Master Degree in Computer Science

Master Degree in Data Science for Economics and Health

Natural Language Processing

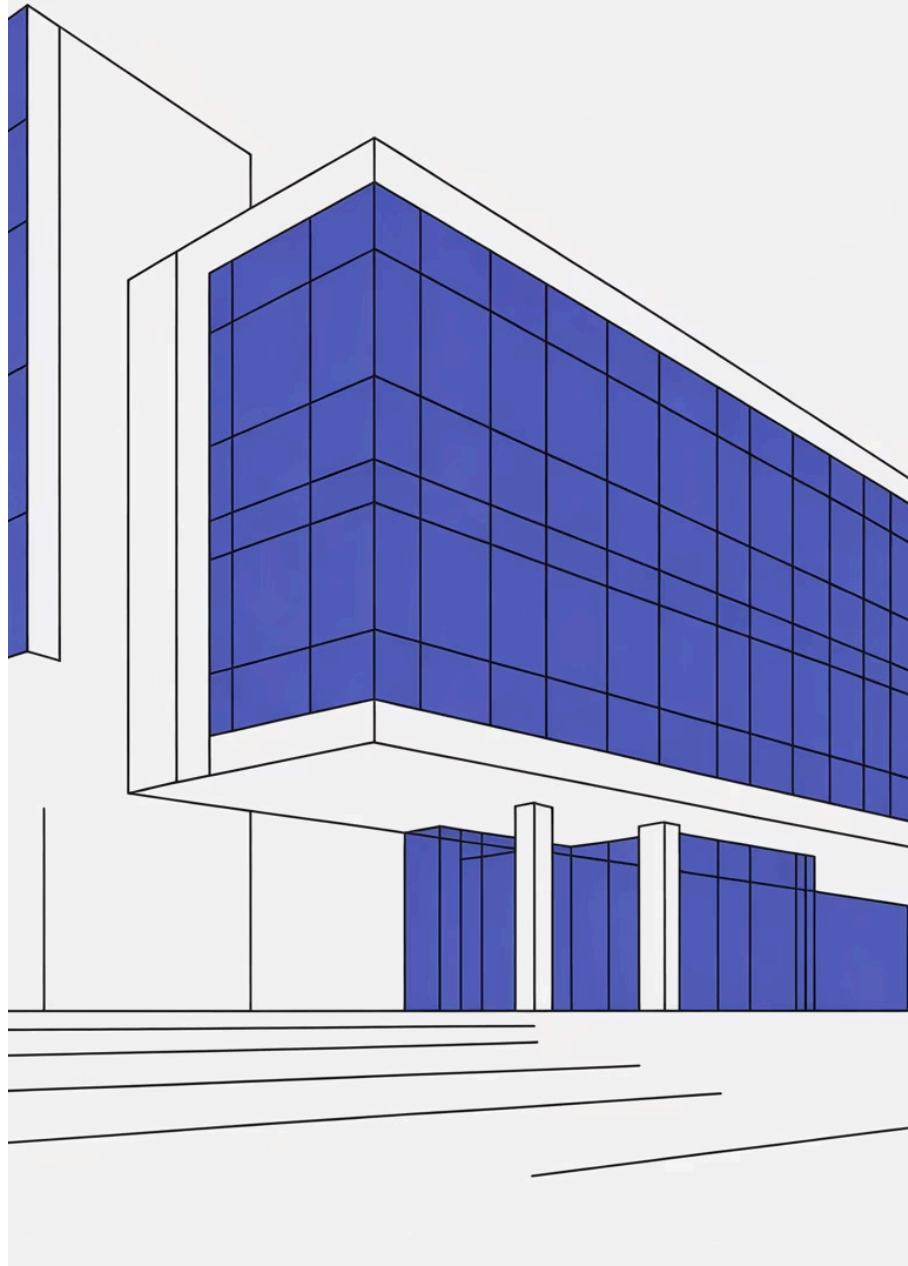
Prof. Alfio Ferrara

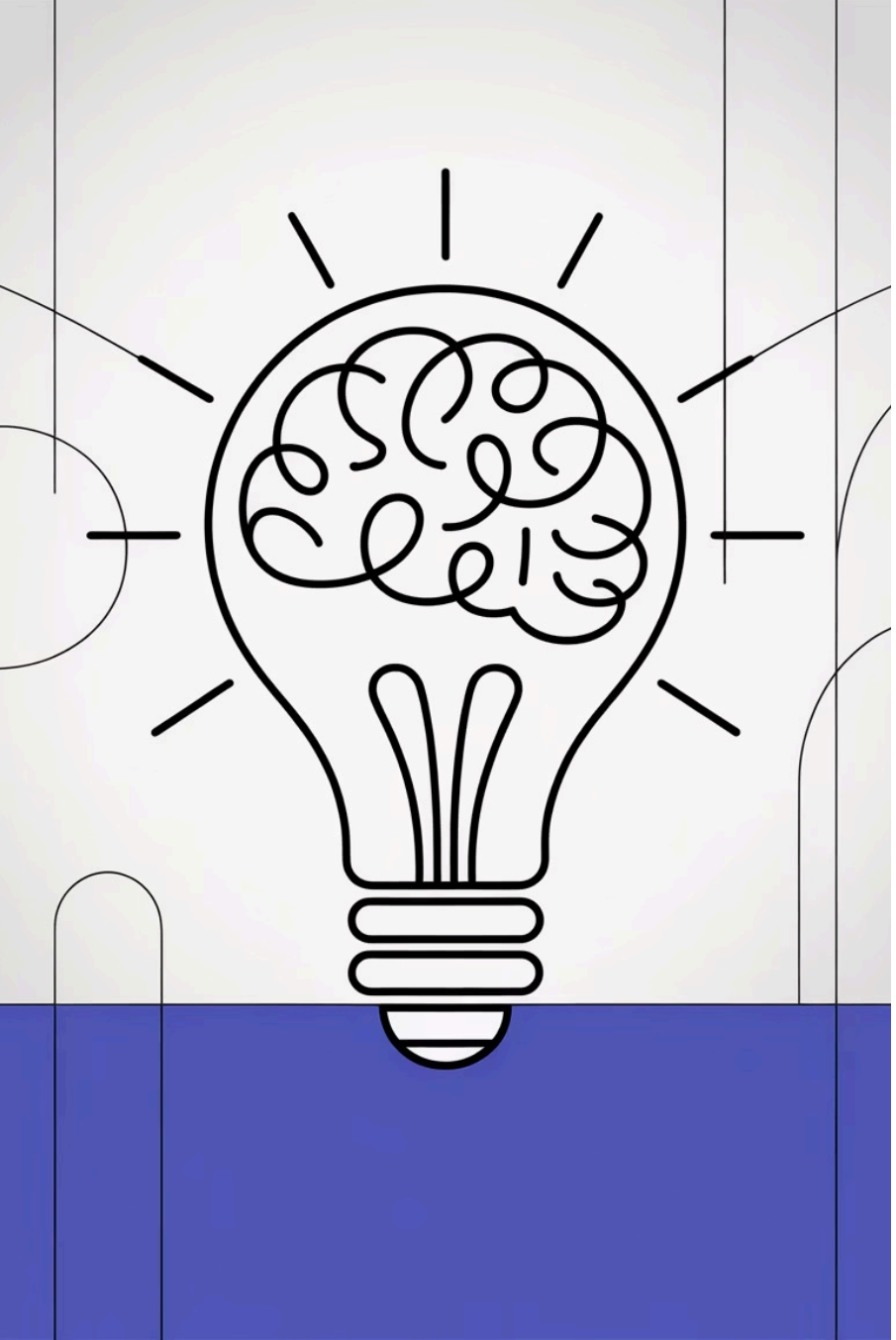
Dott. Sergio Picascia, Dott.ssa Elisabetta Rocchetti

Department of Computer Science, Università degli Studi di Milano

Room 7012 via Celoria 18, 20133 Milano, Italia

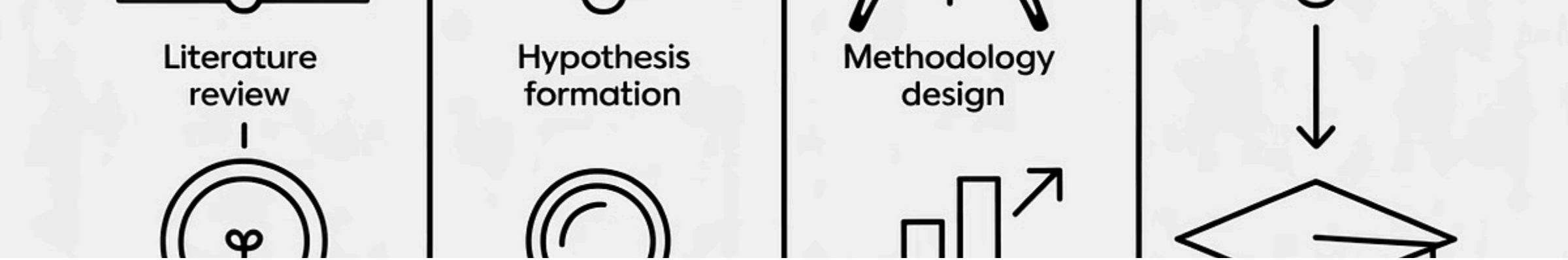
alfio.ferrara@unimi.it





Ideas for Final Projects

Explore cutting-edge research opportunities in Natural Language Processing through innovative project ideas spanning multiple thematic clusters. Each project offers unique challenges and methodological approaches to advance our understanding of language models and their applications.



Instructions

The final project consists in the preparation of a **short study** on one of the topics of the course, identifying a precise research question and measurable objectives. The project will propose a methodology for solving the research question and provide an experimental verification of the results obtained according to results evaluation metrics.

The emphasis is not on obtaining high performance but rather on the **critical discussion of the results obtained** in order to understand the potential effectiveness of the proposed methodology.

01

Documentation

Short article of 4-8 pages using provided templates

02

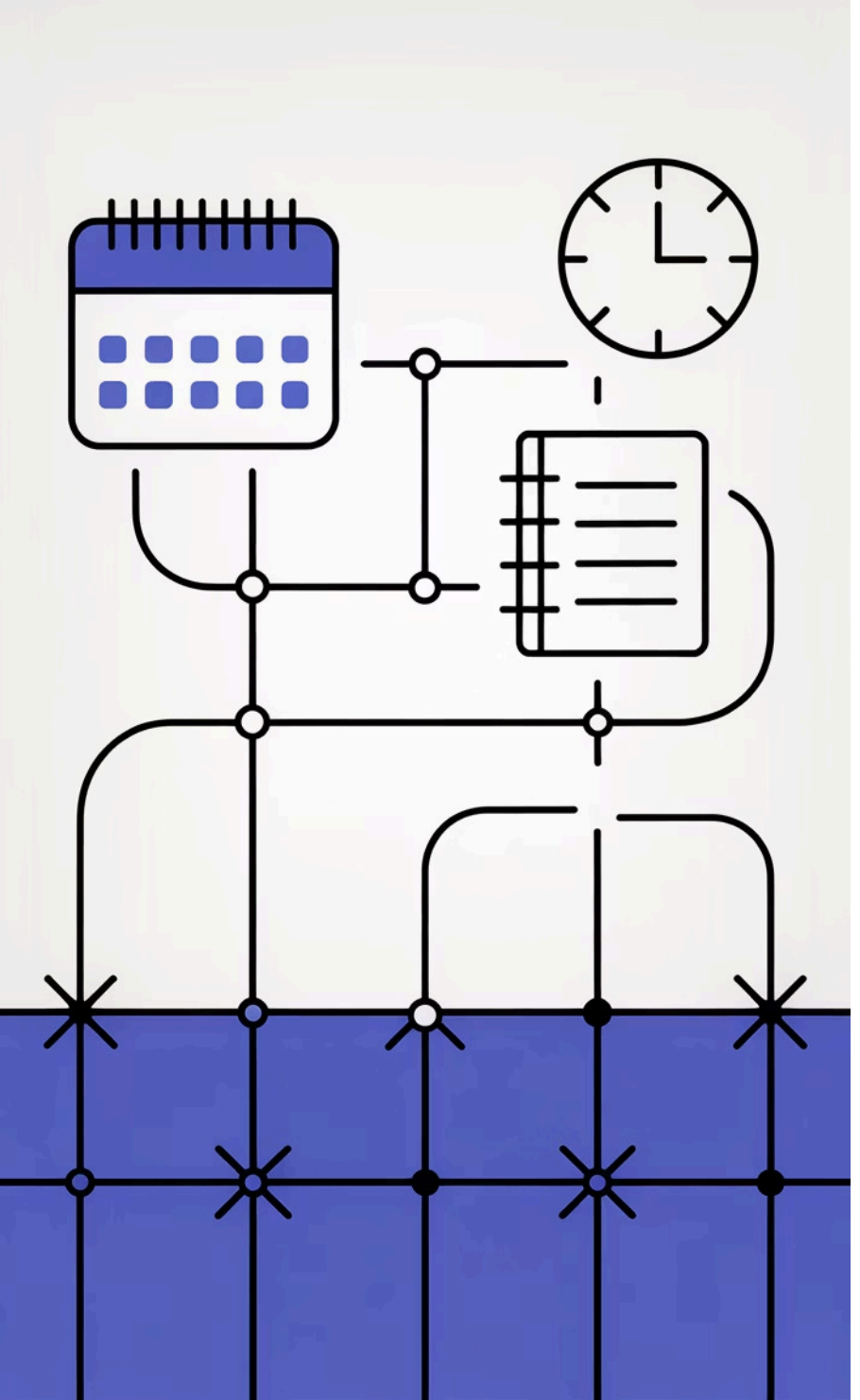
Code Repository

GitHub repository with reproducible experimental results

03

Presentation

10 minutes presentation in English with slides



Procedure

Exam dates are just for the registration of the final grade. The project discussion will be set by appointment, according to the following procedure:

1

Subscribe to Available Date

Register for any available exam date in the system

2

Contact Professor

Reach out when project is finished and ready for discussion

3

Setup Appointment

Schedule and discuss your completed work

Required Information: When contacting Prof. Ferrara, provide: (1) Your subscribed exam date, (2) PDF version of your report, (3) GitHub repository link

If you are **interested in doing your final master thesis** on these topics, the final project may be a preliminary work in view of the thesis. Discuss the contents with Prof. Ferrara during the project discussion.

Structure of the Paper

1. Introduction

Provides an overview of the project and a short discussion on the pertinent literature

2. Research Question & Methodology

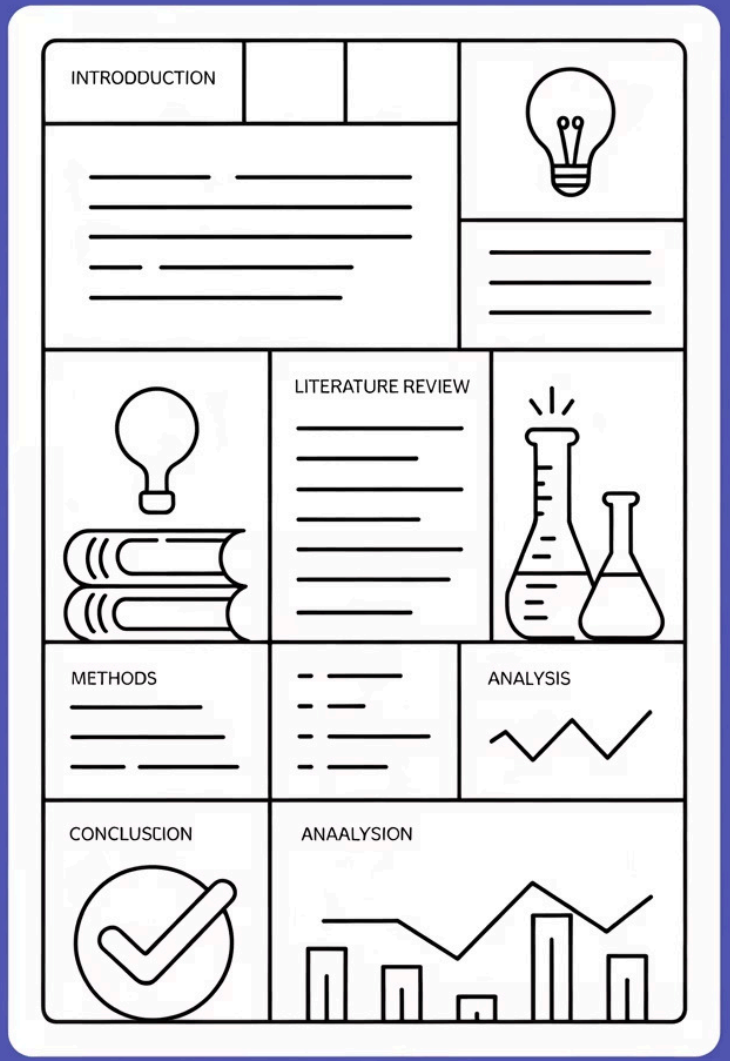
Clear statement of goals, overview of proposed approach, and formal problem definition

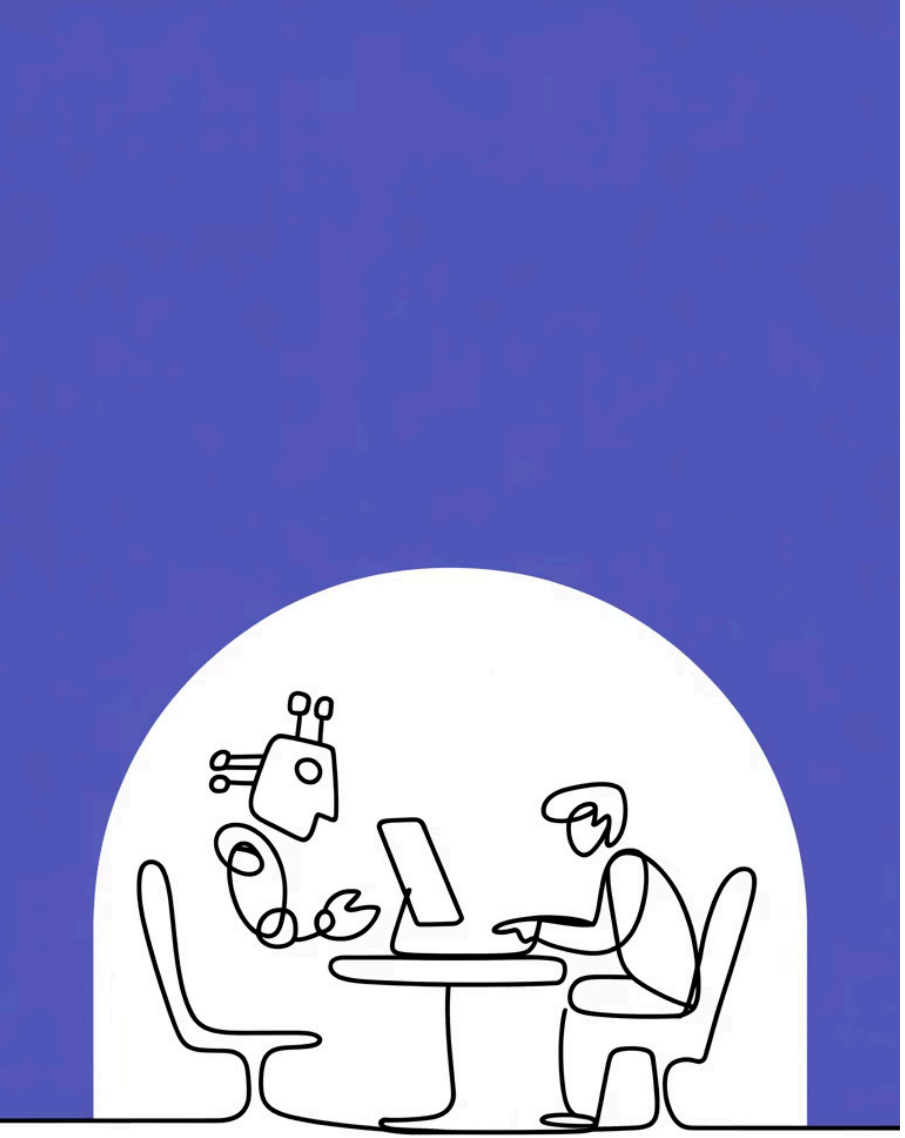
3. Experimental Results

Dataset overview, evaluation metrics, experimental methodology, and results presentation

4. Concluding Remarks

Critical discussion of results and ideas for future work





AI Usage Disclaimer

Parts of this projects have been developed with the assistance of **OpenAI's ChatGPT (GPT-5)**. The AI was used to support the **development of project ideas, the structuring of methodological workflows, the drafting of descriptive texts**, and the **identification of relevant datasets and references**.

All content produced with AI assistance has been **carefully reviewed, edited, and validated** by me. I take full responsibility for the final content and its accuracy, relevance, and academic integrity.

Using AI (for students)

Generative AI tools (such as ChatGPT, Claude, Mistral, or similar models) **may be used in this project**, both as an object of investigation and as a tool to support the development process.

Encouraged Uses

Explore how models function, interact creatively, leverage for inspiration, ideation, drafting, or experimentation

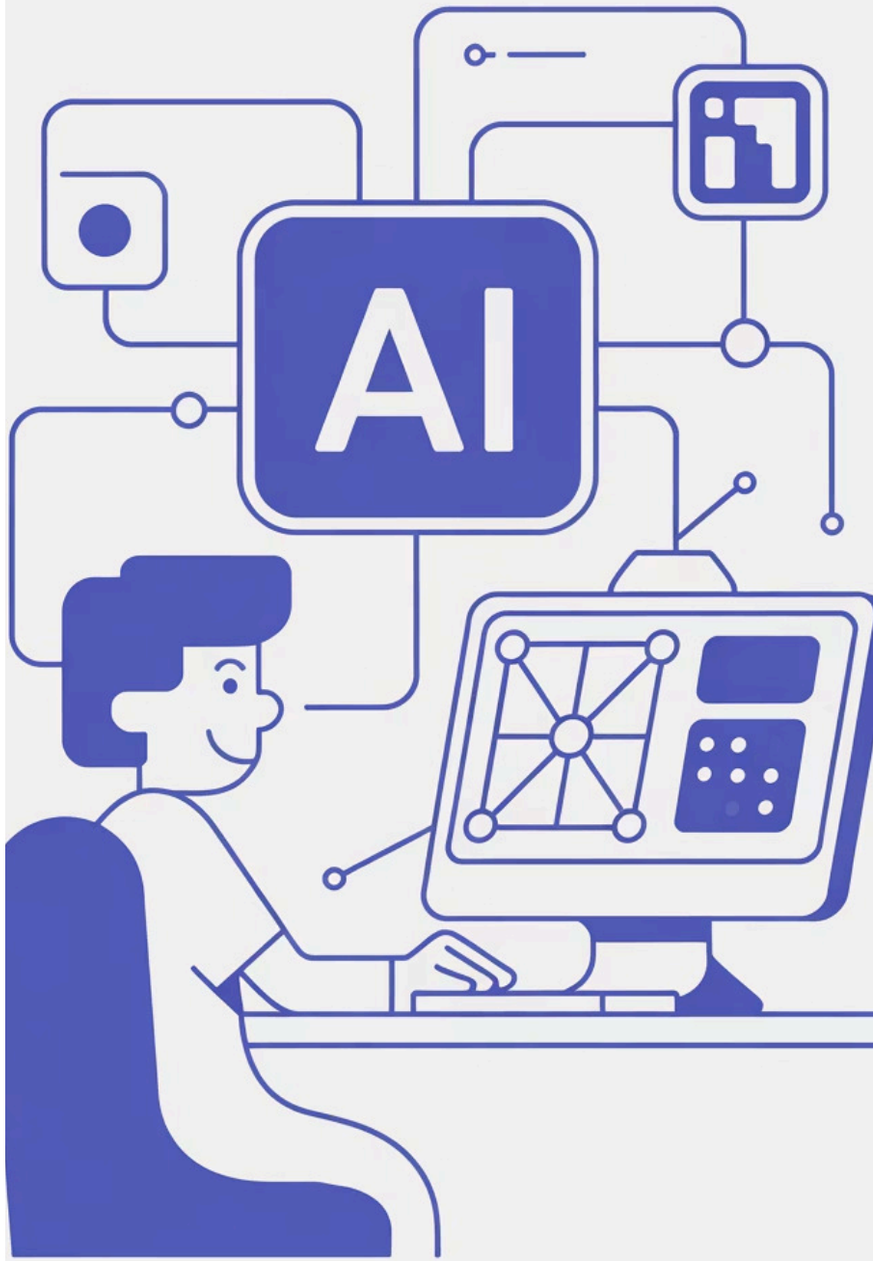
Important Limitation

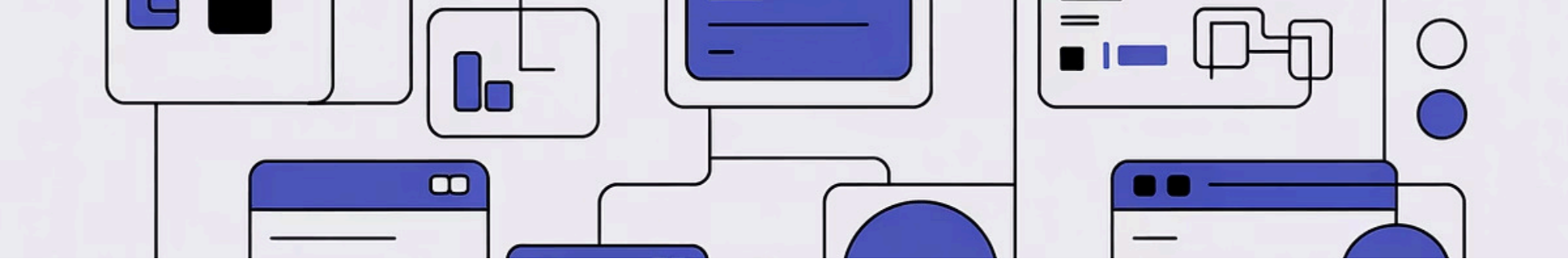
AI should not substitute original work. Student responsibility for structure, reasoning, and understanding remains

Mandatory Disclaimer

Specify which models used, for what purposes, and to what extent outputs were modified or verified

The project will be assessed on output and the **student's ability to explain and justify all choices made**. A final interview will evaluate depth of understanding.





Instructions on Coding

All project code should be written with **clarity, modularity, and reusability** in mind. The implementation should not consist of a single large notebook, but rather follow a structured and maintainable design.

1 Python Modules & OOP

Organize logic into Python modules and packages using object-oriented programming principles

2 Jupyter for Demo

Use notebooks primarily for demonstration, experimentation, and visualization

3 Separation of Concerns

Separate data loading, preprocessing, model interface, evaluation, and visualization

4 Clean Repository

Ensure clean, reproducible, and extensible codebase for replication and future development

Project Ideas

The following are ideas for projects. For each idea, a short description, example of datasets that can be used, and bibliographic references are provided.

Choose Existing Ideas

Students may **choose one of the following** as their project theme, with complete descriptions, methodologies, and references provided.

Propose Your Own

Students can **propose their own idea**, structuring the proposal as those presented in this document. Send project description to Prof. Ferrara.

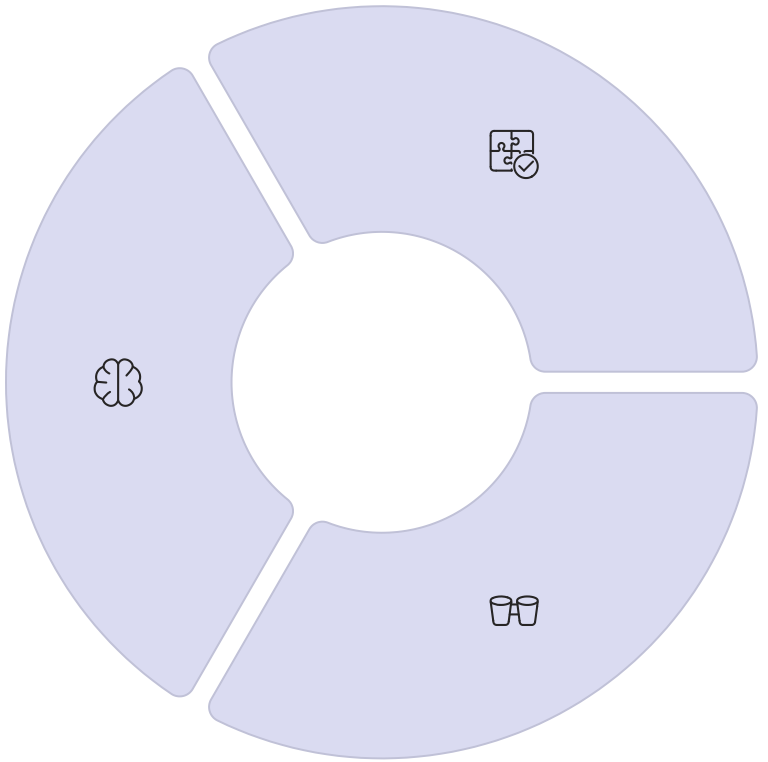
Projects are organized in thematic clusters. The methodological notes, datasets and references are intended as starting guidelines. Students are encouraged to find their own data and references when needed.

Thematic Cluster 1: Reasoning, Logic & Cognition

This cluster investigates the **reasoning and cognitive capacities** of Large Language Models (LLMs). Projects in this group focus on truthfulness, logical inference, and the interaction between symbolic and linguistic knowledge.

Truthfulness

Analyzing how models handle false, incomplete, or contradictory information

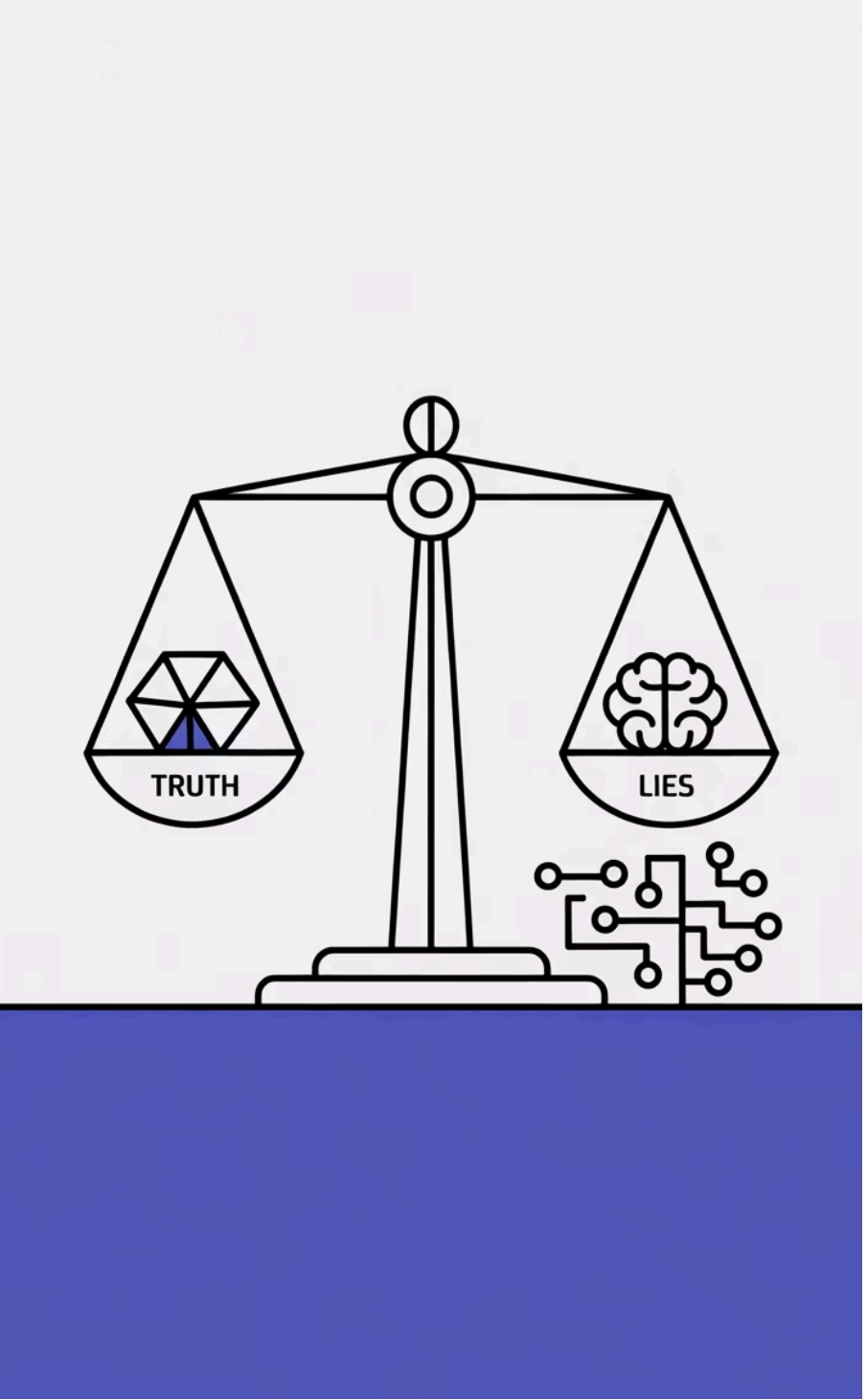


Logical Inference

Examining multi-step reasoning and consistency in logical operations

Symbolic Knowledge

Exploring interaction between symbolic and linguistic knowledge systems



P1. Truth, Lies, and Reasoning Machines

This project investigates how Large Language Models (LLMs) reason when exposed to **false, incomplete, or contradictory information**. The central goal is to understand whether these models can detect inconsistencies, resist misinformation, and maintain logical coherence during multi-step reasoning.

Core Pipeline

Construct controlled reasoning datasets with factual, counterfactual, and contradictory cases. Extract and visualize reasoning chains using factuality metrics.

Expected Outcomes

Quantify logical coherence deterioration under misinformation stress and evaluate self-verification prompts effectiveness.

Methodology

01

Design Reasoning Tasks

Create tasks involving factual and counterfactual statements or inject controlled falsehoods into multi-hop question answering datasets

02

Compare Model Behaviors

Test across setups: baseline (factual), noisy (contradictory), and adversarial (deliberately misleading)

03

Apply Explainability Tools

Use attention visualization, token attribution, or probability tracing to analyze deviation points

04

Implement Self-Correction

Test verification prompts ("Are you sure?", "Check your assumptions") for introspective reasoning

05

Evaluate Outputs

Use logical validity metrics, factual accuracy, and qualitative inspection of reasoning chains

Students are encouraged to analyze the *types of reasoning failures* (e.g., belief persistence, hallucination propagation, circular logic) and discuss how truth distortion affects reasoning reliability.

Dataset & References

TruthfulQA

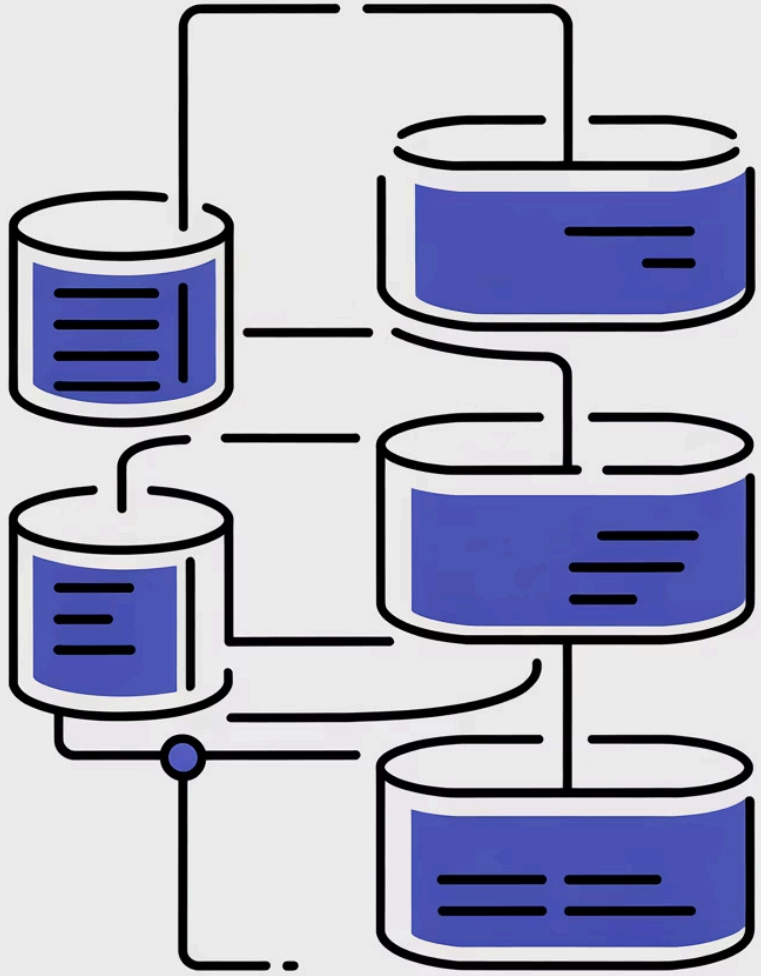
Measuring How Models Mimic Human Falsehoods
https://github.com/sylinrl/TruthfulQA

HotpotQA

Dataset for Diverse, Explainable Multi-hop Question Answering
https://hotpotqa.github.io/

References

- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35, 17359-17372.
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of ACL*, 3214-3252.
- Zhou, X., Wang, Q., Wang, X., Tang, H., & Liu, X. (2023). Large language model soft ideologization via AI-self-consciousness. *arXiv preprint arXiv:2309.16167*.



P2. The Knowledge Translator

This project investigates whether **Large Language Models (LLMs)** can act as *semantic interpreters* for structured knowledge queries. Given a **formal query** and a **textual corpus** as the only knowledge source, the model retrieves, synthesizes, and justifies answers using natural language understanding rather than database reasoning.

1

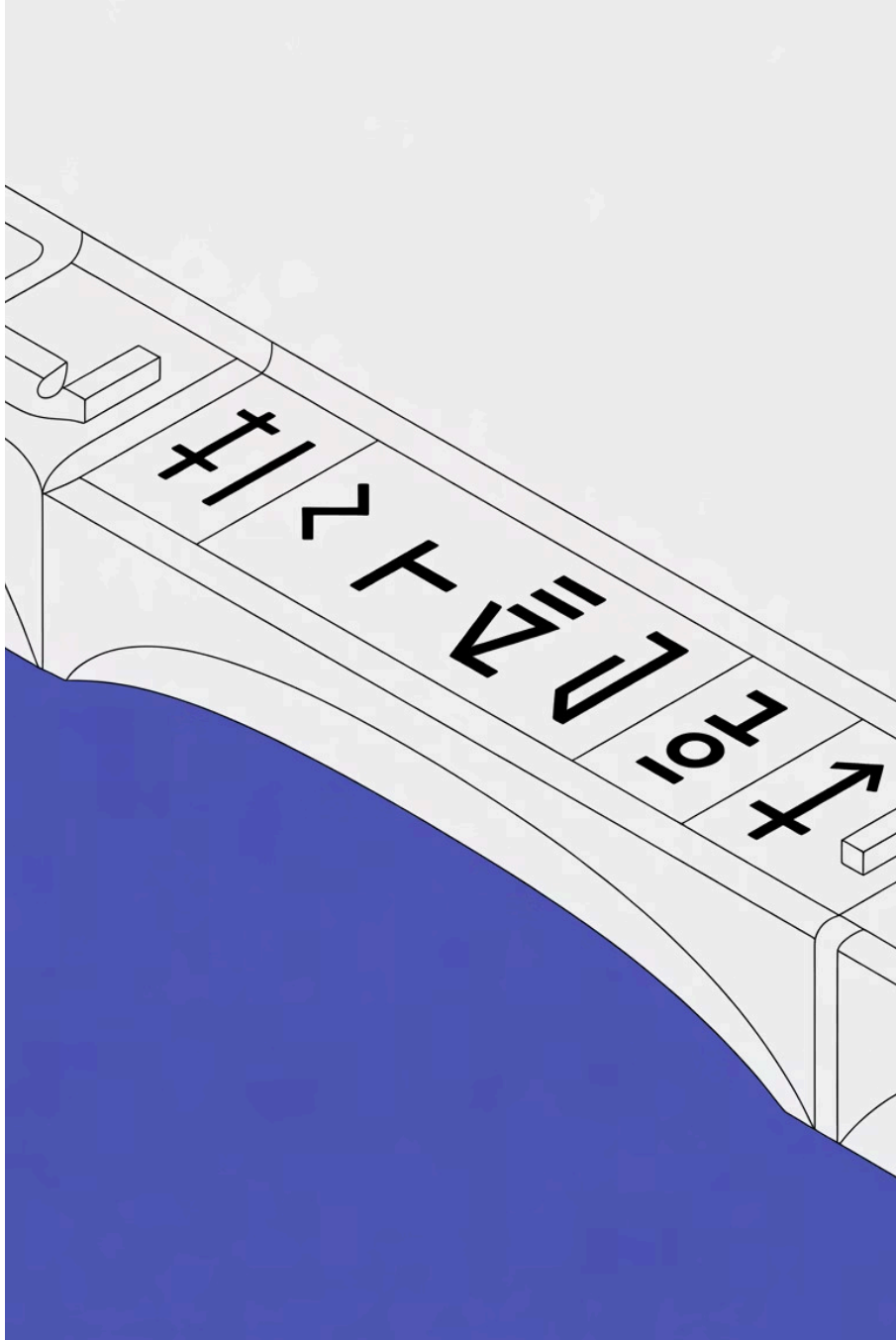
Core Pipeline

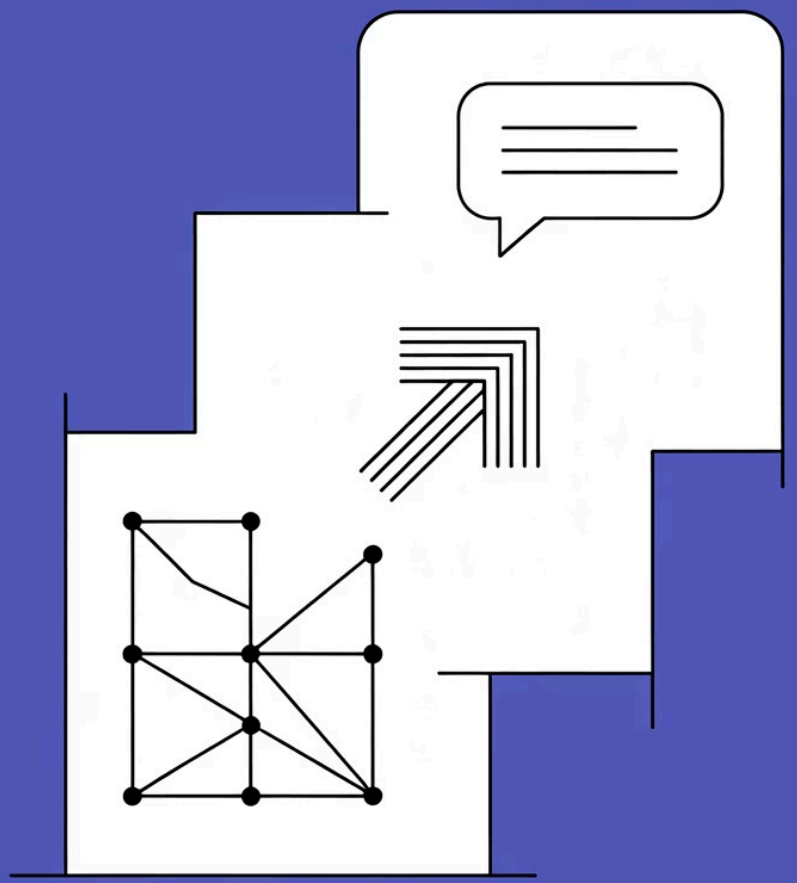
Structured queries from ontologies paired with unstructured textual corpora. LLM interprets queries and returns variable bindings based on textual evidence.

2

Expected Outcomes

Empirical insight into LLM capacity to bridge symbolic queries and natural language reasoning, revealing linguistic inference compensation.





Methodology

Formal Query Definition

Select structured queries from existing ontology or knowledge graph (e.g., ? author wrote ?book WHERE book.genre = 'science fiction')

Textual Knowledge Source

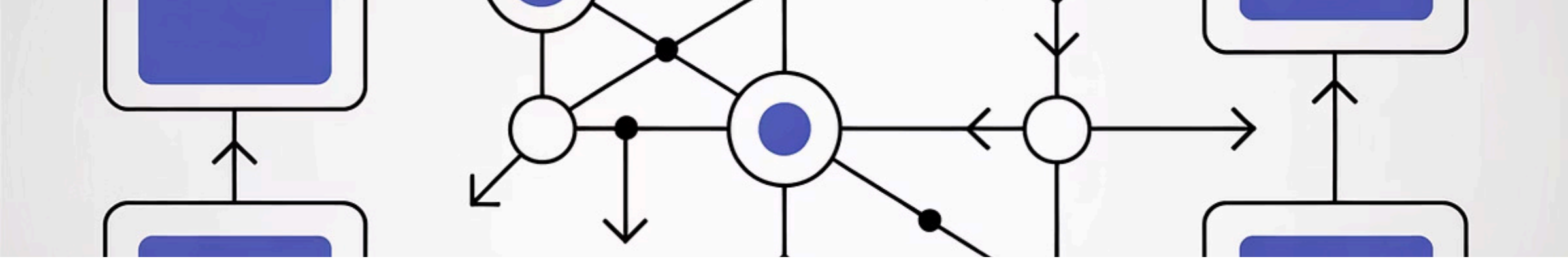
Prepare corpus containing relevant information but not in structured form (Wikipedia articles, domain-specific texts)

Model Task Design

Guide LLM to simulate logical reasoning: identify entities/relations, extract variable assignments, output structured format

Evaluation & Comparison

Measure semantic correctness against baselines, analyze error types: misunderstanding operators, entity mismatch, overgeneralization



Dataset & References

Wikidata / DBpedia

Generate reference triples and queries for evaluation against structured knowledge bases

Wikipedia Corpus

Textual grounding source providing unstructured information for natural language inference

MetaQA Benchmark

Multi-hop question answering over knowledge graphs, adaptable for text-based inference tasks

References

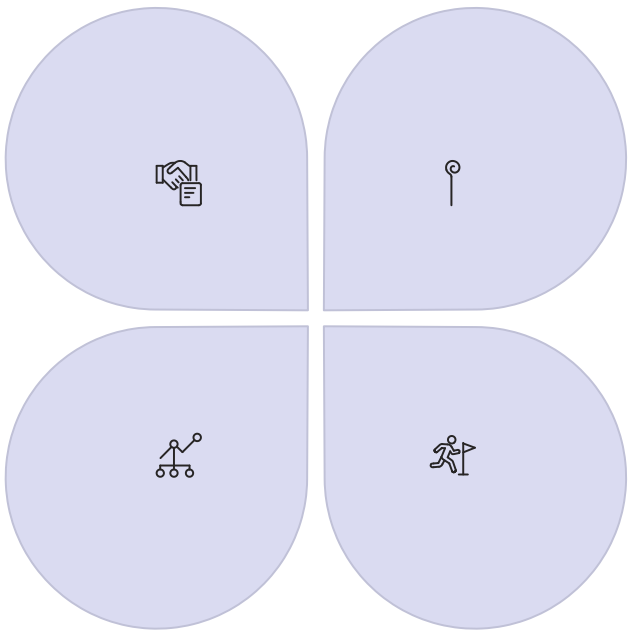
- Saeed, Mohammed, Nicola De Cao, and Paolo Papotti. "Querying large language models with SQL." *arXiv preprint arXiv:2304.00472* (2023).
- Badaro, G., Saeed, M., & Papotti, P. (2023). Transformers for tabular data representation: A survey. *Transactions of the Association for Computational Linguistics*, 11, 227-249.
- Ngonga Ngomo, A. C., et al. (2013). Sorry, i don't speak SPARQL: translating SPARQL queries into natural language. *WWW Conference*, 977-988.

Thematic Cluster 2: Agentic Behavior & Interaction

This cluster explores the **emergent agency** of LLMs in interactive or multi-agent settings. It studies negotiation, cooperation, competition, and coordination between artificial agents, providing insights into pragmatic communication, strategic behaviour, and collective intelligence.

Negotiation

Strategic dialogue and agreement-reaching between autonomous agents



Cooperation

Collaborative behaviors and shared goal achievement

Coordination

Multi-agent alignment and collective intelligence emergence

Competition

Adversarial interactions and strategic advantage-seeking

P3. The Negotiation Arena

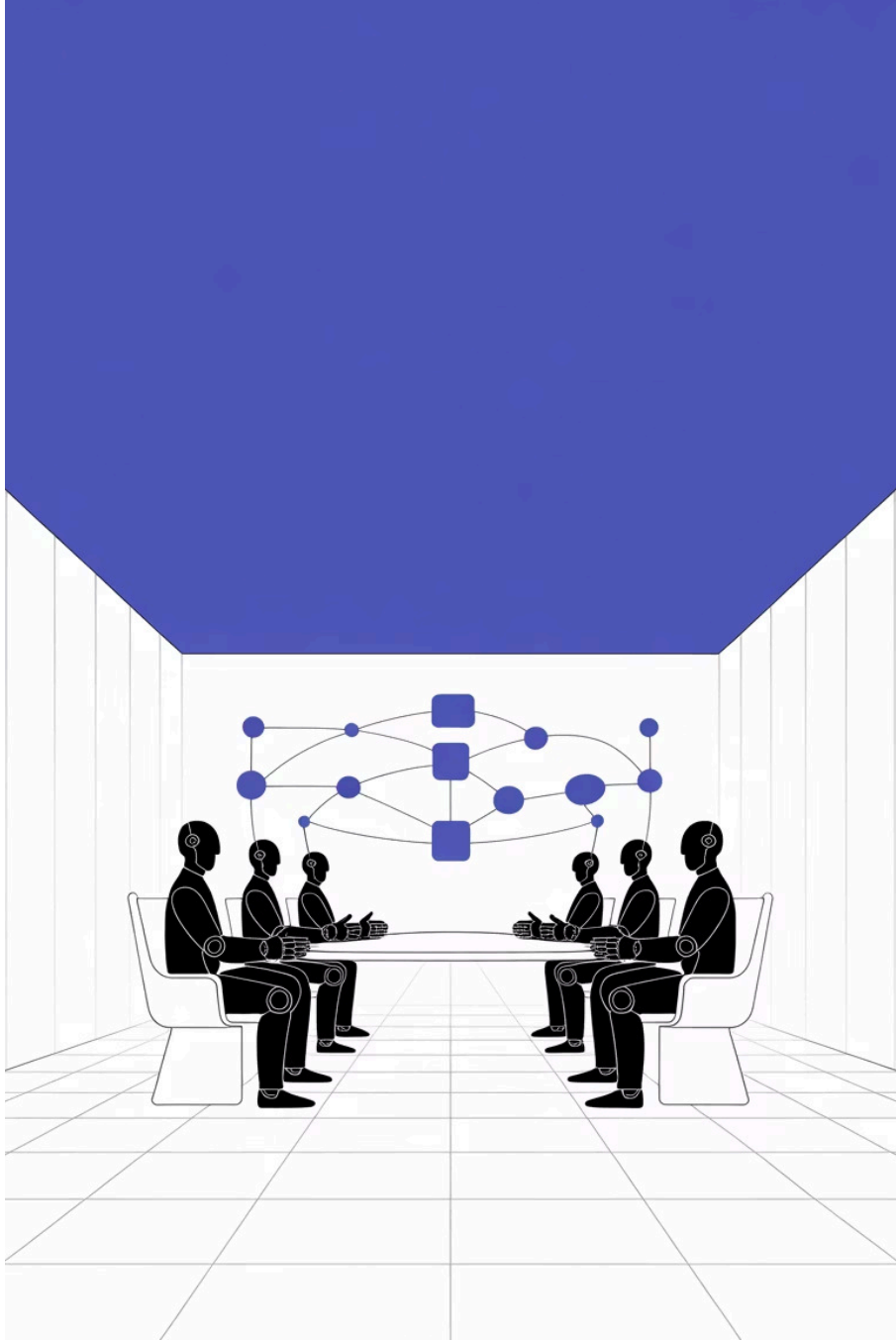
This project investigates how **Large Language Models (LLMs)** behave as autonomous agents engaged in negotiation, cooperation, or strategic dialogue. Two or more models are placed in simulated scenarios where they must **reach an agreement, trade resources, or align on decisions** despite having distinct goals or incomplete information.

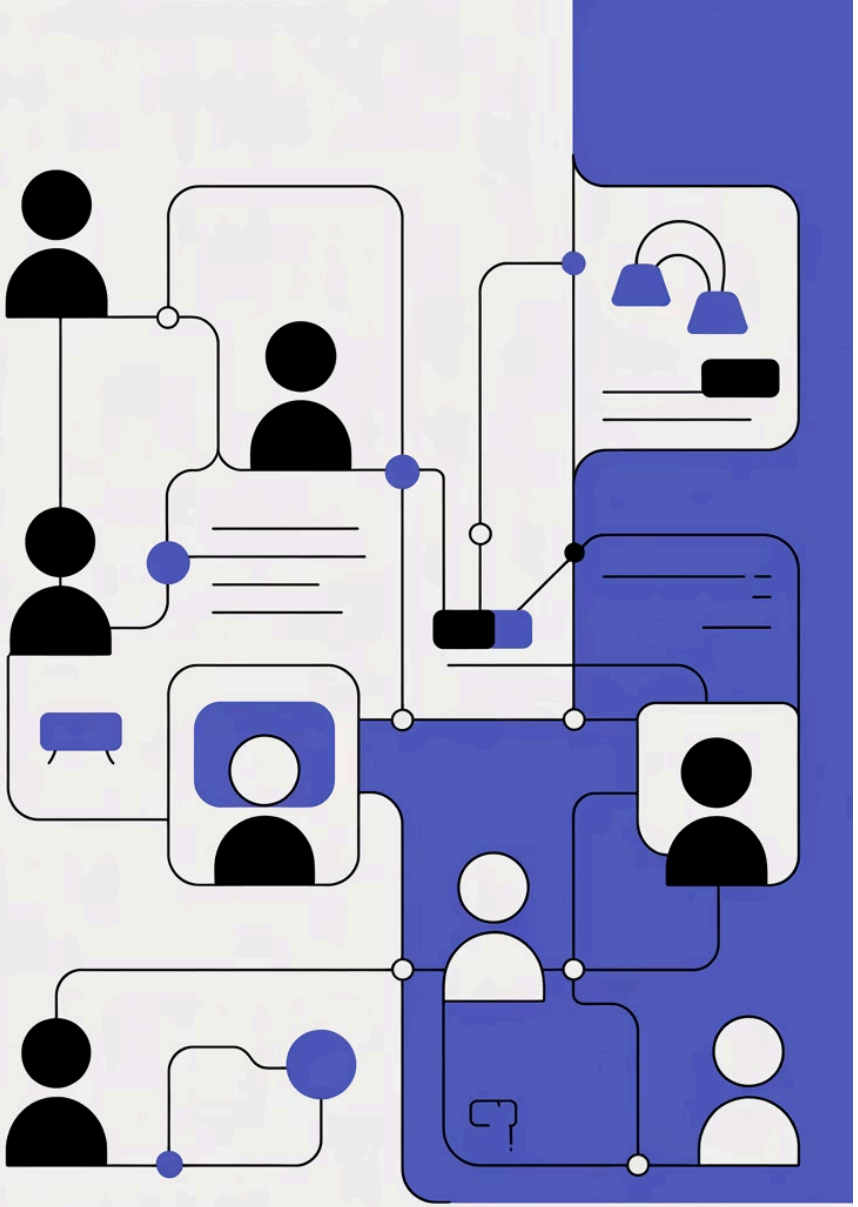
Core Pipeline

Agents instantiated with distinct goals engage in multi-round conversations. Simulations log exchanges, evaluated quantitatively and qualitatively.

Expected Outcomes

Uncover cooperative or adversarial behaviors, identify pragmatic features correlated with negotiation success or failure.





Methodology

1 Scenario Design

Define negotiation settings: resource division, task scheduling, or preference alignment. Each agent receives private information or asymmetric incentives.

2 Agent Configuration

Instantiate LLM agents with distinct personas or objectives. Include optional adjudicator model or human evaluator.

3 Dialogue Simulation

Implement iterative conversation rounds until agreement or impasse. Test cooperative, competitive, and mixed modes.

4 Analysis & Metrics

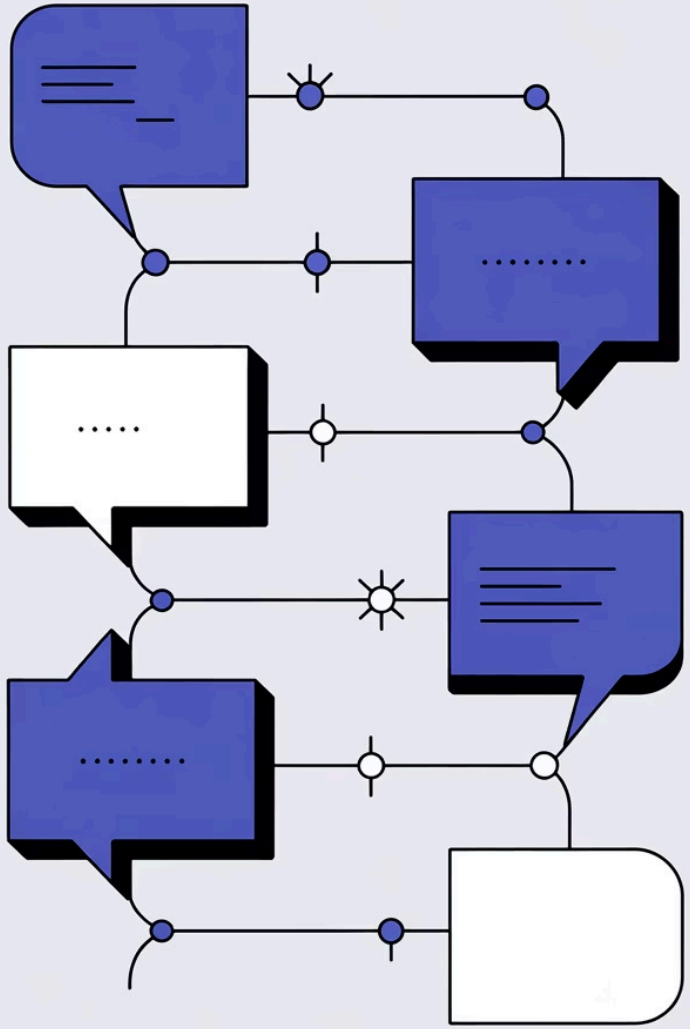
Measure agreement rate, convergence rounds, utility scores, language complexity. Analyze persuasion tactics and emotional tone.

Dataset & References

Dataset: No fixed dataset required; negotiation scenarios can be **synthetically generated** or adapted from existing dialogue datasets.

References

- Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., & Batra, D. (2017). Deal or No Deal? End-to-End Learning of Negotiation Dialogues. *EMNLP*, 2443-2453.
- Akin, S., et al. (2025). Socialized Learning and Emergent Behaviors in Multi-Agent Systems based on Multimodal Large Language Models. *arXiv preprint arXiv:2510.18515*.
- Gupta, P., et al. (2025). The Role of Social Learning and Collective Norm Formation in Fostering Cooperation in LLM Multi-Agent Systems. *arXiv preprint arXiv:2510.14401*.

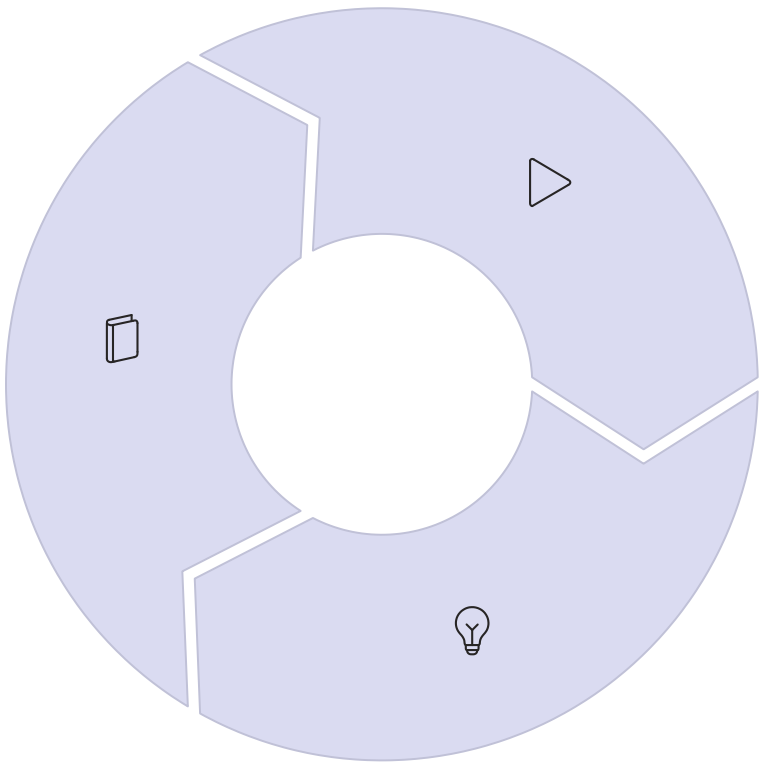


P4. Game of Thoughts

This project explores how **Large Language Models (LLMs)** understand, manipulate, and generate **structured rule systems** — from interpreting existing board games to inventing entirely new ones. By treating games as a proxy for structured reasoning, the goal is to evaluate logical consistency, creativity, and procedural understanding.

Rule Comprehension

Interpret existing game rules, simulate valid moves from intermediate states



Game Simulation

Execute gameplay scenarios while maintaining rule consistency

Creative Generation

Generate new playable games with internal consistency and procedural validity

Methodology

Game Understanding

Provide natural-language rulebooks, ask model to explain rules, identify ambiguities, suggest corrections

Game Simulation

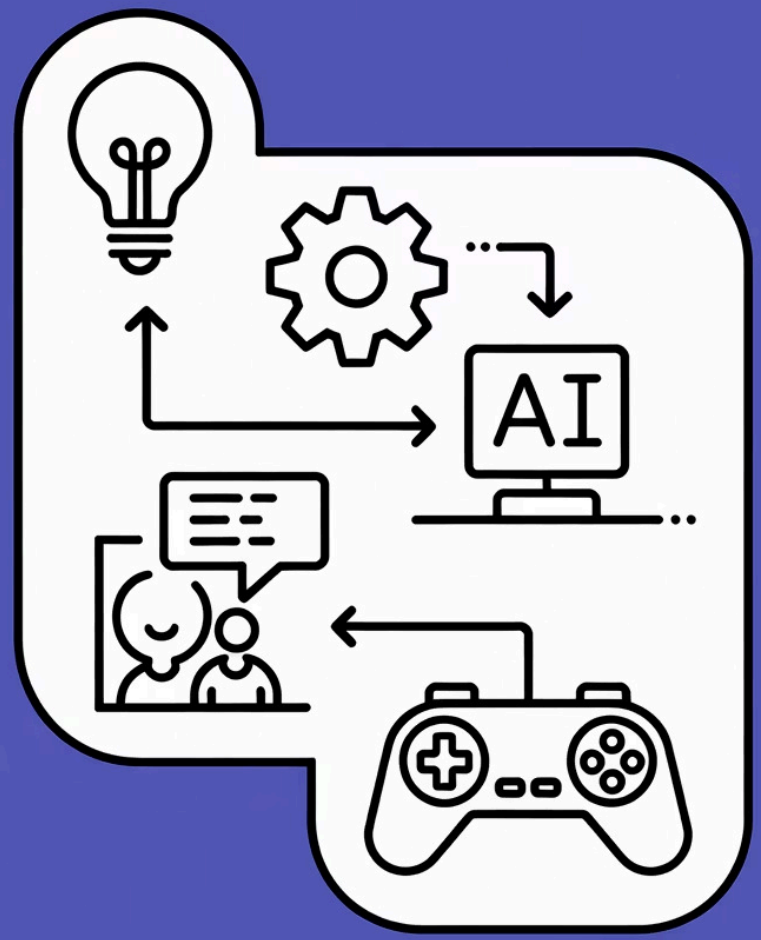
Feed current game state, request next valid move or strategic suggestion, evaluate rule adherence

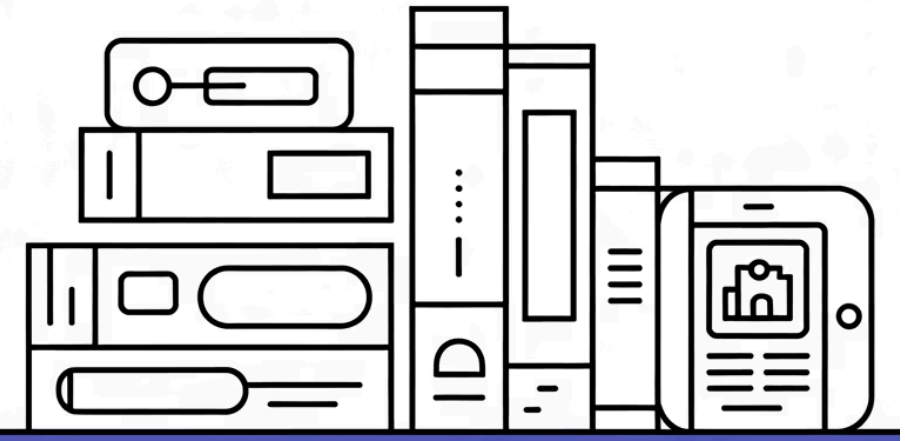
Game Generation

Instruct model to design new game given theme or constraint, assess originality and playability

Evaluation

Compare with ground truth, evaluate internal consistency, balance, clarity, and fun factor





Dataset & References

BoardGameGeek (BGG) API

Metadata and rule summaries for thousands of board games

Official Rulebooks

Freely available online from publishers

Ludii Game Database

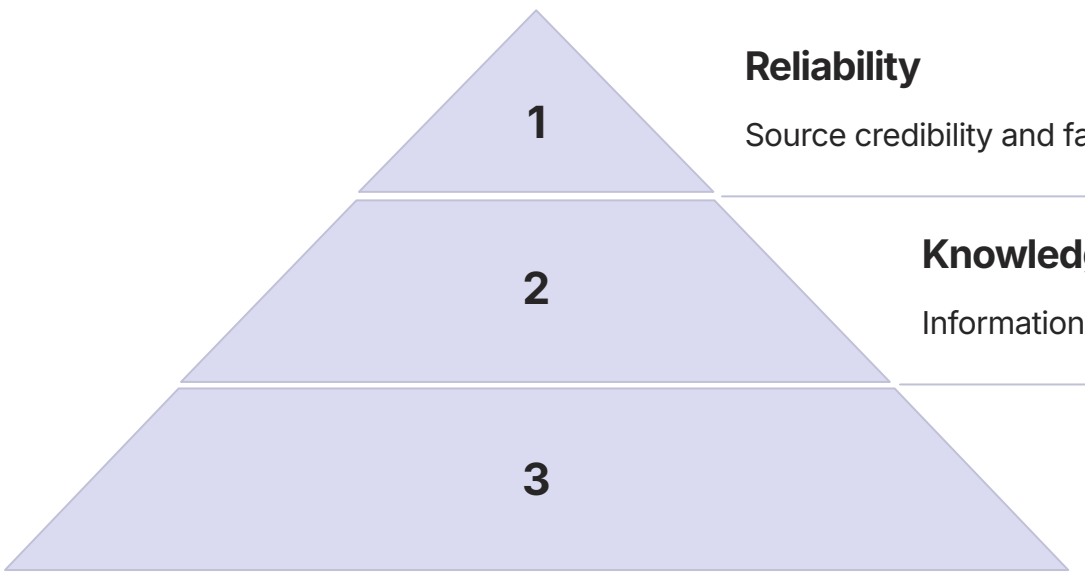
Structured repository of formalized game rules

References

- Todd, G., et al. (2024). Gavel: Generating games via evolution and language models. *Advances in Neural Information Processing Systems*, 37, 110723-110745.
- Hu, C., Zhao, Y., & Liu, J. (2024). Game generation via large language models. *IEEE Conference on Games*, 1-4.
- Piette, E., et al. (2021). General board game concepts. *IEEE Conference on Games*, 01-08.

Thematic Cluster 3: Knowledge, Retrieval & Robustness

This cluster focuses on the **reliability and structure of knowledge** in LLMs and retrieval-augmented systems. Projects analyse the robustness of retrieved information, knowledge compression or distillation, and mechanisms by which models verify, maintain, or adapt factual content across domains.



Reliability

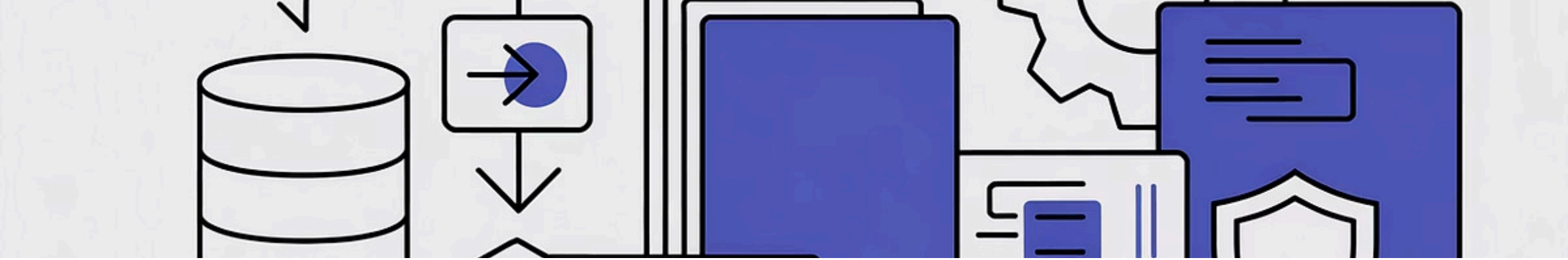
Source credibility and factual verification

Knowledge Structure

Information organization and representation

Robustness Analysis

System resilience against misinformation and conflicting sources



P5. In RAG We Trust?

This project evaluates how **Retrieval-Augmented Generation (RAG)** systems manage source credibility and factual reliability when retrieving information from multiple documents. Instead of designing new RAG architectures, students focus on **quantifying and analyzing reliability** by measuring how models weigh and reconcile conflicting or falsified sources.

Core Pipeline

RAG pipeline with intentional "poisoned" retrieval component. Test robustness against false or conflicting evidence while prompting source reliability assessment.

Expected Outcomes

Quantitative assessment of RAG system misinformation handling, producing metrics and qualitative insights on trust mechanisms.

Methodology

- **Controlled Data Poisoning**

  Introduce intentional falsehoods or contradictions into retrieved document sets

- **Multi-source Verification**

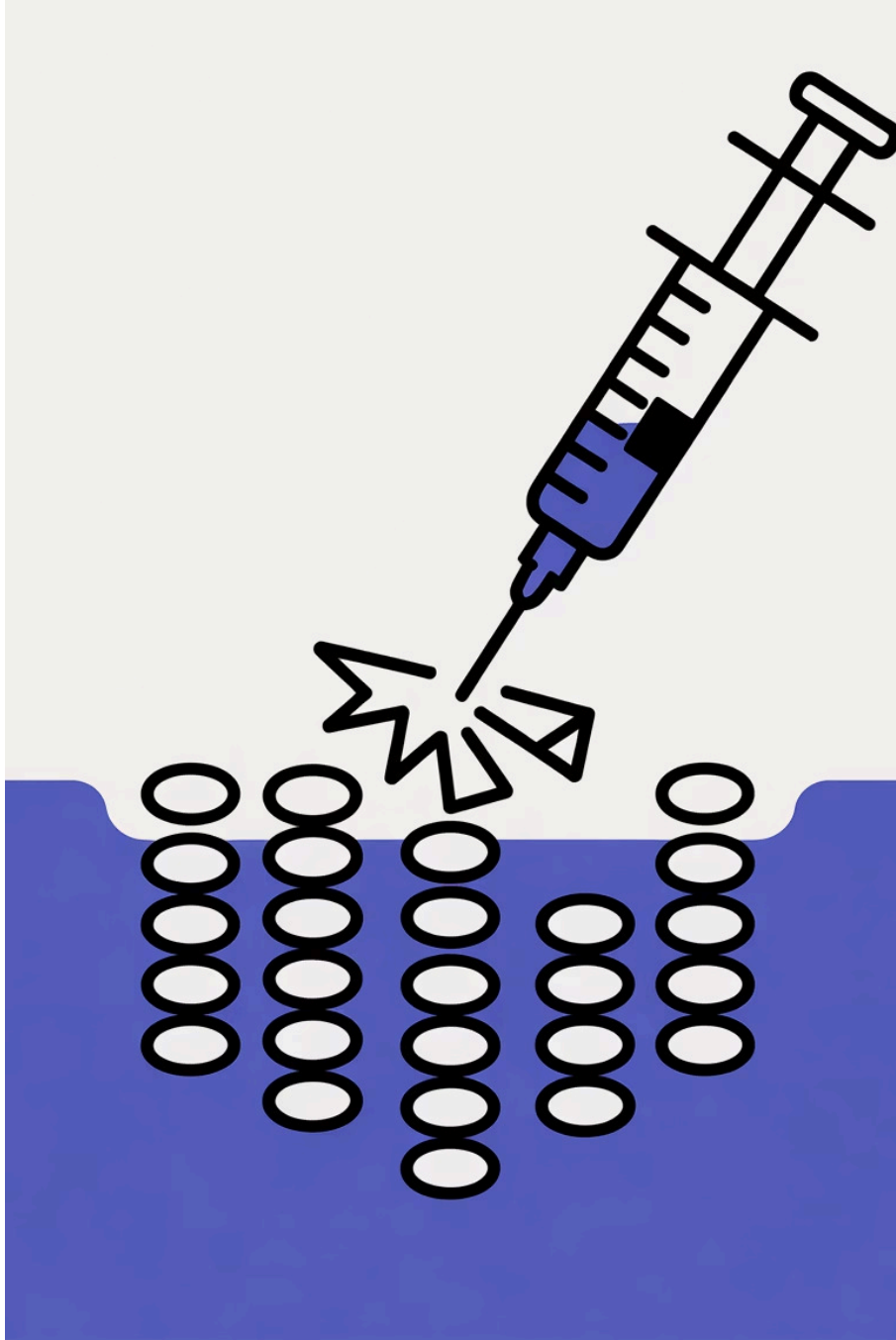
  Measure whether model identifies inconsistencies, seeks confirmation, or hedges answers

- **Prompt Design**

  Test whether meta-prompts improve reliability (e.g., "check consistency across documents")

- **Evaluation**

  Compare factual accuracy, hallucination rate, and self-consistency across models and retrieval settings



Dataset & References

1

FEVER Dataset

Fact-checking and verification benchmark for evidence-based reasoning

2

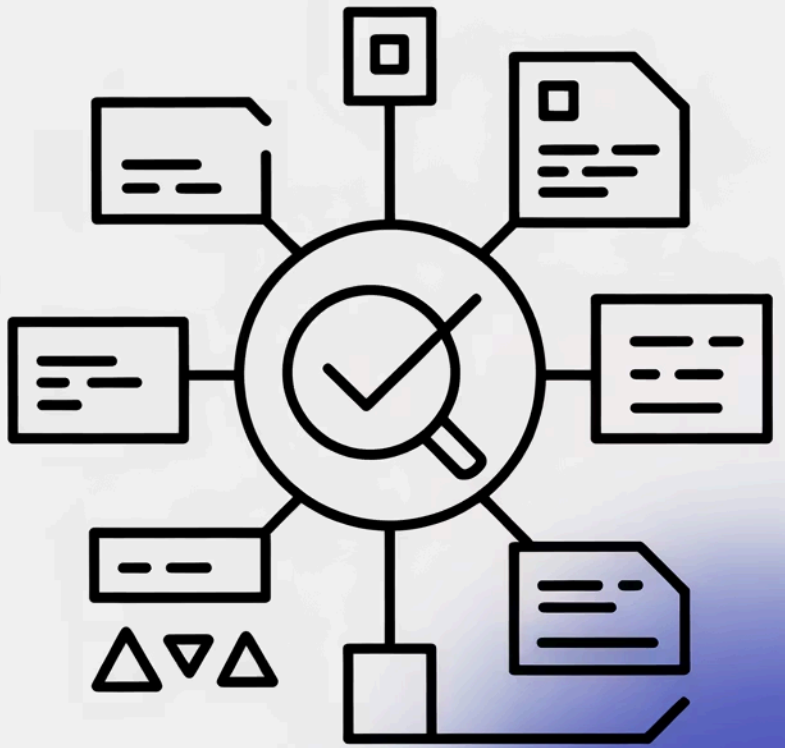
HotpotQA

Multi-hop QA benchmark for evidence reasoning and source verification

3

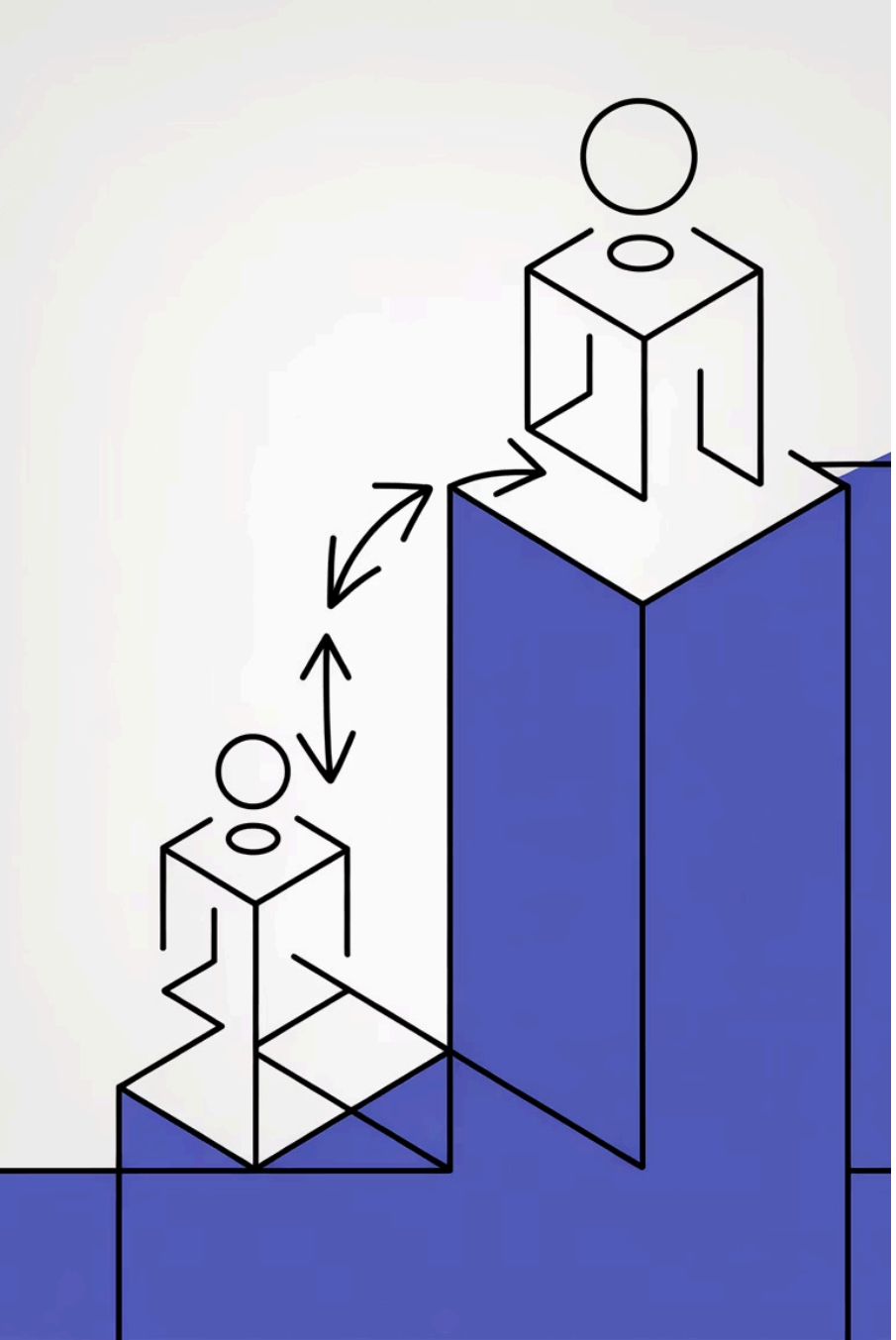
Synthetic Datasets

Custom datasets with deliberately injected false or contradictory passages



References

- Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
- Zhou, Y., et al. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.
- Singal, R., et al. (2024). Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. *FEVER Workshop*, 91-98.



P6. The Apprentice Model

This project explores **knowledge distillation** by training a smaller model to imitate a larger LLM on a specific domain or reasoning task. The aim is to obtain lightweight, domain-specialized "expert" models while analyzing trade-offs between **efficiency, specialization, and generalization**.

1

Core Pipeline

Smaller model trained using outputs or intermediate representations from teacher LLM. Student predictions compared with teacher across benchmarks.

2

Expected Outcomes

Measure trade-offs between model compactness, domain specialisation, and reasoning quality retention.

Methodology

01

Data Collection

Use large LLM to generate or label domain-specific examples, establish as teacher model

02

Model Distillation

Train smaller transformer to reproduce teacher's predictions or reasoning chains

03

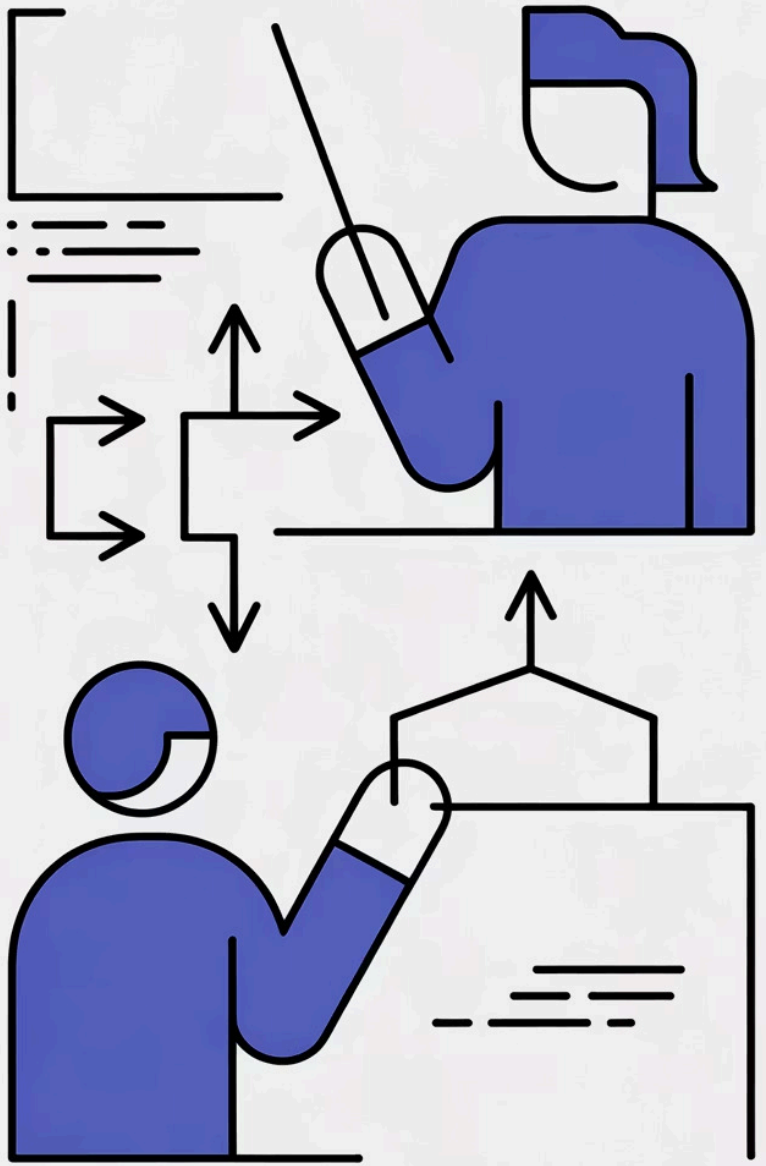
Evaluation

Compare distilled vs. teacher on domain benchmarks: performance drop, interpretability, computational savings

04

Extension

Experiment with multi-teacher distillation or domain adaptation through selective fine-tuning

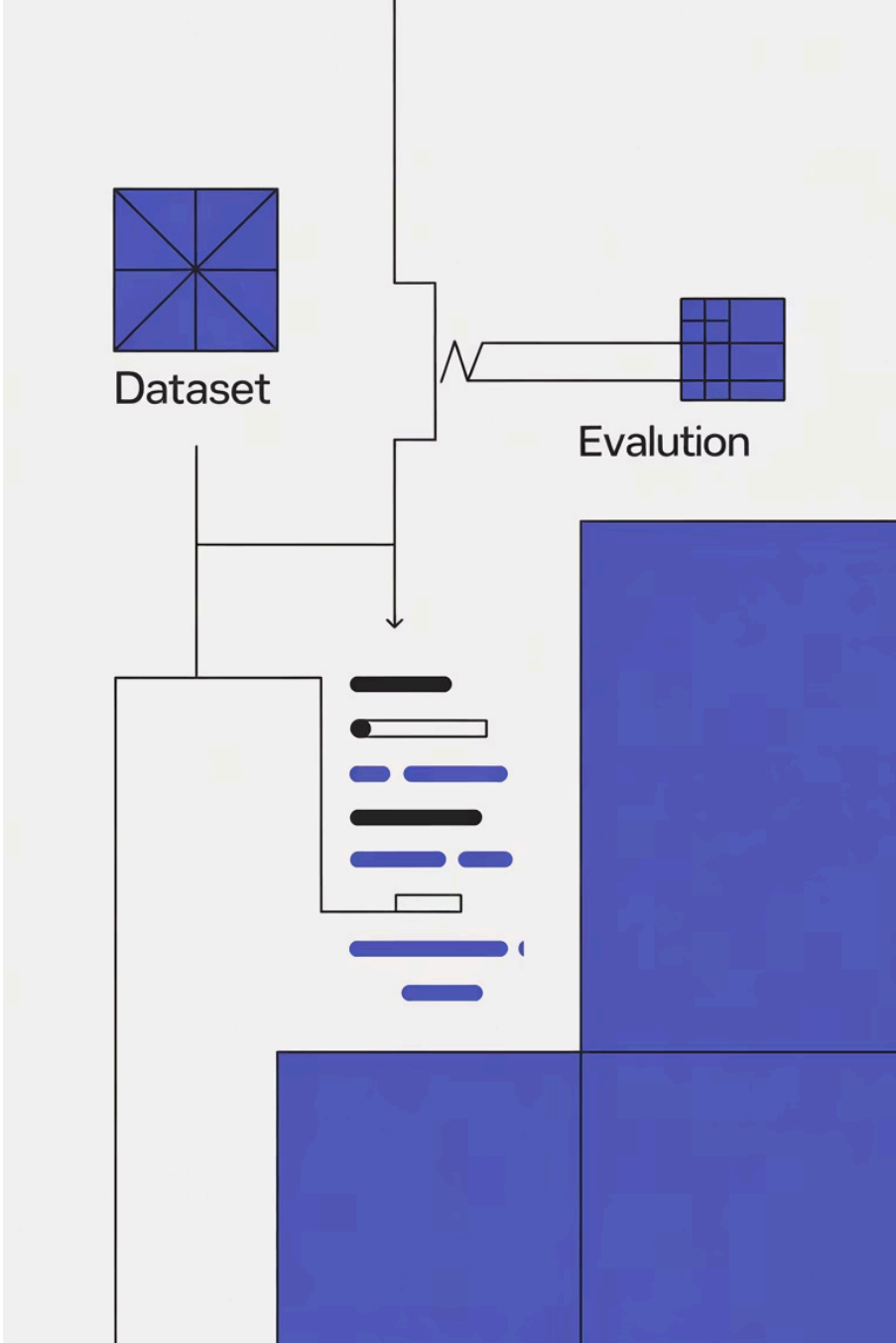


Dataset & References

Dataset: Domain-specific datasets and custom datasets generated by prompting larger models (GPT-4, Claude).

References

- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Ji, G., & Zhu, Z. (2020). Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. *Advances in Neural Information Processing Systems*, 33, 20823-20833.
- Jiao, X., et al. (2020). Tinybert: Distilling bert for natural language understanding. *Findings of EMNLP*, 4163-4174.



P7. Analyzing Thematic Alignment in Scientific Journals

The core objective is to quantitatively assess whether articles published in a specific journal align with its stated **"Aims & Scope"**. Students develop methodology to model the journal's intended focus and compare it against publication content, potentially identifying thematic drift over time or outlier papers.

1

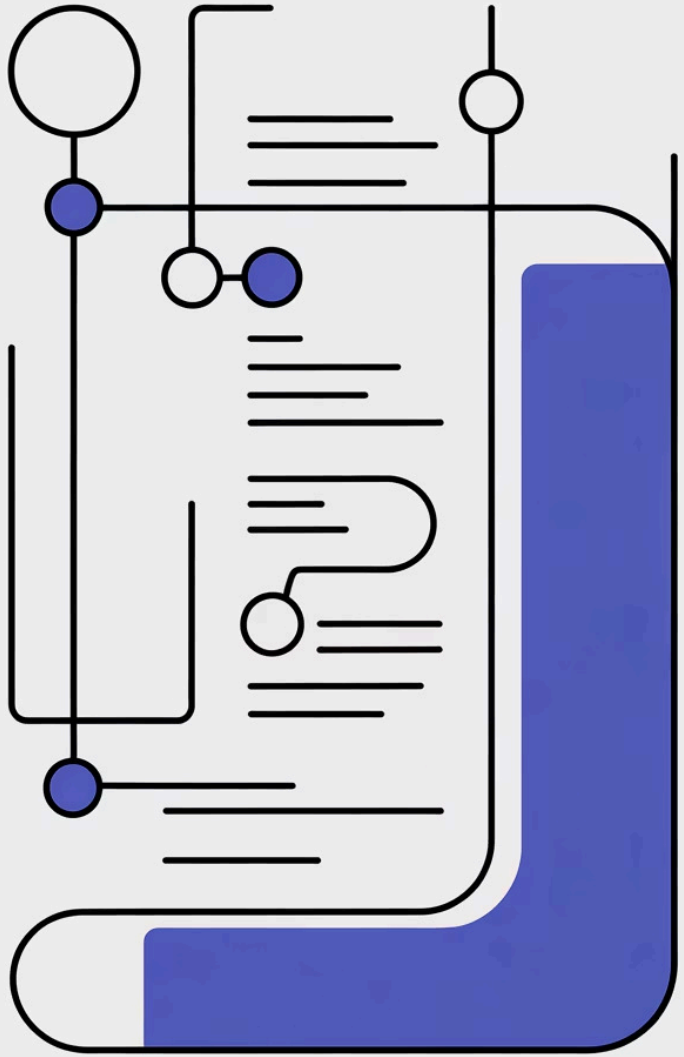
Core Pipeline

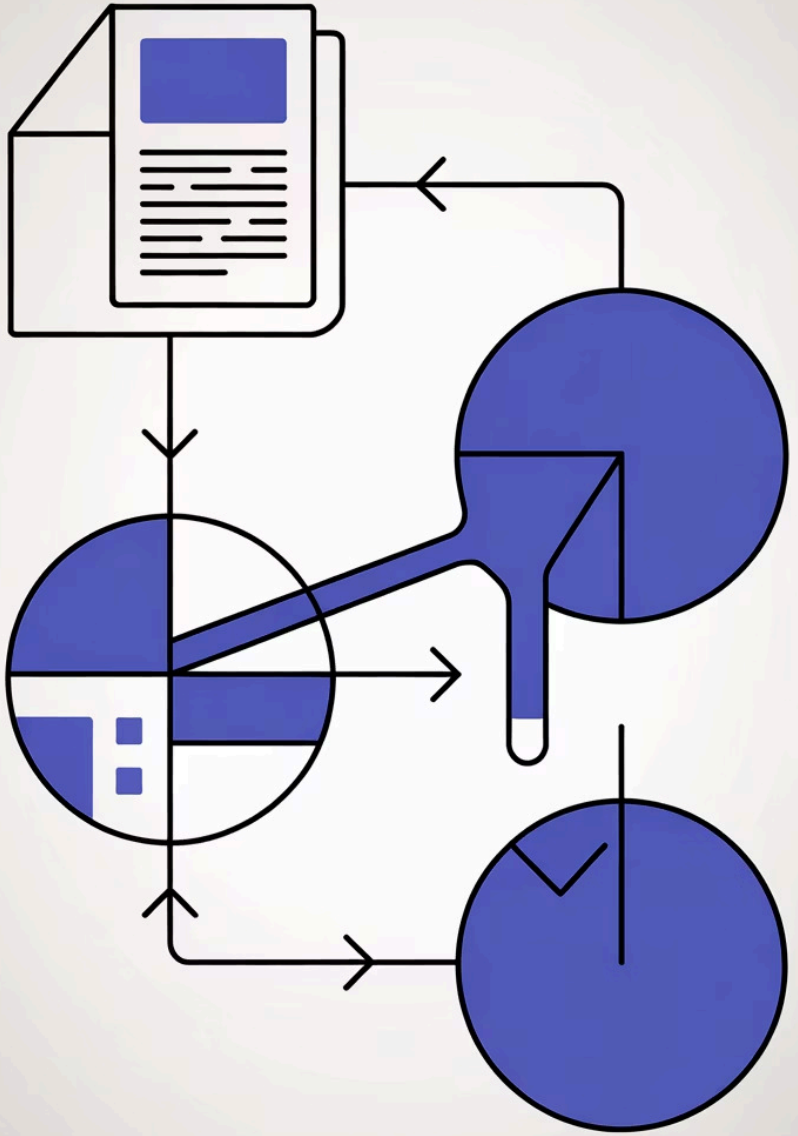
Articles represented as embeddings or topic distributions. "Aims & Scope" serves as thematic reference for alignment scoring.

2

Expected Outcomes

Quantify thematic coherence, identify outlier papers and trends revealing long-term conceptual evolution.





Methodology

Data Curation

Select scientific journal with clearly defined "Aims & Scope" statement as ground truth for intended focus

Content Modeling

Create structured, machine-readable representation capturing core subjects and meaning for both scope and articles

Measure Alignment

Design computational method to measure thematic overlap, generating quantitative alignment scores

Report Findings

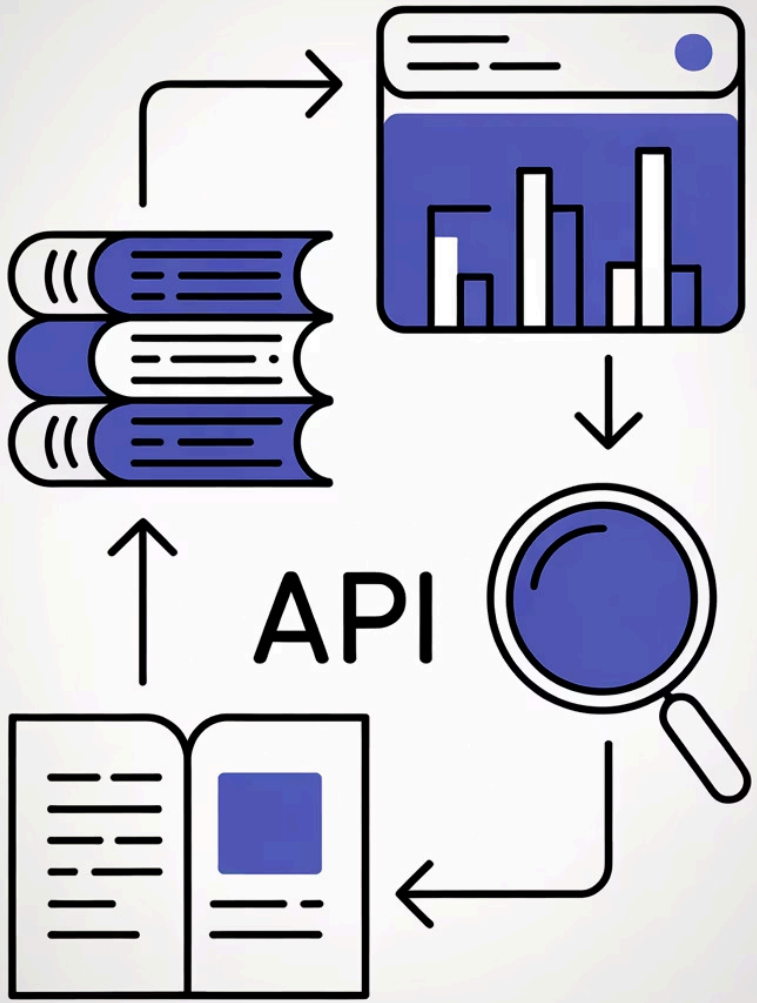
Analyze score distribution, visualize results, detect thematic drift, identify outliers with qualitative validation

Dataset & References

Dataset: Use relevant API (arXiv, Semantic Scholar, PubMed) to collect article abstracts from target journal over 5-10 years.

References

- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP*, 3982-3992.
- Picascia, S., et al. (2025). The Atlas of Data Science Research. *IEEE Access*.
- Hassan-Montero, Y., et al. (2014). Graphical interface of the Scimago Journal and Country Rank. *El profesional de la información*, 23(3).

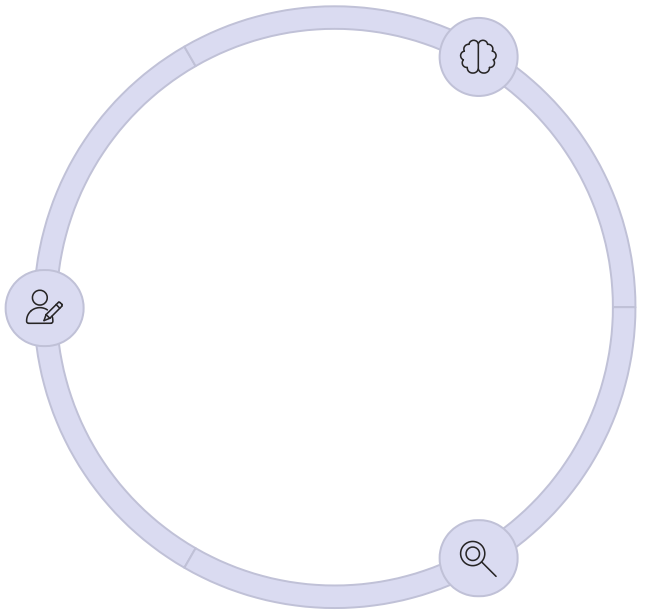


Thematic Cluster 4: Prompt Engineering, Meta-NLP & Instruction Tuning

This cluster examines how the formulation of prompts and fine-tuning procedures shape model behaviour. It investigates **adaptive prompting, meta-learning, and causal inference** to better understand how models internalise instructions, optimise responses, and modify their linguistic and cognitive patterns.

Adaptive Prompting

Self-improving prompt strategies that evolve based on performance feedback



Meta-Learning

Models learning to learn and optimize their own response patterns

Causal Inference

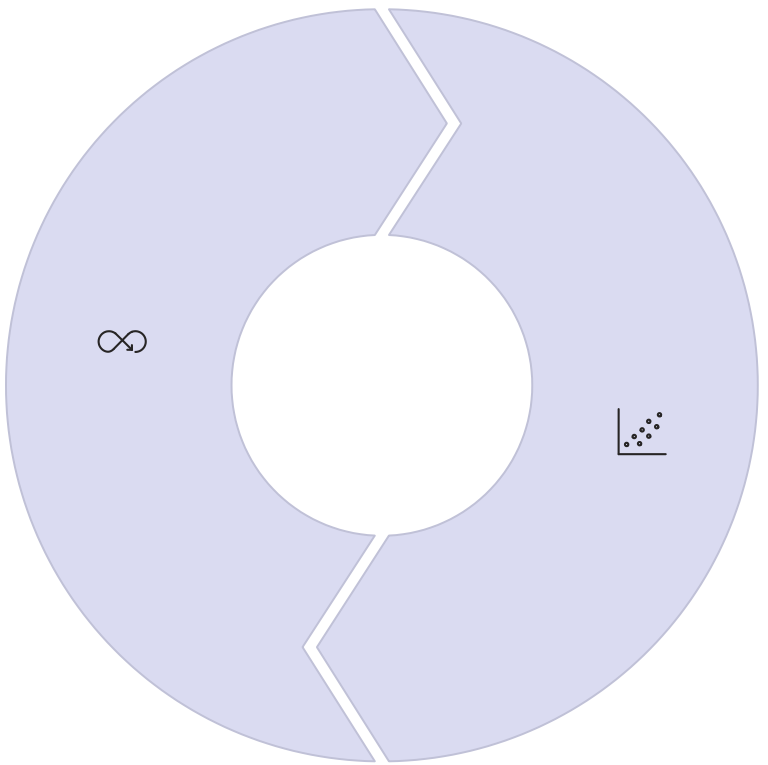
Understanding causal relationships between instructions and behavioral changes

P8. Evolution of a Prompt

This project designs an **automated prompt optimization framework**, where prompts evolve iteratively based on performance feedback. The goal is to study **prompt sensitivity** and develop adaptive, self-improving strategies that balance human control and model autonomy.

Core Pipeline

Feedback-driven optimization loop iteratively refines prompts using performance metrics or LLM self-evaluation



Expected Outcomes

Demonstrate how prompt evolution influences task performance and linguistic structure, revealing emergent meta-learning

Methodology

1 **Task Selection**

Choose concrete NLP task (summarization, reasoning, or classification) for optimization target

2 **Prompt Optimization Loop**

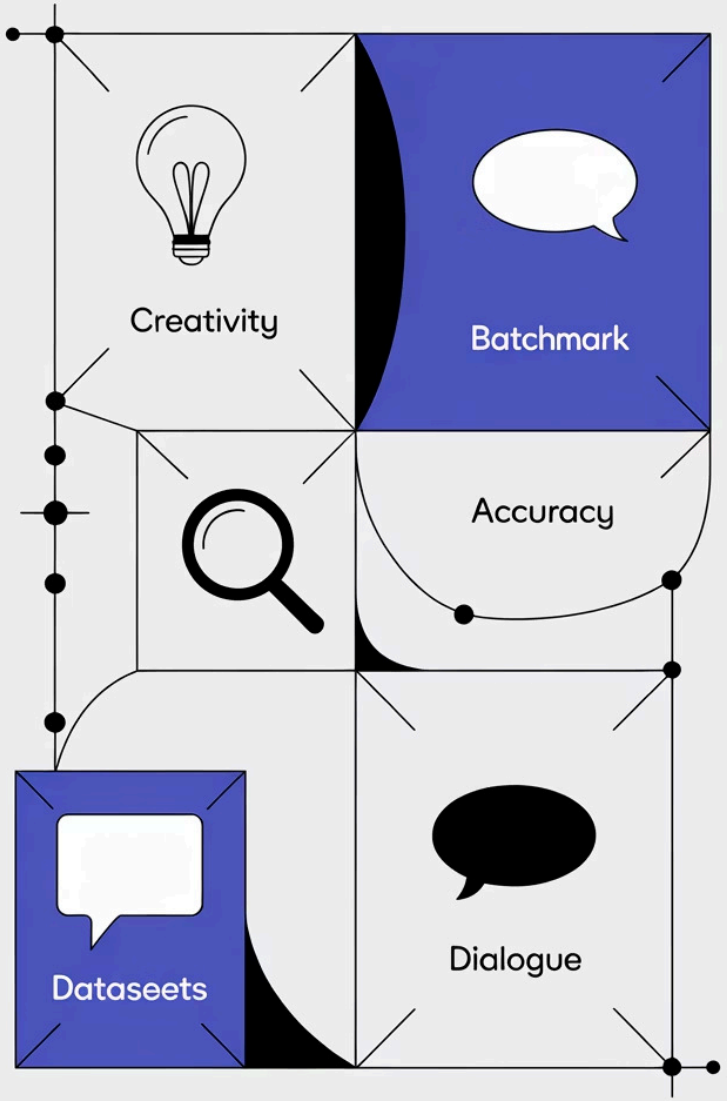
Implement iterative prompt mutation via scoring (accuracy, coherence, or BLEU metrics)

3 **Feedback Mechanism**

Use external metrics or LLM self-evaluation to guide evolutionary process

4 **Comparison**

Benchmark adaptive prompting against manually engineered baselines for effectiveness



Dataset & References

BIG-bench

Challenging reasoning tasks for comprehensive prompt evaluation
https://github.com/google/BIG-bench

Natural-Instructions

Language instructions dataset for LLM prompt optimization
https://github.com/allenai/natural-instructions

References

- Schulhoff, S., et al. (2024). The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.
- Ye, Q., et al. (2024). Prompt engineering a prompt engineer. *Findings of ACL*, 355-385.
- Hsieh, C. J., et al. (2024). Automatic engineering of long prompts. *Findings of ACL*, 10672-10685.

P9. Measuring Total Causal Effects of Instruction Tuning

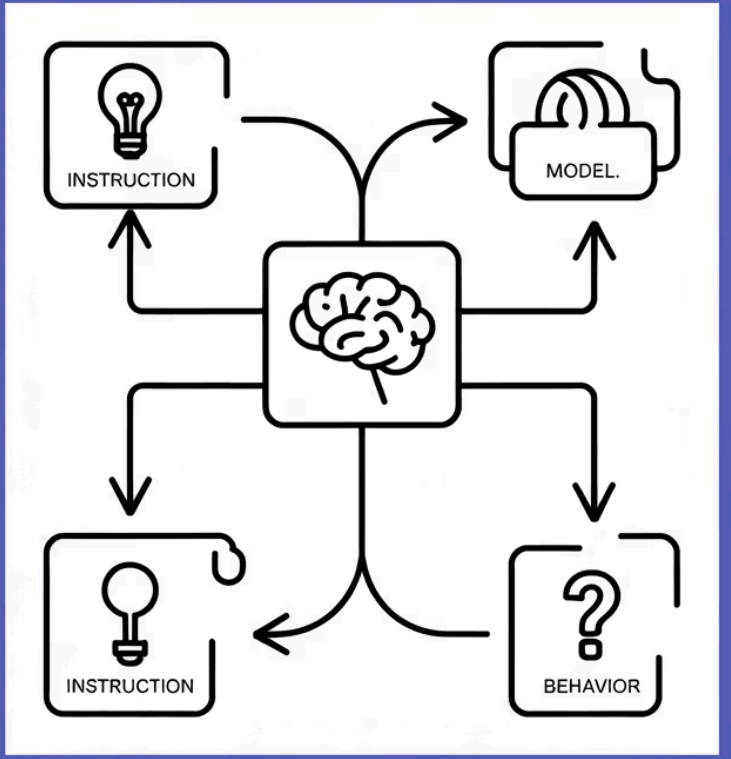
The process of **"instruction tuning"** is a critical step in creating helpful and safe AI assistants. This project frames instruction tuning as a "treatment" and aims to measure its **total causal effect** on a range of model variables, moving beyond simple performance metrics to quantify both intended and unintended changes.

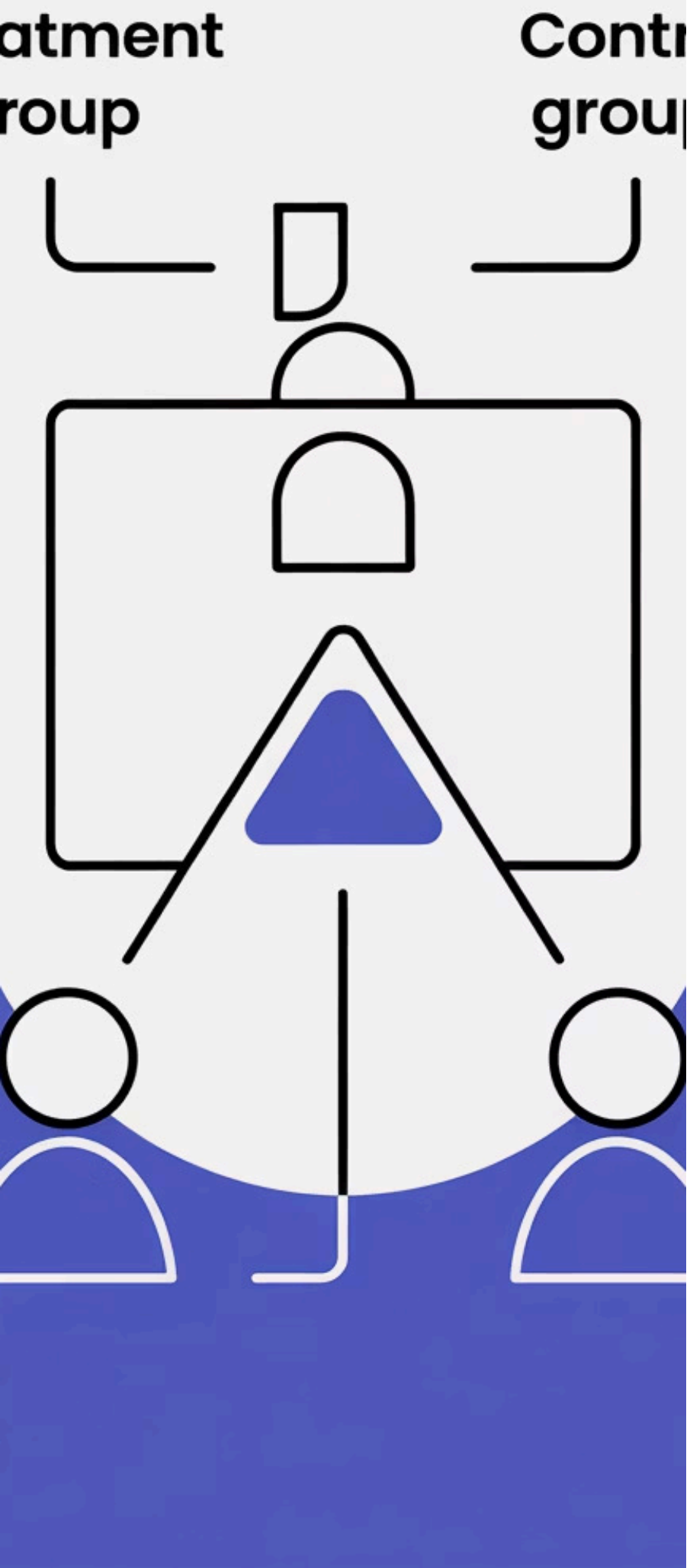
Core Pipeline

Two model versions (base and instruction-tuned) evaluated on predefined variables using causal inference techniques

Expected Outcomes

Quantitative measures of how instruction tuning alters reasoning depth, bias expression, and linguistic features





Methodology & Variables

01

Causal Framework Definition

Define outcome variable, treatment (instruction tuning), treatment/control groups, and total effect measurement

02

Model & Variable Selection

Select open-source model family with base and instruction-tuned versions, choose specific measurable variable

03

Dataset & Prompt Design

Standardized prompts from academic benchmarks or custom-designed for controlled experimental environment

04

Experiments & Analysis

Run prompts through both models, quantify outcomes, calculate total effect with statistical significance testing

Potential Variables for Study

Sycophancy

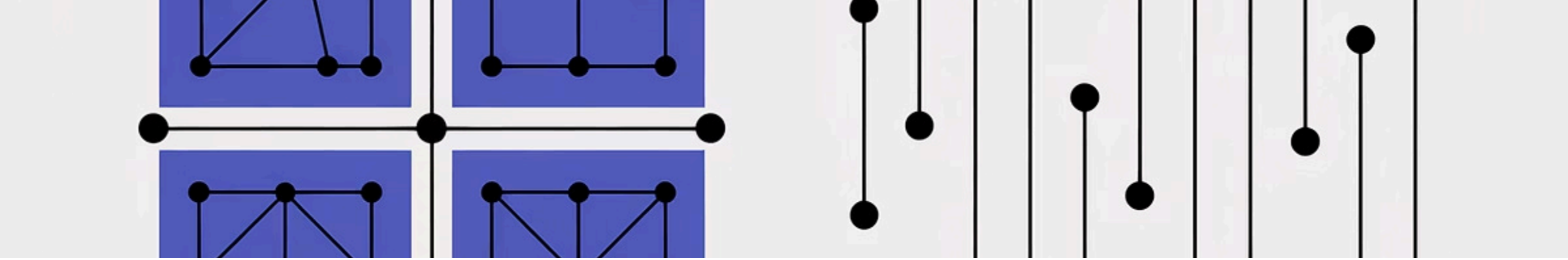
Tendency to agree with user's premise, even if factually incorrect

Lexical Complexity

Vocabulary sophistication measured by Flesch-Kincaid grade level

Logical Reasoning

Performance on standardized logical puzzles or benchmarks



Dataset & References

Dataset: Existing academic benchmarks for reasoning, toxicity, or bias, or custom-designed controlled experimental environments.

References

- Feder, A., et al. (2022). Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10, 1138–1158.
- Vig, J., et al. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. *Advances in Neural Information Processing Systems*, 33, 12388–12401.
- Faulborn, M., et al. (2025). Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models. *ACL*, 31684–31704.

Thematic Cluster 5: Bias, Ethics & Cultural Intelligence

This cluster addresses the **ethical, social, and cultural dimensions** of LLM behaviour. Projects investigate moral alignment, value consistency, and cultural representation, studying how models interpret and express ethical stances or reproduce socio-linguistic biases.

Cultural Intelligence

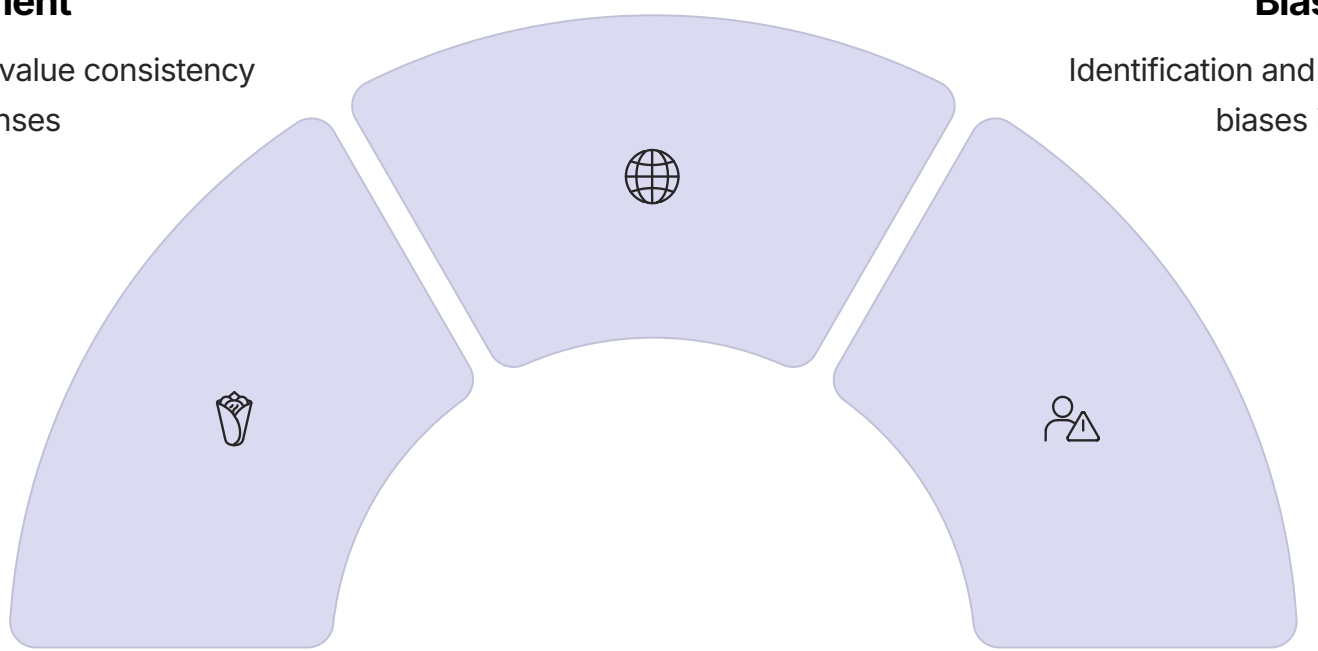
Cross-cultural understanding and representation in language models

Moral Alignment

Ethical decision-making and value consistency in model responses

Bias Detection

Identification and analysis of socio-linguistic biases in model outputs



P10. Right, Wrong, and Everything in Between

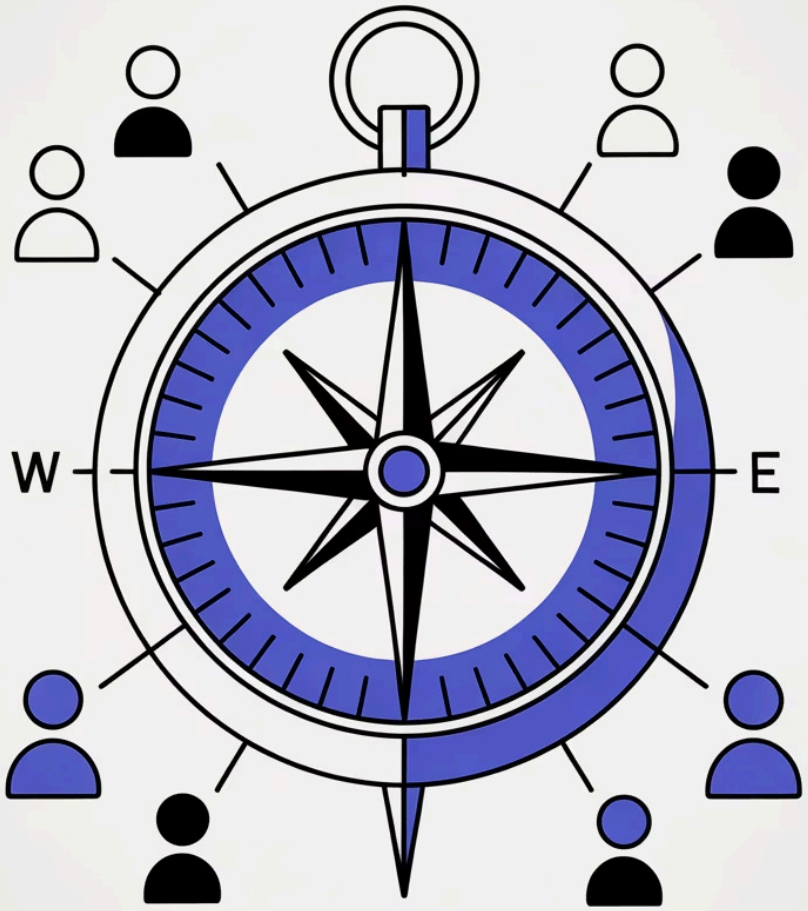
This project examines how LLMs respond to **morally ambiguous or ethically charged prompts**, analyzing *when, how,* and *why* models refuse, reframe, or justify their responses. The aim is to study the intersection between **moral alignment** and **linguistic pragmatics**.

Core Pipeline

Ethically ambiguous prompts curated and submitted to multiple LLMs. Responses analyzed linguistically and semantically.

Expected Outcomes

Reveal how ethical alignment manifests in language, showing navigation of moral grey areas and safety constraints.



Methodology

Scenario Construction

Design ethically sensitive dialogues: fairness dilemmas, medical triage, privacy conflicts

Response Analysis

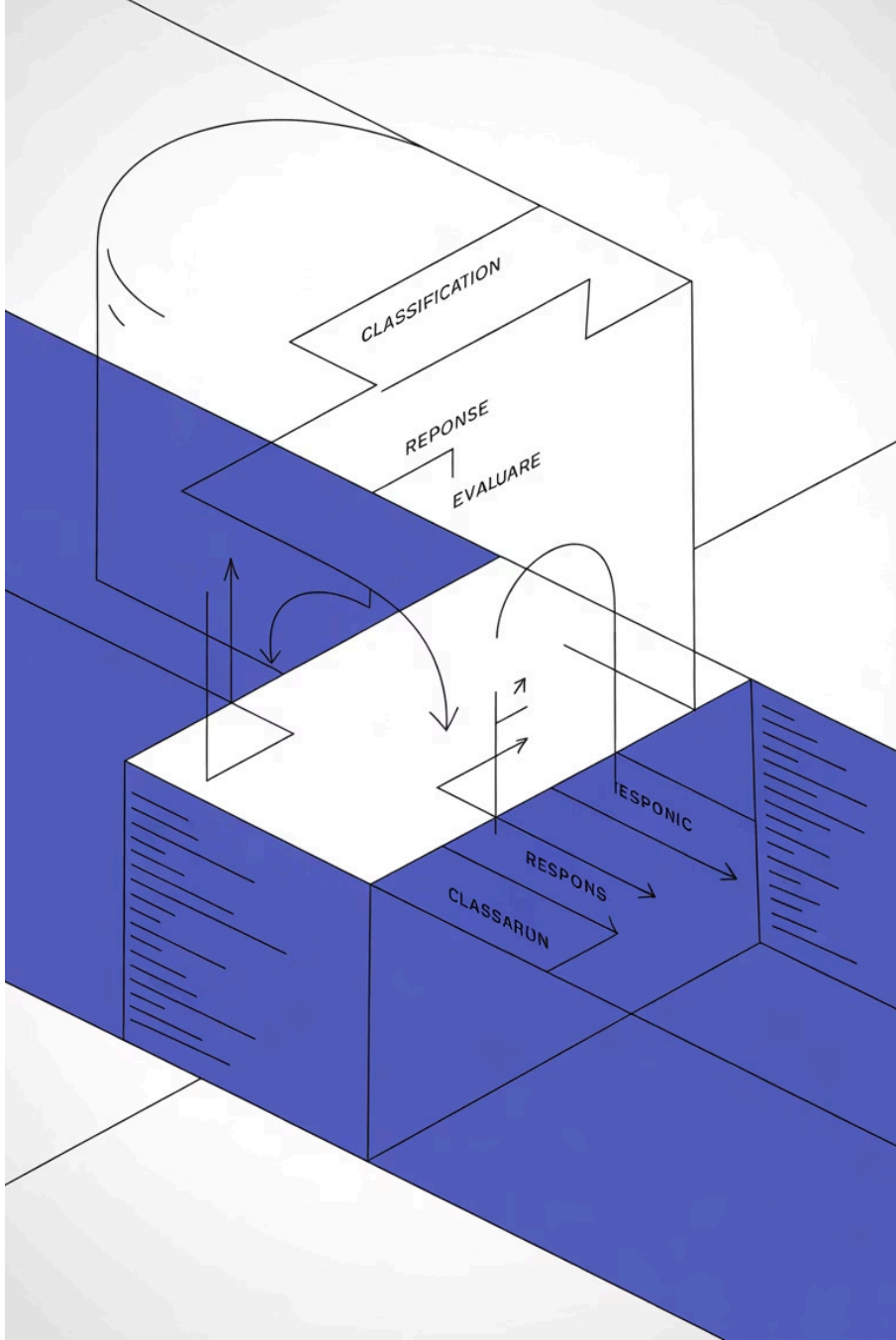
Collect outputs from different LLMs, classify by strategy: refusal, justification, reframing

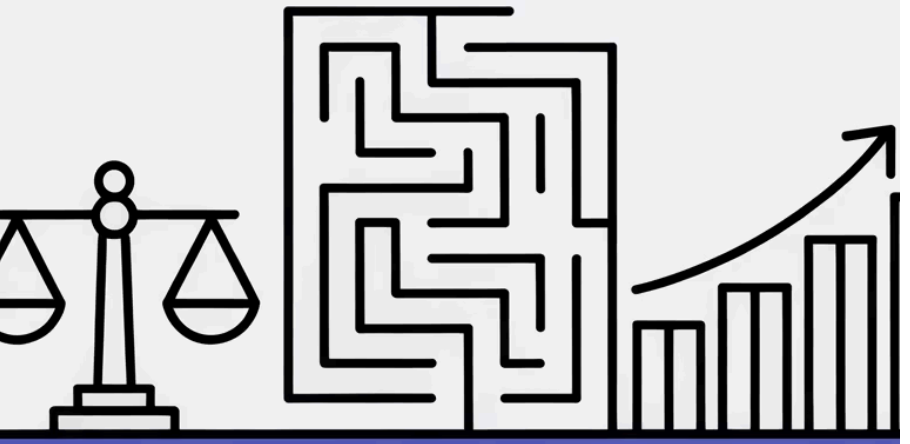
Quantitative Evaluation

Measure response diversity, moral consistency, and sentiment balance across models

Qualitative Analysis

Compare linguistic markers: modality, politeness, uncertainty across models and contexts





Dataset & References

1

ETHICS Dataset

Comprehensive benchmarks for moral reasoning and ethical decision-making evaluation

2

Custom Prompts

Reflecting social or political dilemmas tailored to research objectives

References

- Hendrycks, D., et al. (2020). Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.
- Forbes, M., et al. (2020). Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.
- Bonagiri, V. K., et al. (2024). Sage: Evaluating moral consistency in large language models. *arXiv preprint arXiv:2402.13709*.

P11. Measuring Causal Effects of Prompting Strategies

This project applies a **causal framework** to measure the total effect of different prompting strategies on a model's linguistic and cultural expression. The core idea treats the prompt's formulation as a "treatment" and measures its impact on model response, comparing standard prompts, cultural personas, and native language prompts.

1

Standard Prompt

No references to roles or languages (control group)

2

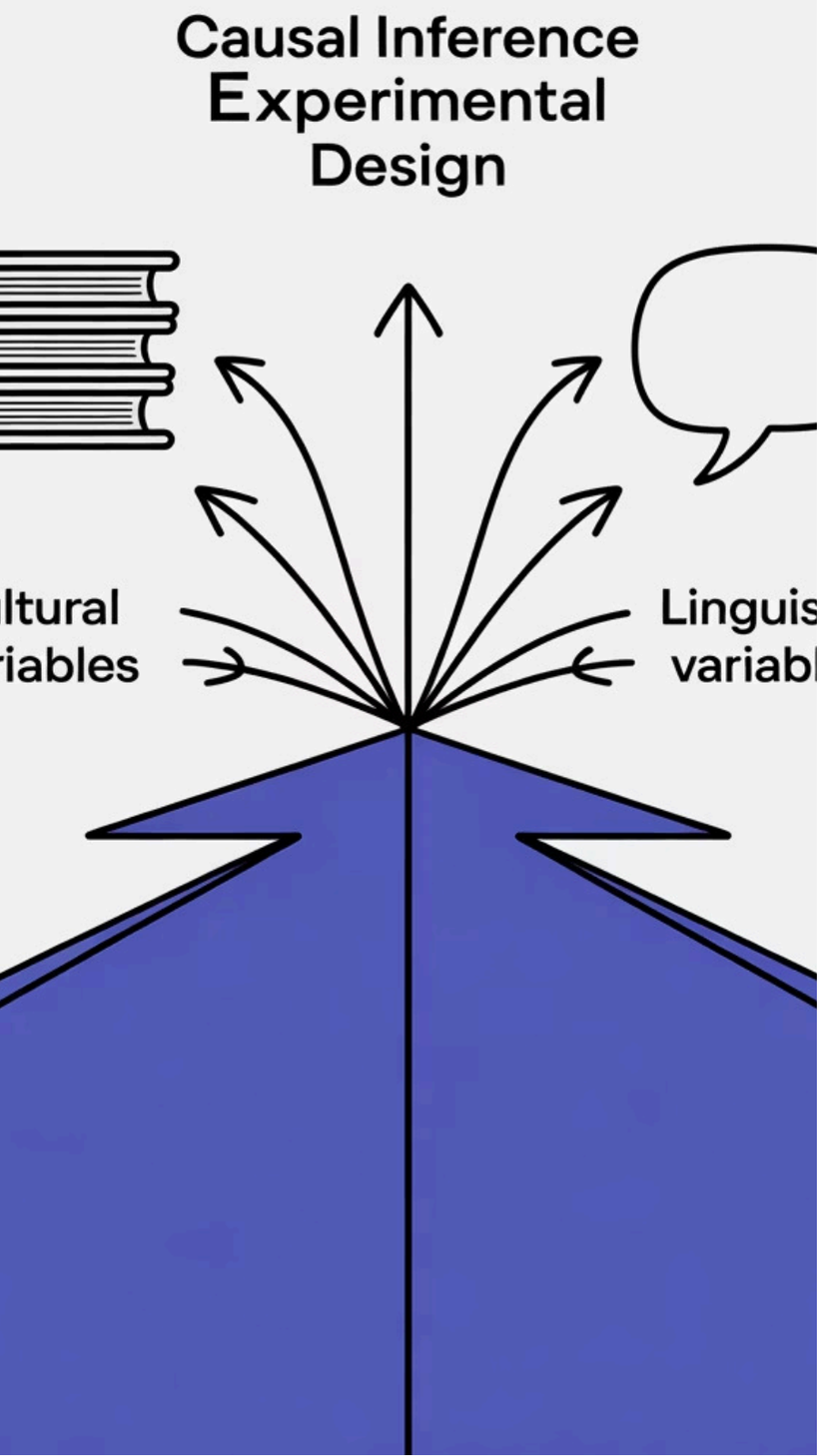
Cultural Persona

Explicit cultural role assignment (e.g., "As a person from Nigeria...")

3

Native Language

Prompts written directly in corresponding native language



Methodology & Variables

Causal Framework

Define outcome variable, treatment (prompting strategy), treatment/control groups, total effect measurement

Model Selection

Select open-source multilingual model, choose specific measurable variable for primary outcome

Experiments

Run prompt sets through LLM, collect responses, score outputs based on chosen variable

Analysis

Calculate total effect by comparing average metric scores, apply statistical significance tests

Potential Variables for Study

1

Stereotype Prevalence

Frequency of stereotypical vs. nuanced descriptions

2

Cultural Knowledge

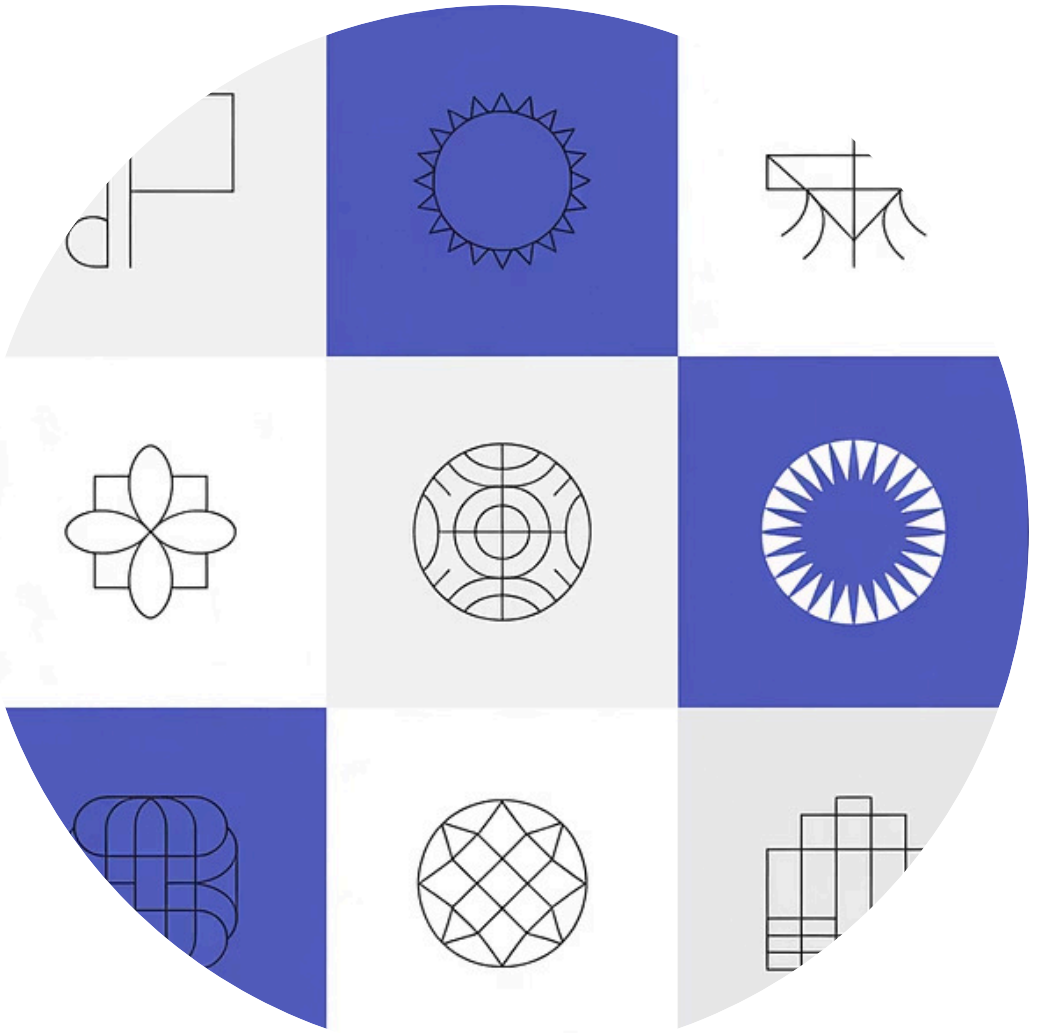
Accuracy of cultural facts and norms

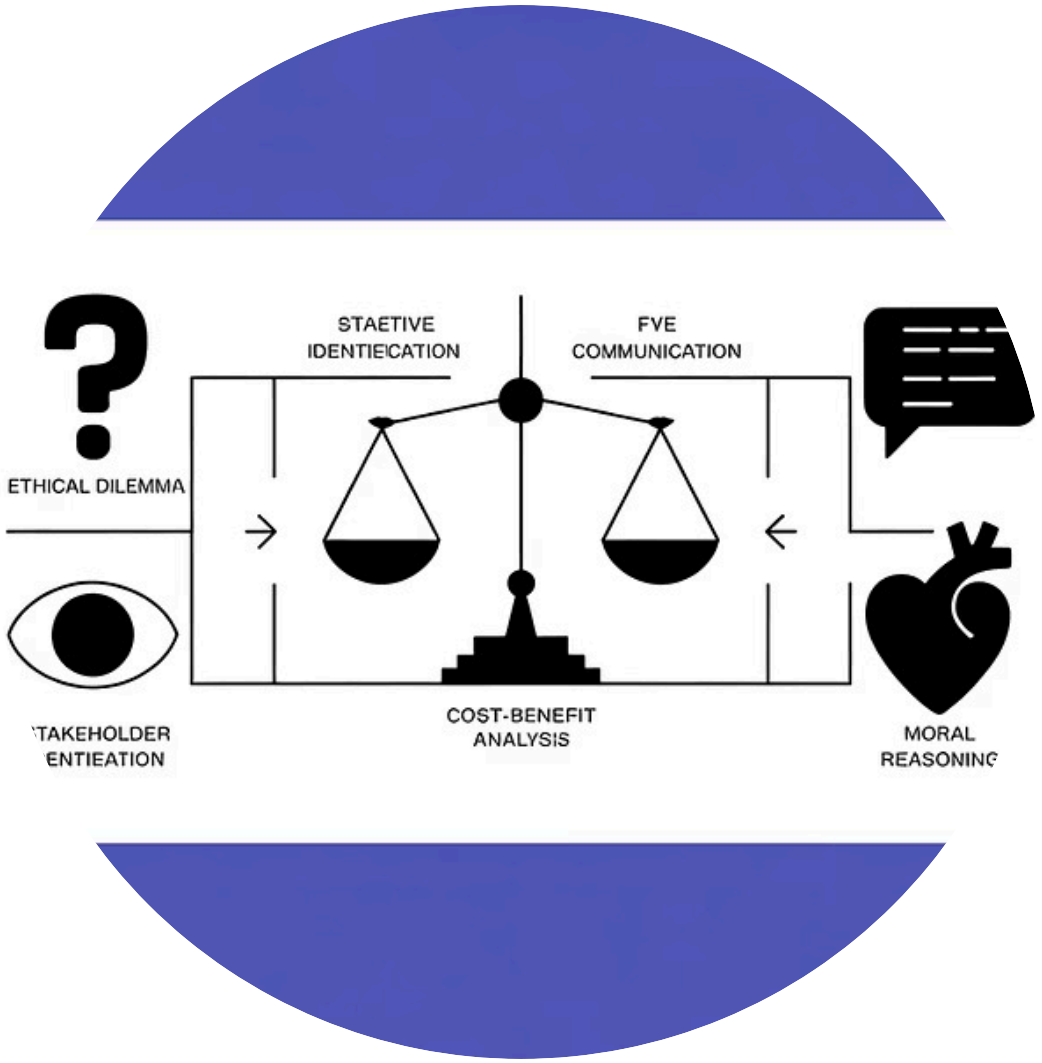
3

Value Expression

Cultural dimensions like individualism vs. collectivism

Dataset & References





BLEnD Benchmark

LLMs on Everyday Knowledge in Diverse Cultures and Languages (NeurIPS 2024)

ETHICS Dataset

Benchmarks for moral reasoning across cultural contexts

References

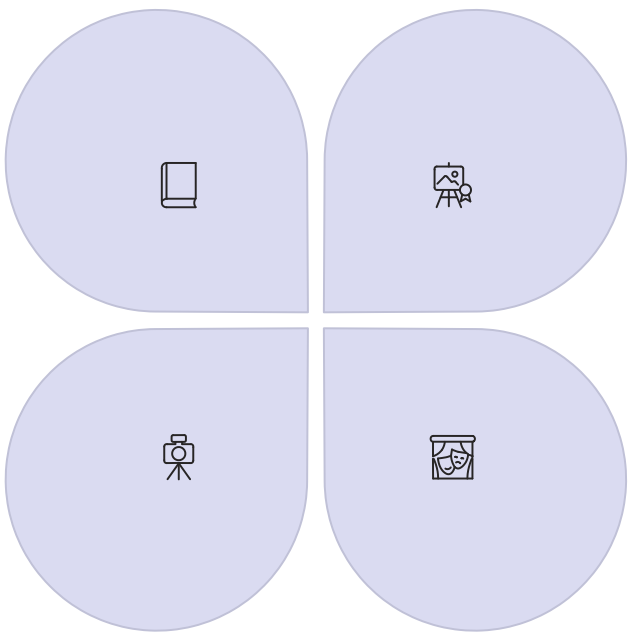
- Feder, A., et al. (2022). Causal Inference in Natural Language Processing. *Transactions of the Association for Computational Linguistics*, 10, 1138–1158.
- Vig, J., et al. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. *Advances in Neural Information Processing Systems*, 33, 12388–12401.
- Faulborn, M., et al. (2025). Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models. *ACL*, 31684–31704.

Thematic Cluster 6: Creativity, Narrative & Style

This cluster explores the **expressive and generative capabilities** of LLMs across creative, narrative, and stylistic domains. It examines how models emulate storytelling, produce cultural artefacts, and develop distinctive stylistic identities that blur the boundaries between computation and creativity.

Narrative Archetypes

Cross-cultural storytelling patterns and universal narrative structures



Stylistic Identity

Distinctive aesthetic voices and creative fingerprints in generated content

Computational Creativity

Emergence of creative expression from probabilistic generation processes

Cultural Translation

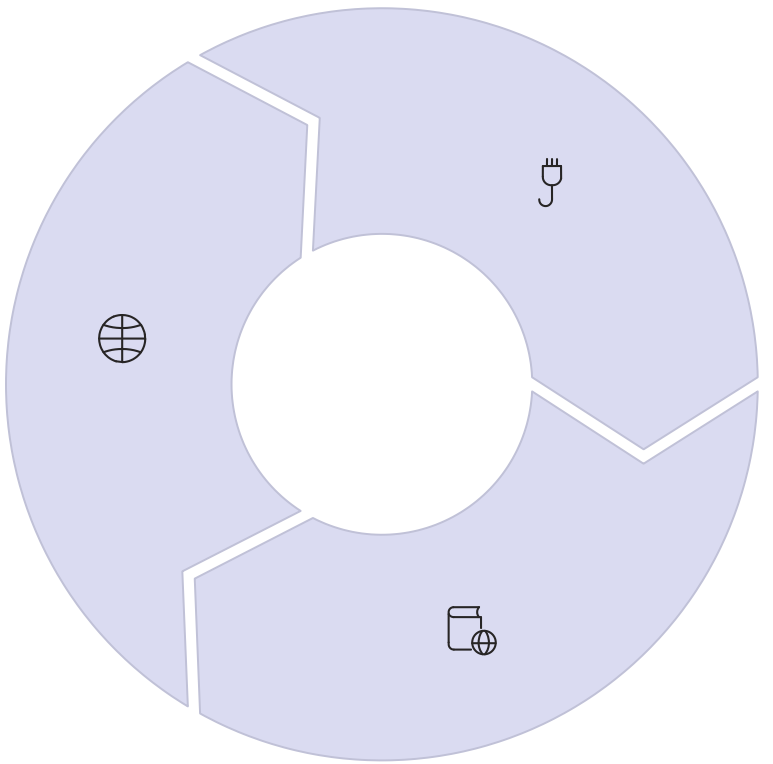
Adaptation of myths and stories across different symbolic systems

P12. Stories We Tell (and the Machines Retell)

This project explores how **Large Language Models (LLMs)** represent and reproduce **narrative archetypes across different cultures and traditions**. By combining computational narratology, linguistic analysis, and network science, the project investigates whether storytelling structures appear as **universal cognitive patterns** or are culturally bound.

Cross-Cultural Analysis

Compare narrative logic of human-written texts with LLM-generated stories

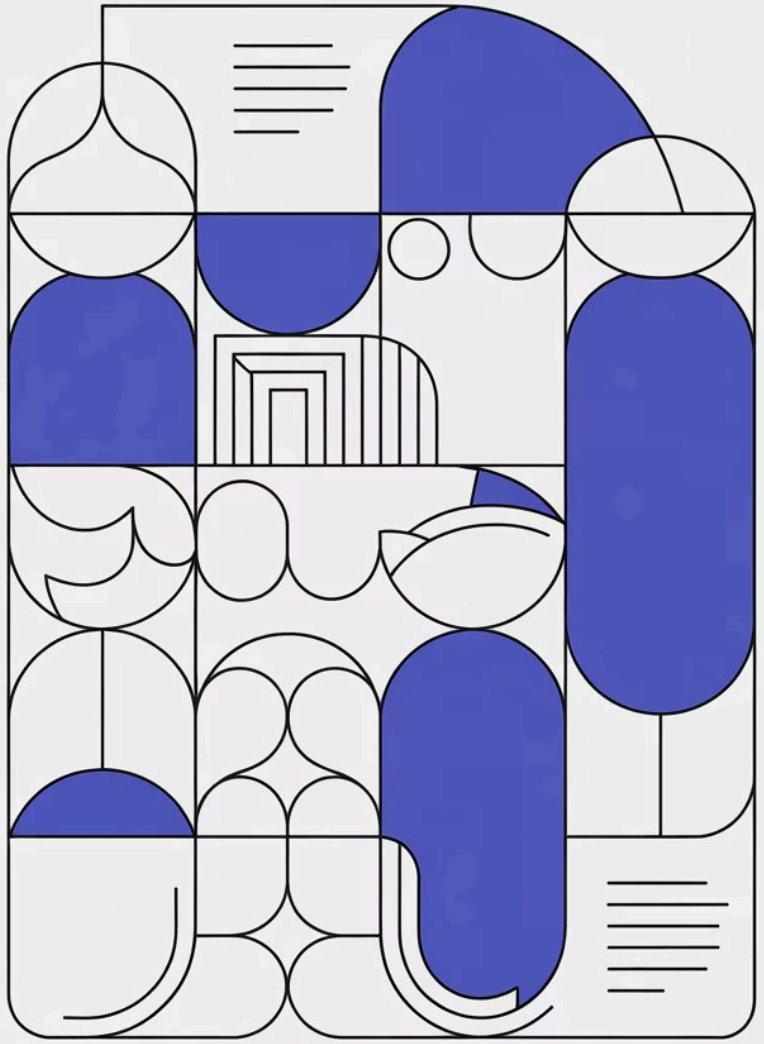


Structural Modeling

Model narratives as structured event graphs revealing deep patterns

Cultural Translation

Test LLMs as cultural translators adapting myths across symbolic systems



Methodology

01

Corpus Design

Build multilingual corpus of myths, folktales from at least three cultural areas (European, East Asian, African, Indigenous)

02

Narrative Structure Modeling

Extract narrative entities and relations, represent as narrative graphs, apply clustering for recurring structures

03

Cross-Cultural Comparison

Ask LLMs to generate retellings in different cultural contexts, evaluate structural pattern preservation

04

Quantitative Analysis

Use graph edit distance, motif overlap, semantic role alignment to quantify structural similarity

Dataset & References

LitBank

Annotated dataset of 100 English-language fiction works for computational humanities tasks

World Folktale Corpora

Scraped from web or collected from various cultural traditions and sources

References

- Valls-Vargas, J., et al. (2016). Predicting proppian narrative functions from stories in natural language. *AAAI Conference on AI and Interactive Digital Entertainment*, 107-113.
- Kumaran, V., et al. (2024). NARRATIVEGENIE: generating narrative beats and dynamic storytelling with large language models. *AAAI Conference on AI and Interactive Digital Entertainment*, 76-86.
- Ranade, P., et al. (2022). Computational understanding of narratives: A survey. *IEEE Access*, 10, 101575-101594.

P13. The Aesthetics of Generation

This project investigates whether **Large Language Models (LLMs)** exhibit a distinctive **aesthetic or stylistic identity** in their generated outputs — a recognizable "voice" that persists across tasks, topics, and prompts. Beyond surface-level metrics, the project asks: *do models have style?*

Core Pipeline

Texts generated by different LLMs under uniform prompts, analyzed for stylistic and lexical variation using stylometric features

Expected Outcomes

Characterize stylistic identity of major LLMs, contributing to emerging field of AI stylistics and authorship attribution

Methodology

1

Corpus Generation

Generate balanced corpus across genres (narration, argumentation, dialogue, description) with constant prompts

2

Stylometric Analysis

Extract quantitative features: lexical diversity, sentence length, POS ratios, syntactic depth, punctuation frequency

3

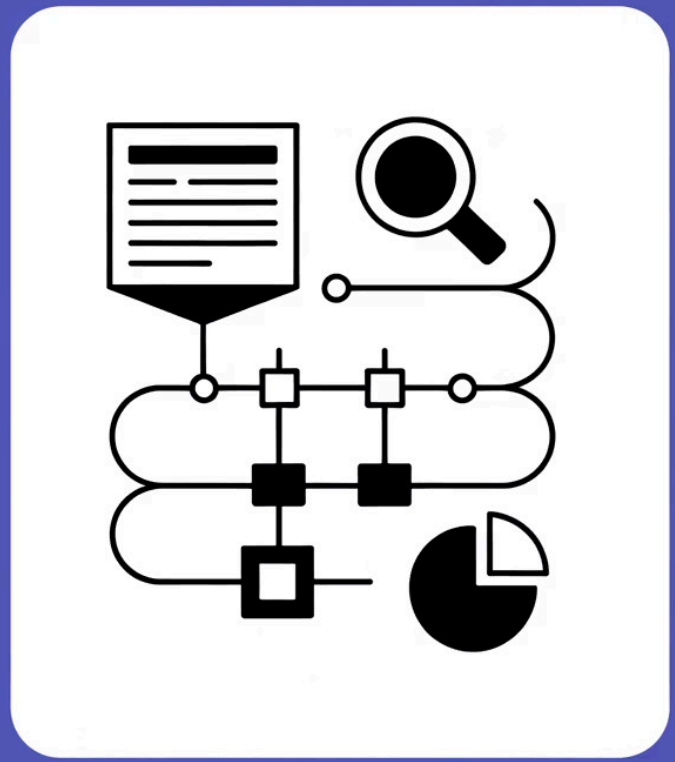
Comparative Evaluation

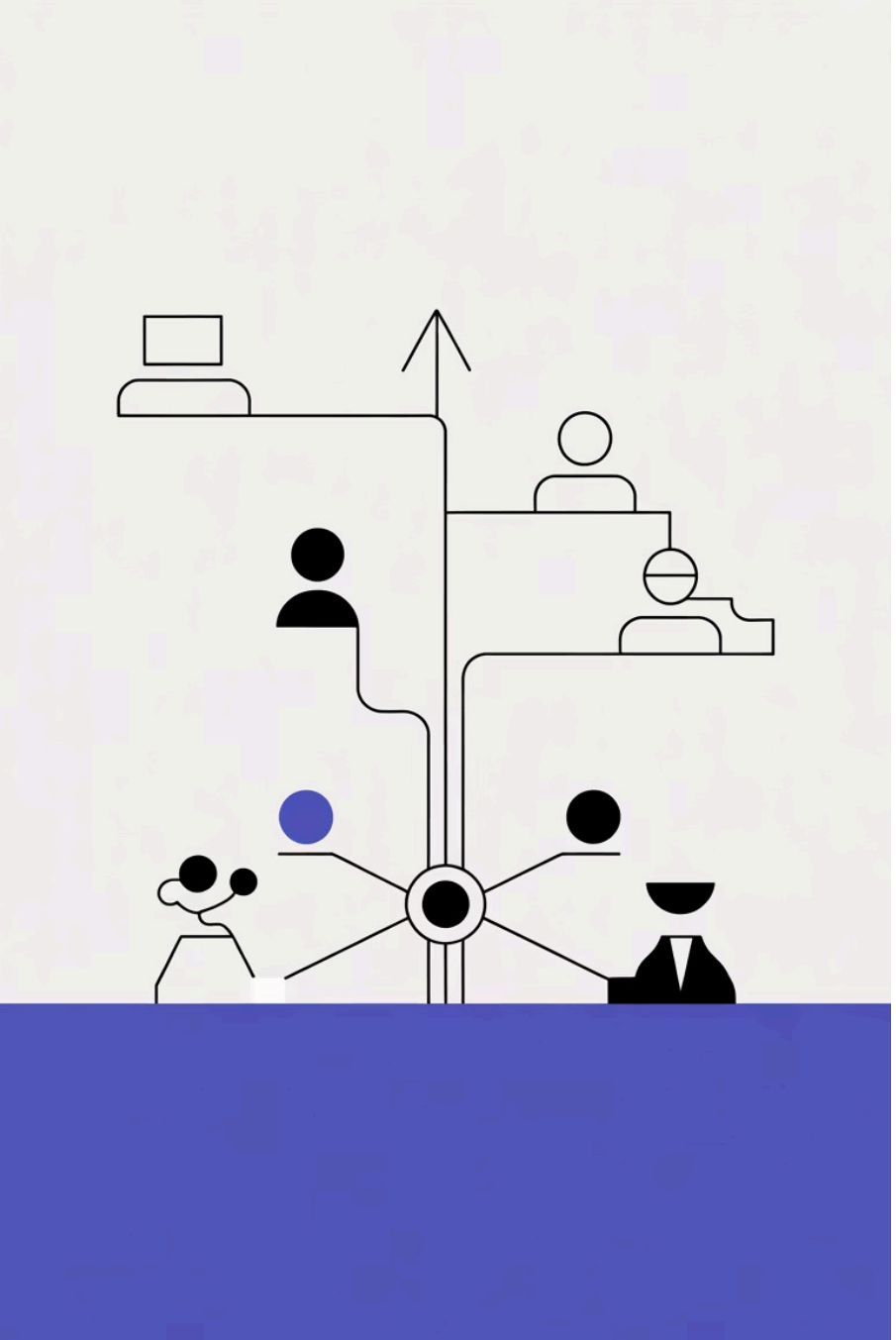
Use authorship attribution tasks to test model identification, evaluate stylistic signature robustness

4

Style Transfer

Experiment with style blending and cross-model paraphrasing for aesthetic transformation analysis





Dataset & References

Dataset: Texts generated from open and closed LLMs (GPT, Claude, LLaMA, Mistral) across multiple genres and prompts.

References

- Opara, C. (2024). StyloAI: Distinguishing AI-generated content with stylometric analysis. *International Conference on AI in Education*, 105-114.
- Kumarage, T., et al. (2023). Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.
- Zellers, R., et al. (2019). Defending against neural fake news. *Advances in neural information processing systems*, 32.
- Okulska, I., et al. (2023). Stylometrix: An open-source multilingual tool for representing stylometric vectors. *arXiv preprint arXiv:2309.12810*.

Thematic Cluster 7: Explainability, Visualization & Model Understanding

This cluster investigates **interpretability and transparency** in language models. Projects aim to visualise and explain the internal dynamics of model reasoning, uncovering how abstract representations of meaning and attention evolve within high-dimensional linguistic spaces.

1

Transparency

Making model reasoning processes visible and interpretable

2

Internal Dynamics

Understanding attention flow, token importance, and activation patterns

3

Semantic Mapping

Visualizing abstract representations in high-dimensional linguistic spaces

4

Interactive Exploration

Dynamic environments for hands-on model interpretation and analysis

P14. Transparent Minds

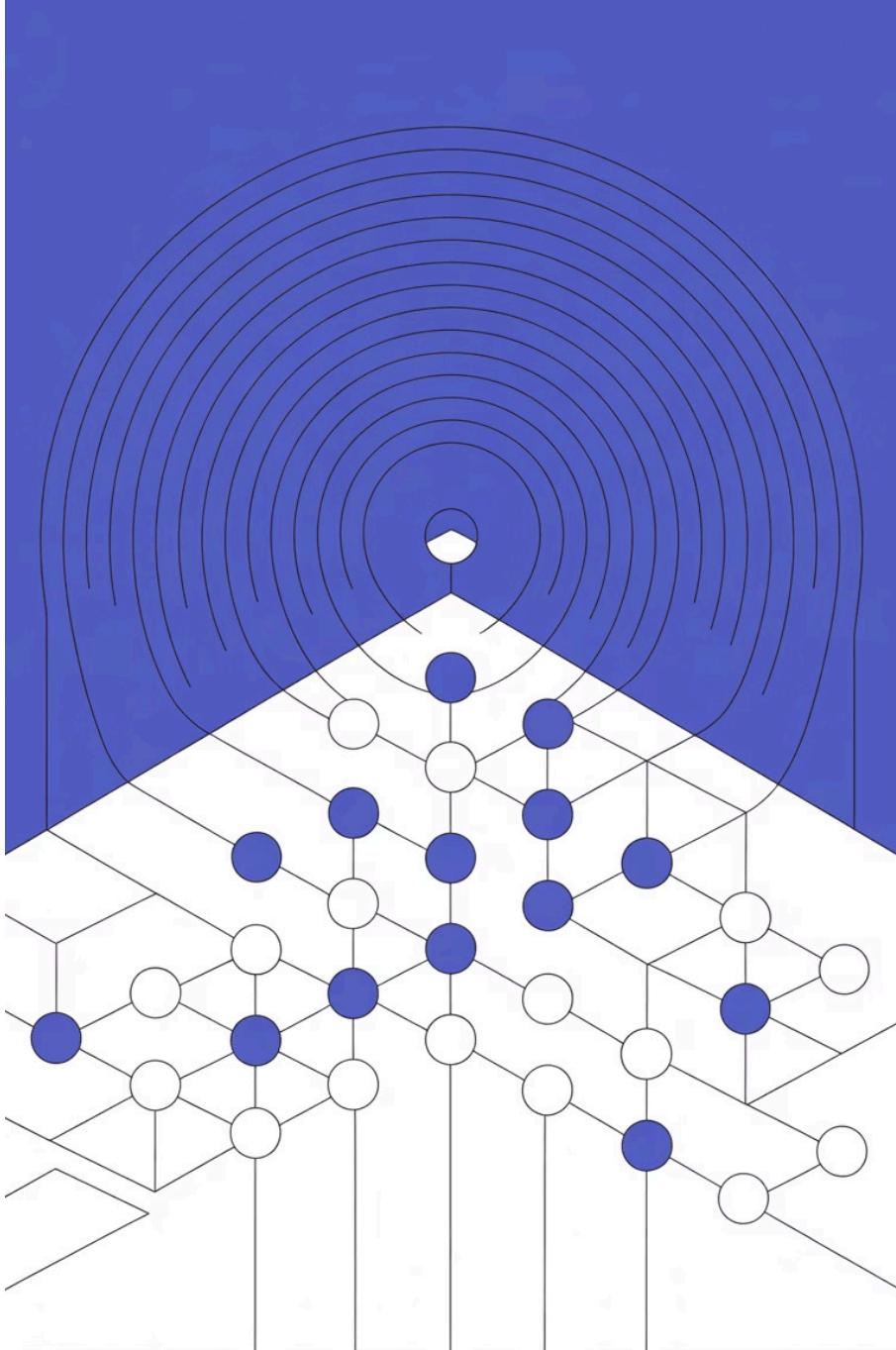
This project aims to design and implement an **interactive toolkit for visualizing and interpreting the reasoning processes** of transformer-based language models. While most explainability studies remain abstract or static, this project focuses on creating a **hands-on, dynamic environment** where users can *see* and *experiment with* the inner mechanics of LLMs.

Core Pipeline

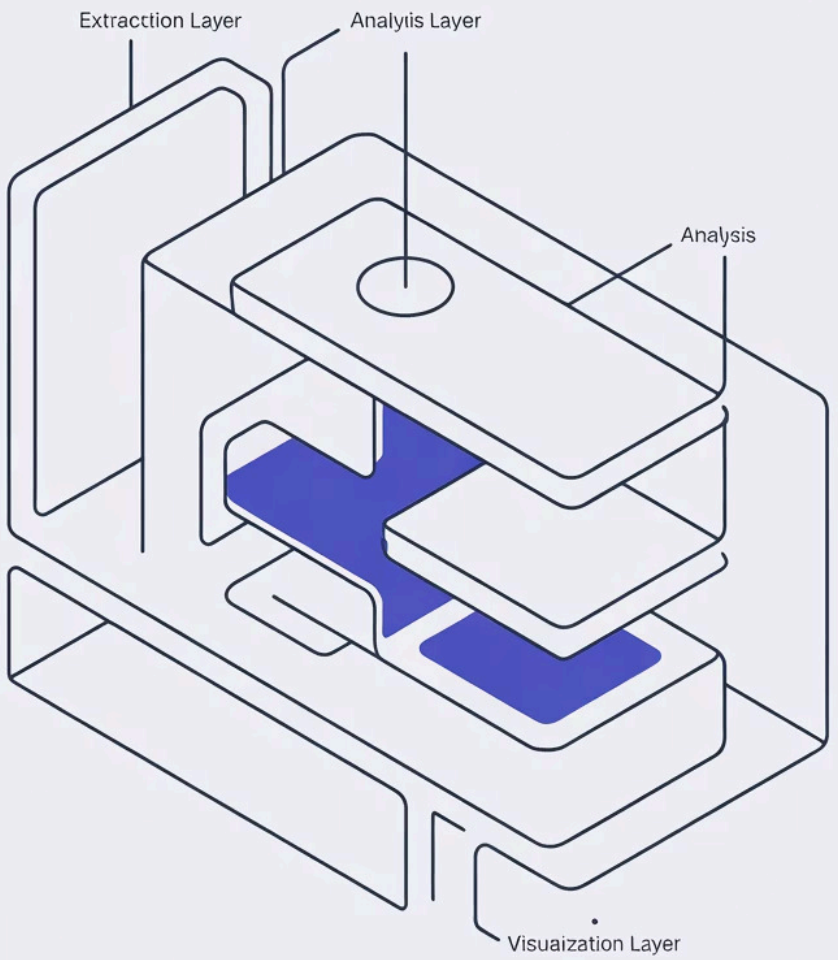
Interactive toolkit for extracting and visualizing internal states (attention weights, gradients, activations) with dynamic interfaces

Expected Outcomes

Working explainability prototype with human-centered evaluation linking technical inspection to cognitive understanding



Explainability Suite Architecture



Methodology

1

Architecture Design

Three-layer system: Extraction (model internals), Analysis (interpretability metrics), Visualization (interactive interface)

2

Model Integration

Select transformer models, implement wrappers for attention matrices, embeddings, activations using existing libraries

3

Interface Development

Build interactive dashboard allowing users to upload text, inspect layers, view attention heatmaps and influence scores

4

Evaluation

Assess visualization methods for human insight, conduct usability experiments, evaluate explanation consistency

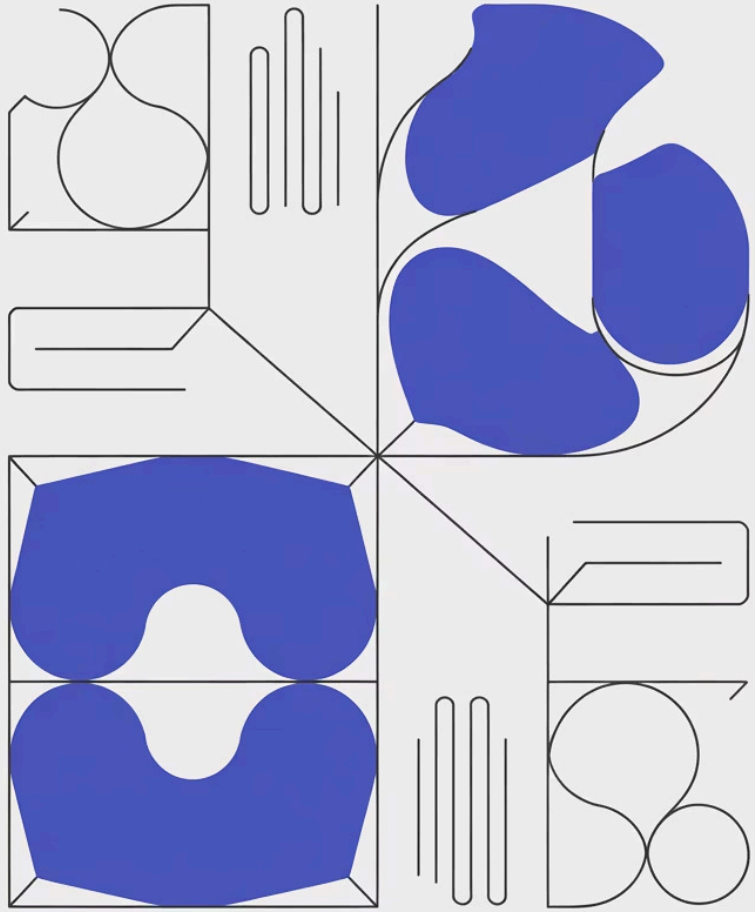
Implementation Guidelines: Encourage modular design with Python modules, OOP principles, documented APIs, and at least one interactive notebook demo.

Dataset & References

Dataset: Any text classification or QA dataset suitable for visual experiments and interpretability analysis.

References

- Grimsley, C., et al. (2020). Why attention is not explanation: Surgical intervention and causal reasoning about neural models. *LREC*, 1780-1790.
- Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.
- Raza, S., et al. (2025). Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*.



P15. Cartographers of the Invisible

This project turns students into **semantic explorers** — "cartographers" mapping how **Large Language Models (LLMs)** represent meaning across linguistic and multimodal spaces. By combining embedding analysis, visualization, and interpretability techniques, the project aims to **make abstract representations visible**.

Core Pipeline

Concept embeddings extracted from LLMs and projected into 2D/3D spaces using dimensionality reduction with interactive visualizations

Expected Outcomes

Intuitive, navigable maps of semantic space offering visual understanding of conceptual relationships and model biases

Methodology & References

01

Concept Selection

Choose semantic domains (emotions, professions, abstract concepts), collect representative textual/visual examples

02

Embedding Extraction

Use pretrained models (BERT, CLIP, LLaMA-2) to extract embeddings, normalize and compute cosine similarity

03

Dimensionality Reduction

Apply PCA, t-SNE, or UMAP for 2D/3D projection, test clustering algorithms for conceptual groupings

04

Interactive Atlas

Build visualization interface allowing cluster exploration, concept search, nearest neighbors, cross-model comparison

LAION-5B

Large-scale text–image dataset for multimodal embedding analysis

ConceptNet

Structured commonsense knowledge base for grounding semantic concepts

References

- Reif, E., et al. (2019). Visualizing and measuring the geometry of BERT. *Advances in neural information processing systems*, 32.
- Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.
- Liang, C. X., et al. (2024). A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*.