Scalable XAI Evaluation Using LLM-Simulated User Populations: A Multi-Strategy Prompting Framework

Tavia Liu, Xinshi Feng
Advised by Torrence Farmer, Chien-Ju Ho

Washington University in St. Louis
James McKelvey School of Engineering

MOTIVATION

- Understanding why AI systems make decisions is critical for trust, yet evaluating explanation quality typically requires expensive and time-consuming human studies [1, 2].
- Current metrics for evaluating AI explanations often fail to capture human perspectives and key explanatory qualities [1, 3].
- Large language models offer a promising approach to simulate human judgment, enabling scalable and cost-effective evaluation of AI explanations [1, 2].

RESEARCH QUESTIONS

- Can LLMs replicate human decision-making in XAI evaluation when prompted with user characteristics?
- Which prompting strategies most effectively capture human decision-making patterns?

METHODOLOGY 1

We test whether LLMs can replicate human personality-behavior relationships by prompting GPT-4o-mini with real participant profiles and comparing predicted behaviors against actual decisions.

Dataset

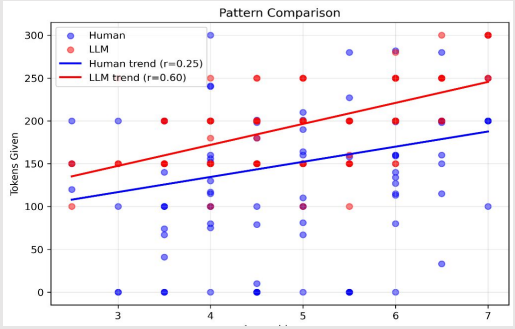
- N = 158 participants from behavioral economics study
- Measures: Big Five personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness)
- Behavioral task: Ultimatum Game offers (how much to offer opponent out of 400 tokens)

Simulation

- Model: GPT-4o-mini
- Prompt: Participant personality profile (Big Five scores)→ Predicted token allocation

RESULT 1

LLM captures *direction* of relationship but *amplifies* effect magnitude.



RESULT 2

LLM shows an improvement in correlation with CULTURAL prompting strategy

METHODOLOGY 2

We tested whether explicitly invoking cultural background in the prompts will improve LLM alignment with human cross-cultural values compared to simply mention of nationality.

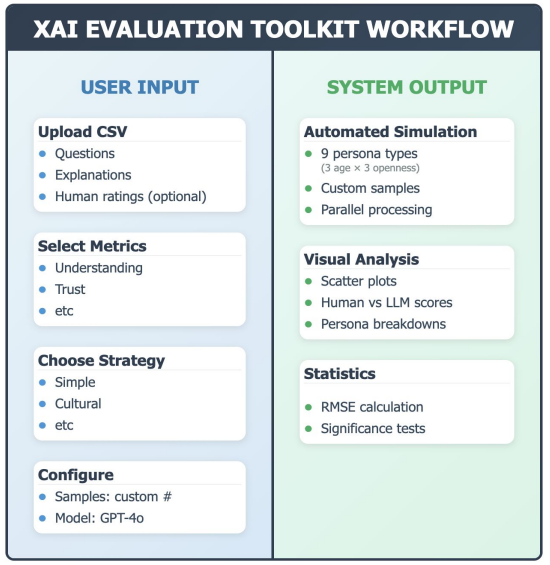
Dataset

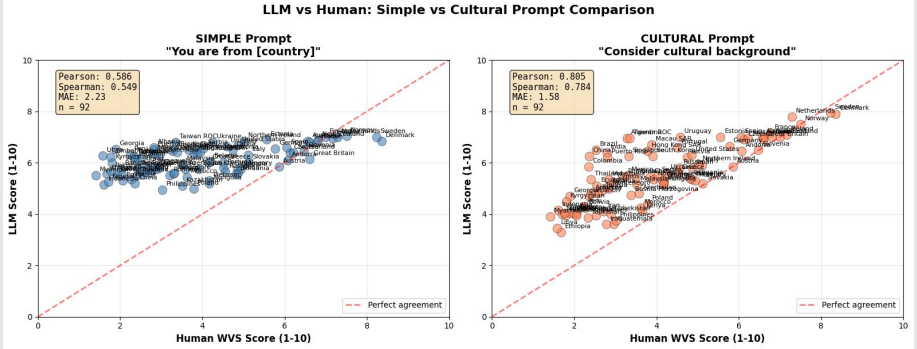
- N = 92 countries from World Value Survey (WVS)
- Measures: Abortion attitudes (1-10 scale, higher = more accepting)

Simulation

- Model: GPT-4o-mini with persona simulation
- Prompting strategies:
  - SIMPLE: "You are from [country]"
  - CULTURAL: " Consider your cultural background from [country]"
- Evaluation: Pearson correlation with human WVS scores

Pipeline of Workflow





CONCLUSIONS

- LLMs can simulate personality-behavior relationships when prompted with individual traits
- Prompting strategy significantly impacts alignment, where cultural framing improves correlation by 37%.
- Explicit context matters for value-based tasks
- **Future work**
  Test additional prompting strategies across diverse datasets
  Explore which strategies work best for different task types

[1] Domnich et al., "Towards Unifying Evaluation of Counterfactual Explanations," AAAI, 2025.
[2] Bombassei De Bona et al., "Evaluating Explanations Through LLMs," NeurIPS Workshop, 2024.
[3] Mittelstadt et al., "Explaining Explanations in AI," FAT*, 2019.

Try the Toolkit

Source Code