P10. Right, Wrong, and Everything in Between

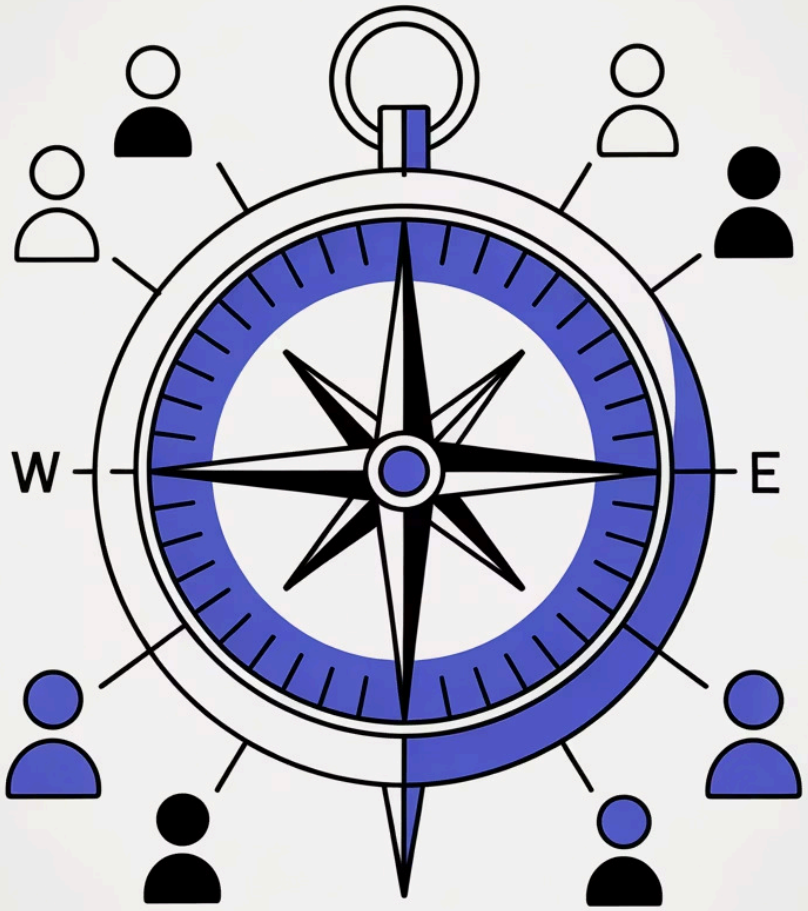
This project examines how LLMs respond to **morally ambiguous or ethically charged prompts**, analyzing *when, how,* and *why* models refuse, reframe, or justify their responses. The aim is to study the intersection between **moral alignment** and **linguistic pragmatics**.

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

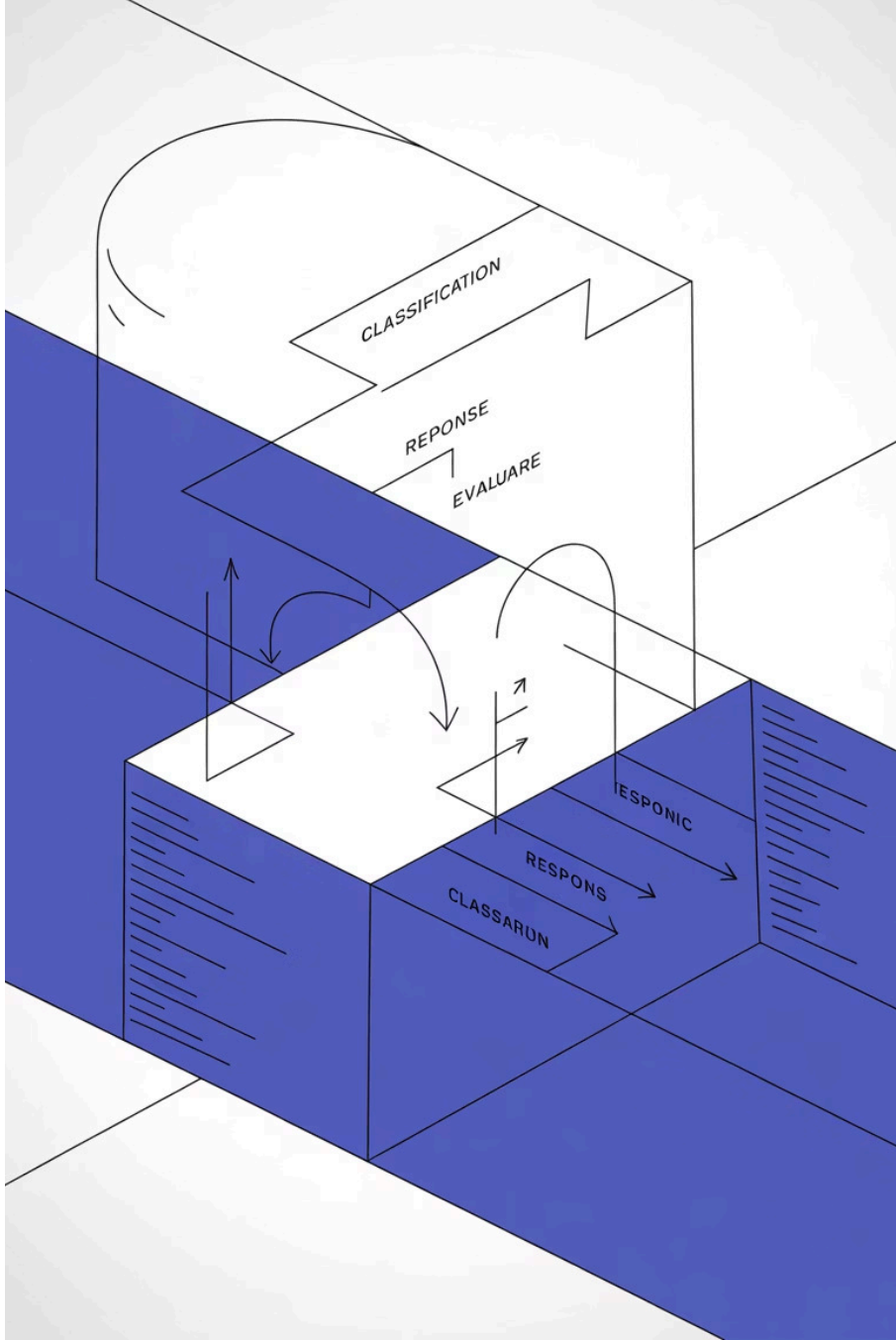
Collect outputs from different LLMs, classify by strategy: refusal, justification, reframing

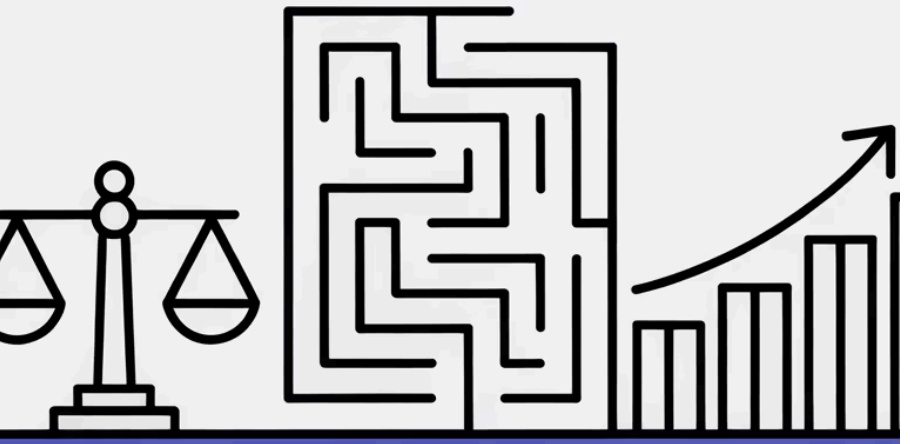
Quantitative Evaluation

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*.