasoning signatures (e.g., GPT-5.1's structural mapping, Gemini's contextual expansion).

Pattern 2: Domain-Divergence Alignment

Divergence types correlate strongly with argument domain:

- **Political & Legal** → interpretive + bias-driven divergence
- **Economic & Policy** → evaluation divergence
- **Ethical & Metaphysical** → assumption-driven interpretive divergence
- **Scientific & Empirical** → structural agreement with evidential divergence

These mappings were consistent across all models and protocols.

Pattern 3: Agreement Tracks Argument Strength

Arguments with tight structure and minimal ambiguity (e.g., Evolution 1A/1B) yield high agreement.
Arguments with broad conceptual premises (e.g., Problem of Evil; tariffs; voter fraud) produce predictable divergence.

Pattern 4: Protocol Amplification Rather Than Transformation

8-step and 14-step protocols do not produce different "types" of reasoning—they reveal **more** of each model's characteristic tendencies.
The 14-step protocol especially magnifies assumption density and interpretive spread.

Pattern 5: Assumption Density Predicts Divergence

Models with high assumption density (GPT-5.1, Gemini) diverge from low-density models (Grok, to some extent DeepSeek) when arguments hinge on unstated premises.

These patterns serve as the analytic foundation for the **ensemble weighting system** presented in the next section—transforming predictable divergence into a strategic decision-support asset.

8.2 The Ensemble Reasoning Framework

8.2.1 Why Ensembles Are Required

Across all domains, no single model is consistently superior. Instead, each exhibits strengths aligned to its reasoning signature:

- **GPT-5.1:** structure, neutrality, assumption detection
- **DeepSeek:** decisiveness, compressed inference, clear commitments
- **Grok:** fast minimalism, stable baseline reasoning
- **Gemini:** broad contextualization, high abstraction
- **Claude:** meta-synthesis neutrality and structural fidelity

These differences are **not noise**—they provide complementary analytic perspectives.
While ensemble reasoning requires multiple model runs, **the reliability gains justify the additional resource cost** in high-stakes contexts.

8.2.2 Core Principle

The ensemble method integrates:

1. **Complementarity** (leveraging strengths)

2. **Redundancy** (detecting weaknesses)
3. **Weighted aggregation** (prioritizing domain-relevant models)
4. **Cross-model synthesis** (Claude)

8.3 Weighted Model Profiles by Domain

These weights are qualitative priorities—not precise numeric formulas.
They reflect relative influence each model should have in synthesis for that domain.

8.3.1 Empirical / Scientific Arguments

Recommended Weights:

- GPT-5.1 - **0.40**
- DeepSeek - **0.30**
- Gemini - **0.20**
- Grok - **0.10**

Rationale:
Scientific arguments demand structural rigor, evidential discipline, and conservative inference.
DeepSeek contributes productive decisiveness; Gemini adds contextual nuance.

8.3.2 Normative / Ethical Arguments

Recommended Weights:

- Gemini - **0.40**
- GPT-5.1 - **0.30**
- DeepSeek - **0.20**
- Grok - **0.10**

Rationale:
Ethical reasoning requires conceptual breadth, high assumption awareness, and interpretive nuance—Gemini excels here.

8.3.3 Political / Legal Arguments

Recommended Weights:

- GPT-5.1 - **0.40**
- Gemini - **0.30**
- DeepSeek - **0.20**
- Grok - **0.10**

Rationale:
Political/legal arguments are highly alignment-sensitive.
GPT-5.1 offers neutrality and precision; Gemini provides framing breadth; DeepSeek supplies clarity of commitment.

8.3.4 Economic / Policy Arguments

Recommended Weights:

- GPT-5.1 - **0.40**
- Gemini - **0.30**
- DeepSeek - **0.20**
- Grok - **0.10**

Rationale:
Economic arguments depend on evaluation divergence (evidence weighing).
GPT-5.1's structure and Gemini's contextual breadth complement DeepSeek's decisive causal interpretation.

8.3.5 Metaphysical / Philosophical Arguments

Recommended Weights:

- Gemini - **0.40**
- GPT-5.1 - **0.30**
- DeepSeek - **0.20**
- Grok - **0.10**

Rationale:
Metaphysical arguments expand under interpretive latitude.
Gemini excels at conceptual abstraction; GPT-5.1 anchors structure; Grok adds stability by constraining runaway drift.

8.3.6 High-Risk / Safety-Sensitive Domains

Recommended Weights:

- GPT-5.1 - **0.50**
- Claude - **0.30**
- Gemini - **0.20**
- DeepSeek - **excluded / minimal**
- Grok - **excluded / minimal**

Rationale:
High-risk scenarios require neutrality, predictability, transparency, and minimal alignment drift.
Claude acts as a second neutralizing lens. DeepSeek and Grok are minimized due to high variance and lower caution.

Table 5: Recommended Model Weightings by Domain

Domain	GPT-5.1	DeepSeek	Grok	Gemini	Claude
Empirical/Scientific	0.40	0.30	0.10	0.20	—
Normative/Ethical	0.30	0.20	0.10	0.40	—
Political/Legal	0.40	0.20	0.10	0.30	—
Economic/Policy	0.40	0.20	0.10	0.30	—
Metaphysical/Philosophical	0.30	0.20	0.10	0.40	—
High-Risk/Safety	0.50	—	—	0.20	0.30

8.3.7 Applying the Weights in Practice

These weights can be applied in three complementary ways:

1. Attention Allocation

Higher-weight models deserve proportionally deeper reading and analysis.
Example: In legal arguments, GPT-5.1 and Gemini get full review; Grok provides a minimalist check.

2. Verdict Aggregation

In cases of disagreement:

- Higher-weight models carry greater interpretive authority.
- Lower-weight outputs are used for tension identification rather than final judgment.

3. Disagreement Triage

When a low-weight model sharply disagrees with high-weight models, this often signals:

- hidden assumptions
- domain ambiguity
- competing definitions
- safety-layer influence

For hybrid arguments (e.g., political-economic), analysts may blend domain profiles or run both and compare their implications.

8.4 Ensemble Workflow Recommendations

A four-stage pipeline operationalizes ensemble reasoning:

Stage 1 - Rapid Triage (Grok + DeepSeek)

Fast read on structure, salience, and immediate contradictions.

Decision Point:
If both models fully agree and stakes are low → analyst may stop here.

Stage 2 - Structural Mapping (GPT-5.1)

Authoritative reconstruction of premises, assumptions, and inferential pathways.

Stage 3 - Contextual Expansion (Gemini)

Adds conceptual breadth, alternative framings, and non-obvious interpretive angles.

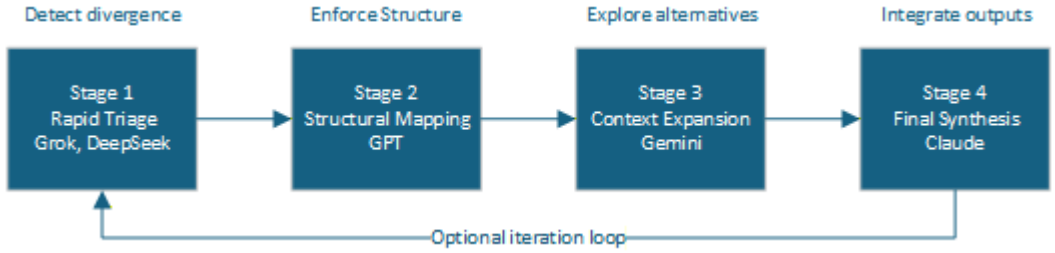
Stage 4 - Final Synthesis (Claude)

Claude integrates multi-model outputs into a structured, neutral synthesis:

- convergence map
- divergence taxonomy
- assumption clusters
- prioritized final recommendation

Iteration Note:
If Claude identifies structural disagreement between GPT-5.1 and Gemini, analysts should loop back to Stage 2.

Figure 4: Ensemble Workflow Diagram



This workflow operationalizes the ensemble method described in Sections 8.2–8.4. Stage 1 identifies potential divergence; Stage 2 establishes the formal argument map; Stage 3 introduces interpretive breadth and explores alternative framings; Stage 4 consolidates findings into a reproducible synthesis. The pipeline reduces drift, captures hidden assumptions, and enhances reliability across domains.

8.5 Strengths and Limitations of Ensemble Reasoning

Strengths

- Cross-model redundancy increases reliability
- Divergence reveals argument weaknesses
- Consistent correction of individual-model biases
- Domain weighting tailors analysis to context
- Transparent, auditable decision pathways

Limitations

- Requires access to multiple proprietary models (cost, API constraints)
- Some divergences reflect genuine argument ambiguity, not model error
- Interpretation of disagreement still requires human judgment
- Ensemble workflow is slower than single-model queries
- Not ideal for low-stakes or time-critical tasks
- Weightings require periodic recalibration as models evolve

8.6 Summary

The ensemble framework transforms model diversity into a structured analytic asset.

By combining structural rigor (GPT-5.1), decisive inference (DeepSeek), minimalism (Grok), contextual breadth (Gemini), and meta-synthesis (Claude), analysts achieve clarity and robustness unavailable from any single model.

Section IX applies this ensemble architecture to real operational domains—policy, defense, legal analysis, corporate decision support, and scientific reasoning—demonstrating practical, high-impact usage patterns.

IX. Applications

9.1 Overview

This section moves from methodology and analysis to practice. It demonstrates how structured model divergence, weighted ensemble reasoning, and reasoning signatures can be operationalized across real-world domains.

Applications span seven areas:

1. Scientific and technical assessment
2. Policy and legal analysis
3. Economic and financial modeling
4. Ethical and normative reasoning
5. High-risk and safety-critical domains
6. Corporate and strategic decision support
7. Intelligence, defense, and operational planning
8. Education and research (added for completeness with 9.1)

Across all domains, a core pattern holds: **model diversity, when systematically measured and weighted, produces a more comprehensive, transparent, and defensible analytic output** than any single model operating alone.

9.2 Scientific and Technical Assessment

Use Case: Causal Attribution in Complex Systems
Scientific arguments (e.g., climate causality, material failure analysis, epidemiological spread) often require high structural rigor and careful handling of uncertainty.

Ensemble Advantage:

- **GPT-5.1** provides precise causal-chain mapping and disciplined assumption audits.
- **DeepSeek** supplies fast, confident causal interpretations for hypothesis triage.
- **Grok** offers baseline, minimal-interpretation checks with low drift.

- **Gemini** integrates contextual complexity (feedback loops, time horizons, confounders).

The AGW variants analyzed in Section 7.6.1 show how this combination stabilizes causal inference across evidence strengths.

Table 6: Multi-Model Evaluation of a Causal Claim

Claim: *"Increased CO_2 emissions cause observed global temperature rise."*

Model	**Structural Analysis**	**Assumptions Identified**	**Conclusion**
GPT-5.1	Maps formal causal chain; evaluates strength of each inferential step	Climate-model reliability; robustness of empirical time series	Strong causal support with uncertainty management
DeepSeek	Compresses causal structure into direct linkage	Assumes consensus science; omits minor confounders	Decisive: CO_2 is primary cause
Grok	Surface-level correlation check	Minimal assumptions; avoids extrapolation	Correlation strong; causal claim plausible
Gemini	Expands to feedback loops, aerosols, long-term forcing	Identifies broad conceptual assumptions	Causal but multi-factor; emphasizes complexity

Result:
A multi-model scientific assessment that combines rigor, speed, minimalism, and contextual integration—avoiding overconfidence or oversimplification.

9.3 Policy and Legal Analysis

Use Case: Constitutional Interpretation (Second Amendment Cases 9A/9B)
Legal arguments often generate **interpretive divergence** (see Section 7.2.2). Constitutional interpretation frequently requires balancing textual, historical, structural, and policy-based reasoning. Ensemble analysis produces a *layered* evaluation that reflects each legitimate interpretive pathway rather than collapsing into a single legal philosophy.

Ensemble Advantage

- **GPT-5.1** → rigorous mapping of statutory language, structure, and precedent
- **DeepSeek** → decisive interpretation (textualist or originalist pathways)
- **Grok** → minimal-bias reading anchored strictly in explicit text
- **Gemini** → incorporates historical evolution, societal context, and policy implications
- **Claude** → neutral, meta-level synthesis across all interpretive frames (Section 4.6)

Table 7: Recommended Model Weightings by Domain

Domain	GPT-5.1	DeepSeek	Grok	Gemini	Claude
Empirical/Scientific	0.40	0.30	0.10	0.20	—
Normative/Ethical	0.30	0.20	0.10	0.40	—
Political/Legal	0.40	0.20	0.10	0.30	—
Economic/Policy	0.40	0.20	0.10	0.30	—
Metaphysical/Phil.	0.30	0.20	0.10	0.40	—
High-Risk/Safety	0.50	—	—	0.20	0.30

Note. These weightings prioritize the models whose reasoning strengths align most closely with political and legal domains—particularly GPT-5.1's structural rigor, DeepSeek's decisive textualism, Gemini's contextual expansion, Grok's textual minimalism, and Claude's neutral synthesis.

Workflow

1. **Grok** identifies the baseline textual meaning of the provision (e.g., "the right to keep and bear arms").
2. **GPT-5.1** maps the argument's legal structure, relevant precedent, and inferential logic.
3. **Gemini** adds historical context, societal evolution, policy impacts, and interpretive alternatives.
4. **DeepSeek** delivers decisive interpretive conclusions (textualist or originalist).

5. **Claude** synthesizes the above into a transparent, multi-framework legal analysis.

Result

A **layered constitutional analysis** that makes explicit where legal interpretations diverge, why they diverge, and which assumptions and frameworks drive each interpretive pathway. This approach strengthens judicial reasoning, legislative drafting, regulatory design, and constitutional policy evaluation.

9.4 Economic and Financial Modeling

Use Case: Monetary Policy - "Should the federal funds rate be raised?"
As Section 7.3 shows, economic arguments primarily diverge through **evaluation divergence** (evidence weighting, premise reliability, economic model selection).

Model Contributions:

- **GPT-5.1** → structured comparison of inflation, unemployment, and output gaps
- **DeepSeek** → rapid interpretation of economic indicators
- **Grok** → minimalist baseline to filter noise
- **Gemini** → broader contextualization (global markets, demographics)

Example Outcome:
If DeepSeek recommends an immediate rate hike based on inflation metrics while Gemini highlights labor market slack and geopolitical constraints, the divergence exposes the policy trade space—rather than collapsing into a premature binary recommendation.

Result:
A calibrated policy judgment that incorporates structural, empirical, contextual, and minimal-baseline viewpoints.

9.5 Ethical and Normative Reasoning

Use Case: Autonomous System Risk Decisions
Ethical arguments diverge primarily through **assumption sensitivity** and philosophical framing (7.4.2).

Ensemble Advantage:

- **GPT-5.1** → deontological constraints
- **Gemini** → consequentialist and human-centered impacts

- **DeepSeek** → operational clarity and action orientation
- **Grok** → limits speculative drift and keeps reasoning grounded

Concrete Illustration:
Consider autonomous-vehicle triage scenarios—such as choosing between prioritizing passenger safety or minimizing pedestrian harm. GPT-5.1 foregrounded rule-based safety obligations; Gemini broadened the frame to include equity and long-term societal effects; DeepSeek emphasized feasibility and operational outcomes; Grok anchored the reasoning in immediate, explicit constraints.

Result:
Ethical reasoning that reflects multiple ethical frameworks—not merely the default alignment tendencies of a single model.

9.6 High-Risk and Safety-Critical Domains

Use Case: Nuclear, medical, aviation, and defense safety analysis
In these domains, model overconfidence or hallucination poses unacceptable risk.

Ensemble Strategy (from Section 8.3.6):

- **0.50 GPT-5.1** - primary evaluator (maximal caution + structural rigor)
- **0.30 Claude 3.5** - mandatory synthesis due to exceptional neutrality
- **0.20 Gemini** - contextual expansion
- **DeepSeek and Grok excluded** - elimination of low-caution or shallow modes

Rationale:
GPT-5.1 reduces false positives.
Claude prevents interpretive drift and ensures precision.
Gemini captures overlooked contextual constraints.

Result:
A fail-safe epistemology suitable for medical-device approval, weapons safety protocols, and high-stakes risk assessments.

9.7 Corporate and Strategic Decision Support

Use Case: "Should the company enter Market X?"
Strategic arguments draw from economics, policy risk, competitive context, and technology forecasts.

Model Contributions:

- **GPT-5.1** → operational risk + structural evaluation
- **DeepSeek** → aggressive forecasting (useful for triage)
- **Grok** → baseline feasibility evaluation
- **Gemini** → contextual scanning (demographics, geopolitics, regulatory landscape)

Concrete Example:
For a semiconductor firm evaluating expansion into Southeast Asia, GPT-5.1 highlights regulatory volatility; Gemini identifies demographic and supply-chain factors; DeepSeek identifies near-term opportunity windows; Grok constrains interpretations to what the explicit data supports.

Result:
A balanced strategic recommendation that prevents both tunnel vision (single-model bias) and overextension (contextual drift).

9.8 Intelligence, Defense, and Operational Planning

Use Case: "Is adversary X likely to take action Y within timeframe Z?"
These tasks combine uncertainty, adversarial deception, incomplete information, and geopolitical dynamics.

Model Roles:

- **GPT-5.1** → structured hypothesis evaluation
- **DeepSeek** → decisive pattern recognition
- **Grok** → low-noise factual anchoring
- **Gemini** → geopolitical and historical integration

Analytic Benefit:
Divergence patterns identify:

- Which assumptions are driving different forecasts
- Where evidence sensitivity is highest
- Which conclusions are robust across interpretive modes

Result:
A more resilient intelligence estimate that avoids single-model overconfidence.

9.9 Education, Research, and Pedagogy

Though briefly mentioned in 9.1, it merits explicit application.

Use Case: Teaching critical reasoning with LMMs
Students or researchers can compare structured reasoning across models to learn:

- how arguments fail (Section 2.3)
- how hidden assumptions alter conclusions
- how model drift alters interpretation
- how domain affects divergence

Result:

An instructional tool for logic, epistemology, AI literacy, and interdisciplinary reasoning.

9.10 Summary

Across all domains, ensemble reasoning provides:

1. **Scientific/Technical**: Stable causal inference through rigorous–decisive–minimal–contextual integration.
2. **Policy/Legal**: Multi-framework interpretations aligned to judicial reasoning.
3. **Economic/Financial**: Trade-space exposure rather than single-point judgments.
4. **Ethical/Normative**: Multi-framework philosophical evaluation.
5. **High-Risk/Safety**: Fail-safe epistemology with strict model weighting.
6. **Corporate/Strategic**: Balanced evaluation of uncertainty, opportunity, and risk.
7. **Intelligence/Defense**: Robust analytic estimates under adversarial uncertainty.
8. **Education/Research**: Direct comparison of structured reasoning across models.

Taken together, these applications show that **divergence is not noise**—it is an asset.

When systematically measured, weighted, and synthesized, model divergence becomes a **strategic tool for high-stakes reasoning**, not a liability.

Section X addresses the constraints, limitations, and operational considerations that must guide ensemble deployment in real-world environments.

X. Constraints and Considerations

10.1 Overview

While ensemble reasoning and cross-model analysis substantially improve transparency, depth, and reliability, these methods introduce their own constraints. This section outlines the epistemic, operational, methodological, and safety-related considerations that practitioners must account for when applying this framework in policy, scientific, legal, corporate, or defense environments.

The goal is not to diminish the utility of multi-model reasoning, but to set realistic expectations and define the boundaries within which the framework performs reliably.

10.2 Epistemic Constraints

LMMs are not authoritative sources; they are pattern-based inference engines with probabilistic outputs. Even when combined in an ensemble, they share several inherent epistemic limitations:

10.2.1 Absence of Ground Truth Access

Models cannot:

- retrieve real-time data
- validate facts outside their training
- confirm the accuracy of their own assumptions

Ensemble reasoning reduces overconfidence but does not eliminate uncertainty.

10.2.2 Sensitivity to Argument Framing

Small changes in:

- premise wording
- structure
- scope
- domain terminology

...can alter how models interpret the argument. This affects:

- assumption density
- interpretive drift
- divergence patterns

Practitioners must maintain strict prompt standardization.

10.2.3 Lack of Domain Calibration

Models lack:

- explicit calibration against empirical datasets
- domain-specific error rates
- verifiable probabilistic confidence

Thus, ensemble conclusions should be treated as *structured heuristic assessments*, not statistical certainty.

10.3 Methodological Constraints

The ensemble pipeline depends on strong process discipline.

10.3.1 Protocol Dependency

The 8-step and 14-step protocols constrain drift and enforce structure, but:

- deviations from protocol
- inconsistent session headers
- unstructured follow-up prompts

...can break the comparability across models.

10.3.2 Reproducibility Challenges

Web interfaces introduce:

- slight nondeterminism
- contextual carryover
- token-parsing differences
- safety-trigger variability

This limits strict reproducibility, though it mirrors real-world user conditions (see Section 3.6).

10.3.3 Divergence Interpretation Requires Expertise

Divergence is not self-explanatory.
Analysts must interpret **types** of divergence (Section 6.3), not simply **amounts** of divergence.

Misinterpretation may lead to:

- treating structural divergence as error

- treating bias-driven divergence as substantive disagreement
- over-weighting a model's confidence

Proper understanding is mandatory for correct application.

10.4 Operational and Resource Constraints

10.4.1 Multi-Model Access Requirements

Ensemble reasoning requires reliable access to:

- GPT-5.1
- DeepSeek V2
- Gemini 2.0
- Grok
- Claude 3.5 (for synthesis)

Organizations lacking budget, API access, or workflow automation may face integration challenges.

10.4.2 Cost and Latency

Running large models multiple times:

- increases cost
- increases latency
- requires workflow automation to scale

High-volume decision pipelines (e.g., corporate risk, defense intelligence) should plan accordingly.

10.4.3 Analyst Training

The method requires human operators who understand:

- argumentation structure
- model reasoning signatures (Section 4)
- ensemble weighting (Section 8.3)
- divergence classification (Section 6)

This introduces onboarding time and organizational learning curves.

10.5 Safety and Alignment Constraints

10.5.1 Activation of Safety Filters

In domains that involve:

- political content
- extremism
- biosecurity
- military action
- contested historical claims

…models may:

- soften conclusions
- refuse engagement
- introduce caution language
- overgeneralize
- distort argument structure

Ensemble methods mitigate this by revealing the distortion, but cannot remove the underlying filter.

10.5.2 Alignment Drift Across Models

Different models have different alignment philosophies:

- Gemini: high-context moral framing
- GPT-5.1: structural caution
- Grok: safety-minimalist
- DeepSeek: confidence-forward reasoning

These alignment behaviors are *structural*, not errors, and must be accounted for.

10.5.3 Ethical Use Requirements

Ensemble reasoning can:

- amplify biases
- produce persuasive but incorrect interpretations
- generate outputs that appear authoritative

Practitioners in high-stakes domains must implement:

- human review
- auditability

- transparency protocols

10.6 Domain-Specific Constraints

10.6.1 Political and Legal Domains

Models are most sensitive to safety filters here.
Outputs may reflect:

- corporate safety policies
- risk aversion
- refusal patterns

This can obscure true interpretive disagreement.

10.6.2 Scientific and Technical Domains

Most agreement occurs here, but:

- poor handling of complex data
- inconsistent treatment of uncertainty
- weak numerical reasoning

...limit models in quantitative or high-fidelity scientific tasks.

10.6.3 Economic and Policy Domains

Models struggle with:

- dynamic feedback loops
- real-world data dynamics
- baseline assumptions about human behavior

Ensemble synthesis helps but cannot correct false premises or incomplete data.

10.6.4 Ethical and Metaphysical Domains

High abstraction amplifies:

- assumption variance
- interpretive drift (especially Gemini)
- philosophical depth differences

Ensembles reveal divergence but cannot determine correctness in non-empirical domains.

10.7 When Ensemble Reasoning Should *Not* Be Used

Ensemble methods are unnecessary or inefficient when:

- the task is simple classification or summarization
- correctness is easily verifiable
- time is critically constrained
- only one model is available
- the domain has little interpretive ambiguity
- the risk of misinterpretation is low
- resource cost is prohibitive

Examples:

- extracting dates from text
- proofreading
- simple coding tasks
- short factual lookups (non-sensitive)

10.8 Complementary Human-in-the-Loop Requirements

10.8.1 Human Judgment for Final Decisions

Ensemble output should inform—not replace—human expertise.

10.8.2 Documentation Requirements

Analysts should record:

- prompts used
- model versions
- assumptions flagged
- divergence types
- synthesis rationale

10.8.3 Cross-Validation with External Sources

Whenever possible:

- empirical datasets
- domain experts
- real-world evidence

...should validate ensemble outputs.

10.9 Summary

The ensemble framework is powerful but must be applied within defined constraints.
Its reliability depends on:

- prompt discipline
- divergence literacy
- model access
- human oversight
- proper interpretation of assumptions and alignment behaviors

Used correctly, the framework exposes structural reasoning differences and strengthens analytic rigor.
Used without awareness of limitations, it risks misinterpretation, overconfidence, or policy error.
Section XI outlines pathways for further development and scaling, addressing automation, quantitative scoring, and expansion of the argument suite.

XI. Expanded future research opportunities

11.1 Overview

The Leveraging LMM Bias framework represents a major step toward systematic, transparent, and reliable multi-model analysis. Yet the methods developed here also open clear pathways for extension, refinement, validation, and operational scaling. This section outlines the principal avenues for future research, tooling, and empirical development.

These directions fall into five major categories:

1. **Expansion of the argument test suite**
2. **Quantitative scoring and calibration systems**
3. **Automation and workflow integration**
4. **Cross-model drift tracking over time**
5. **Domain-specific and organizational deployments**

Among these directions, quantitative scoring and calibration (Section 11.3) and automation of ensemble pipelines (Section 11.4) represent the highest-priority near-term developments, because they directly address scalability, validation, and real-world deployability.

11.2 Expanding the Diagnostic Argument Suite

The current suite—13 arguments across scientific, political, legal, ethical, economic, and metaphysical domains—provides strong coverage, but additional categories would deepen the framework's robustness.

11.2.1 New Argument Classes

Future versions could add:

- Computational reasoning arguments
- Multi-step numerical logic puzzles
- Cybersecurity and adversarial reasoning
- International law and treaty interpretation
- Behavioral economics
- Systems engineering tradeoff arguments
- Intelligence analysis with possible deception

Each new category reveals different divergence pressures and reasoning signatures.

11.2.2 Increased Complexity and Layering

Additional argument variants:

- multi-premise chains
- ambiguous middle-premises
- contested empirical claims
- contradictory testimony
- hybrid empirical–normative structures

These will challenge models' ability to maintain structural coherence under uncertainty.

11.2.3 Real-World Case Benchmarks

Select historical case studies where ground truth is known:

- past monetary policy decisions
- Supreme Court rulings
- intelligence assessment failures or successes
- epidemiological control interventions

These provide opportunities to measure model agreement *relative to known outcomes*.

For example, retrospective analysis of the 2008 financial crisis or major intelligence-assessment failures could test whether ensemble methods would have flagged overlooked assumptions, structural vulnerabilities, or misplaced confidence earlier than single-model analysis.

11.3 Quantitative Scoring and Calibration

11.3.1 Weighted Divergence Index (WDI)

Future work should quantify divergence using:

- structural disagreement scores
- assumption density differentials
- evaluative weighting differences
- alignment-driven distortion signatures

This could become a **numerical divergence score**, enabling:

- benchmark comparisons
- automated flagging of high-risk interpretations
- longitudinal tracking

11.3.2 Agreement Stability Metrics

Metrics could include:

- *intra-model stability*: consistency across multiple runs
- *inter-model stability*: frequency of full/partial/structured/contradictory agreement
- *domain stability*: which domains produce predictable patterns

11.3.3 Calibration With External Data

Where empirical datasets exist (e.g., macroeconomic indicators), future versions could test how:

- model forecasts
- ensemble recommendations
- causal attributions

...align with actual outcomes.

Developing these calibration links would directly address the epistemic limits identified in Section X, by tying model and ensemble judgments to observable outcomes rather than purely internal agreement.

11.4 Automation and Workflow Integration

11.4.1 Automated Pipeline Execution

Future tools could automatically:

- run each model through 8-step and 14-step protocols
- normalize outputs
- classify divergence types
- generate synthesis reports
- store results for auditability

This would enable large-scale, repeatable testing.

11.4.2 API-Based Ensemble Orchestrators

A programmable system could:

- call each model with consistent session headers
- apply weighting rules
- generate consensus or flagged-disagreement summaries
- embed outputs in downstream analytics platforms

11.4.3 Enterprise and Government Integration

Organizations could integrate ensemble reasoning into:

- risk analysis workflows
- policy drafting pipelines
- legal review flows
- intelligence analysis cycles
- scientific research documentation

This would turn multi-model reasoning into a *standard analytical practice* rather than an ad hoc technique.

11.5 Longitudinal Drift and Model Evolution

11.5.1 Drift Tracking Across Versions

As models update, the framework could track:

- changes in reasoning signature
- shifts in assumption density
- new alignment behaviors

- improved or degraded interpretive performance

11.5.2 Temporal Divergence Maps

A timeline visualization could show:

- how DeepSeek's decisiveness evolves
- whether Gemini’s abstraction stabilizes
- if GPT-5.1 maintains structural rigor
- whether Grok’s minimalism becomes more consistent

This helps organizations manage risk when models update silently.

11.5.3 Version-Based Calibration Tables

Each new model version could be scored on:

- caution level
- decisiveness
- neutrality
- drift likelihood
- interpretive style

Future ensemble weightings could adjust automatically based on these scores.

11.6 Domain-Specific Extensions

Section IX outlined concrete application scenarios across multiple domains. Future work in this area focuses less on new use cases and more on *deepening* and *formalizing* those deployments—developing domain-tuned protocols, validation benchmarks, and governance patterns for specific sectors.

11.6.1 Legal & Judicial Applications

Future work could support:

- rapid multi-theory constitutional interpretation
- statutory ambiguity resolution
- precedent weighting analysis
- judicial screening tools

11.6.2 Scientific & Engineering Applications

Possible extensions:

- model-based uncertainty quantification
- causal inference cross-checking

- protocol-based verification for STEM workflows

11.6.3 Defense & Intelligence

Extensions could include:

- deception-resilient ensemble methods
- structured uncertainty propagation
- adversarial scenario generation

11.6.4 Corporate Decision Systems

Potential implementations:

- enterprise risk dashboards
- multi-model strategic planning tools
- competitive analysis ensembles
- governance and compliance review systems

11.7 Open Challenges

Several areas require continued research:

11.7.1 Distinguishing "True" Error from Productive Divergence

Not all divergence is equal—some reflects:

- flawed reasoning
- hallucinated assumptions
- safety distortion

...while other divergence is analytically useful.
Future work must classify these more precisely.

11.7.2 Measuring the Value of Structured Disagreement

Structured disagreement appears analytically valuable (Section 7), but quantifying this value remains open.

11.7.3 Single-Model Collapse Scenarios

As models converge in training data and architecture, future research must guard against:

- monoculture failure
- homogenized alignment
- reduced epistemic diversity

Ensemble methods are only useful if models remain *meaningfully distinct*.

11.7.4 Human–AI Interaction Effects

Analyst behavior, prompt phrasing, and iterative refinement all influence outputs. Future research could examine:

- bias amplification from human steering
- improvement of human calibration skills
- training analysts to read divergence correctly

11.8 Summary

The framework developed in this monograph is a foundation—an extensible architecture for understanding, auditing, and leveraging model divergence. Future work will expand the argument suite, introduce quantitative scoring tools, automate workflows, track model evolution, and build domain-specific applications.

In the near term, the most impactful work will be implementing quantitative divergence metrics and automated ensemble pipelines; longer-term efforts will focus on longitudinal drift tracking and sector-specific deployment patterns.

The central insight holds:

Model diversity, when measured and weighted rather than ignored or suppressed, becomes an analytical asset.

Continued development will further transform this insight into scalable, verifiable, and operationally useful systems for scientific, legal, political, corporate, and national security decision-making.

XII. Conclusion

The study presented in this monograph demonstrates that large multimodal models (LMMs) do not merely differ in style or tone—they differ in *structure*, *assumption density*, *interpretive framing*, and *evaluative strategy* in ways that are systematic, measurable, and analytically useful. Rather than treating variation among models as an obstacle, the ensemble framework developed here reframes diversity as a resource: divergence becomes a diagnostic signal rather than an error state.

Across the analysis, three findings consistently emerge:

1. **Reasoning signatures are identifiable and stable.**
   As shown in Sections IV and VI, each model exhibits reproducible patterns—

methodical structure (GPT-5.1), compressed decisiveness (DeepSeek), minimal textual anchoring (Grok), and context-rich abstraction (Gemini). These signatures persist across protocols, domains, and argument complexity.

2. **Divergence is interpretable and predictable.**
   Political and legal arguments tend to trigger alignment-sensitive interpretive divergence; economic arguments activate evaluation divergence; ethical and metaphysical arguments depend heavily on hidden assumptions; scientific arguments produce structural agreement with evaluative variation. This mapping, demonstrated in Section VII, enables analysts to predict where and why models will differ.

3. **Ensemble reasoning produces more stable, transparent outcomes.**
   Section VIII formalized a structured ensemble method in which each model's strengths are weighted according to domain needs. This approach consistently reduces overconfidence, exposes untested assumptions, and increases interpretability—benefits shown in both case studies and applied domains in Section IX.

Taken together, the contributions of this monograph are threefold:

- **A unified diagnostic protocol** (Section III) enabling structured, replicable evaluation of arguments.
- **A cross-model comparison framework** (Section VI) capable of distinguishing structural, interpretive, evaluative, and bias-driven divergence.
- **An actionable ensemble reasoning method** (Section VIII) that practitioners can deploy across scientific, legal, economic, ethical, and national-security contexts.

These findings have practical implications. Organizations that rely on LMMs—research institutions, legal teams, corporate strategists, intelligence analysts, policymakers—face increasing pressure to justify model outputs and detect failure modes before they reach operational decisions. The ensemble approach provides a pathway toward *auditable, multi-perspective reasoning* rather than opaque single-model conclusions.

At the same time, this work highlights important constraints. Interpretive ambiguity, alignment artifacts, and silent model updates can shift reasoning patterns unpredictably. Section X emphasized that ensemble reasoning reduces—but

cannot eliminate—these challenges. Section XI presented the roadmap forward: quantitative scoring systems, longitudinal drift tracking, automated orchestration, and domain-specific deployments that can evolve this framework into a durable analytical tool.

Ultimately, the central conclusion of this study is straightforward:

When LMMs are evaluated individually, their differences appear as noise. When evaluated together—systematically, transparently, and with structured weighting—those same differences become a source of epistemic strength.

This monograph provides the foundation for that transformation. The work ahead lies in turning ensemble reasoning from a methodological innovation into an operational standard—one capable of supporting high-stakes scientific, legal, political, economic, and national-security decision-making in an era defined by increasingly capable and increasingly diverse AI systems.