# 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

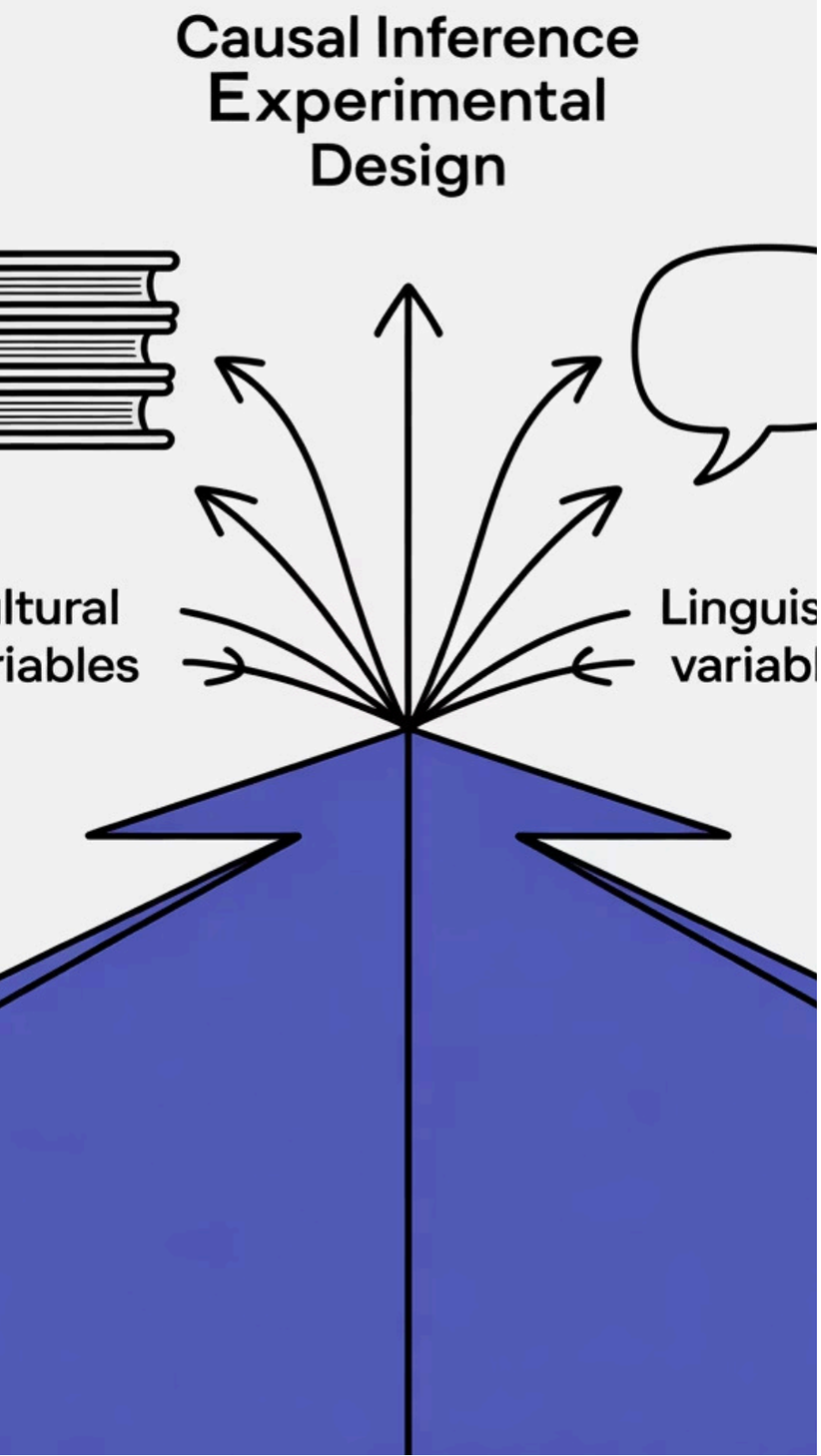
### Cultural Persona

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

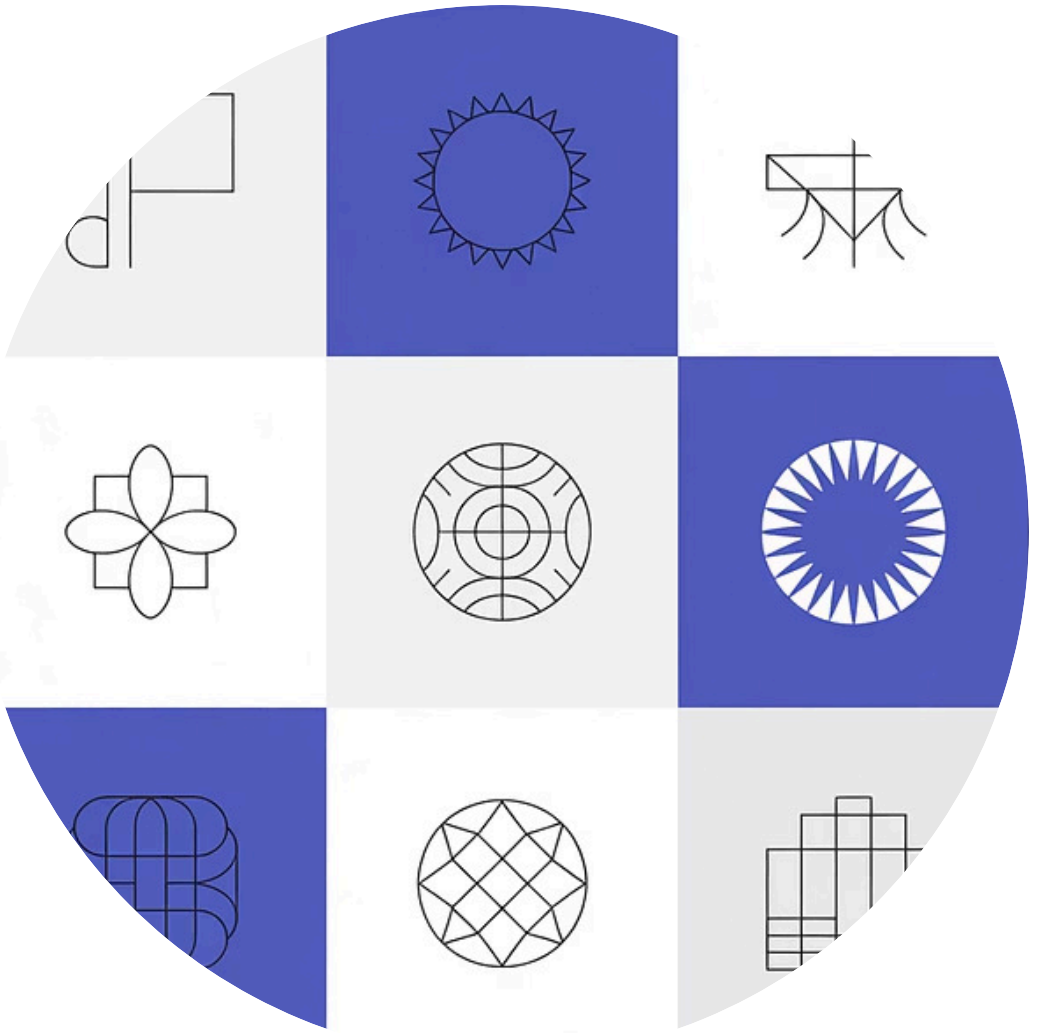
Accuracy of cultural facts and norms

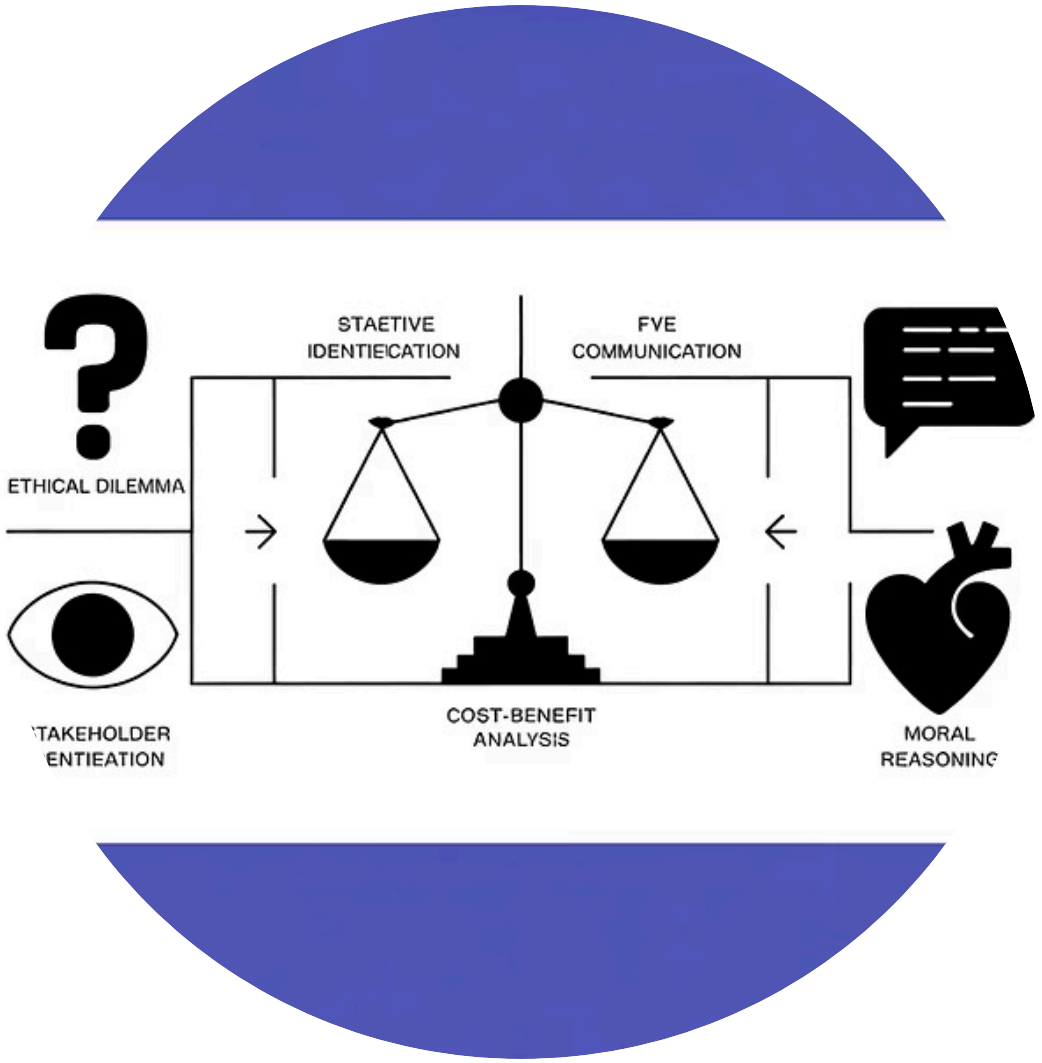
**3**

### 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.