P9. Measuring Total Causal Effects of Instruction Tuning

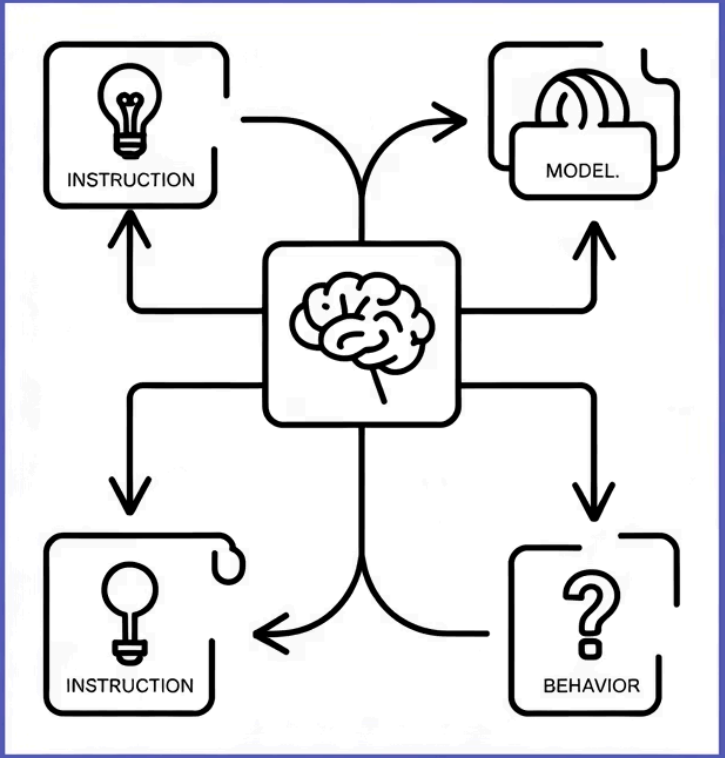
The process of **"instruction tuning"** is a critical step in creating helpful and safe AI assistants. This project frames instruction tuning as a "treatment" and aims to measure its **total causal effect** on a range of model variables, moving beyond simple performance metrics to quantify both intended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

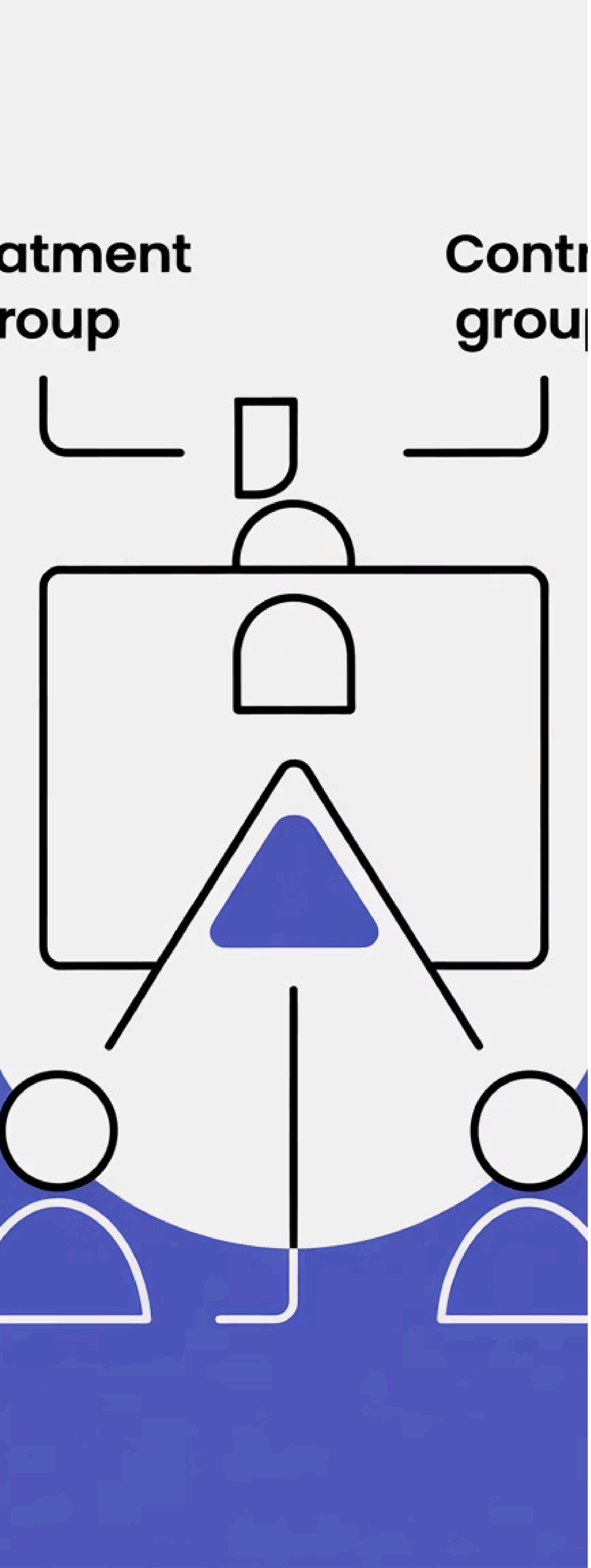
Core Pipeline

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

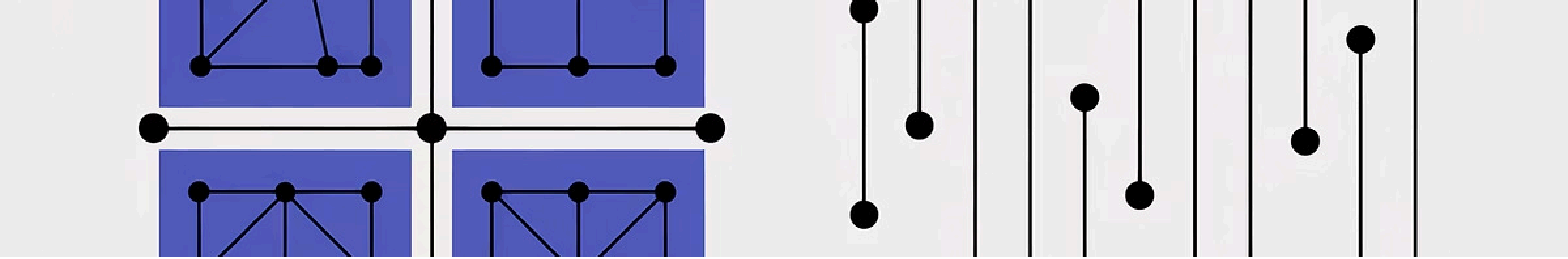
☆

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.