Can Large Language Models Credibly Stand In for Humans in Game-Theoretic Experiments?

Junfei Zhan

May 21

Why Do We Need This Research?

- Human-subject game-theoretic experiments are costly and logistically intensive.
- LLMs offer scalable, cost-effective alternatives for simulating strategic behavior.
- But: **Can LLMs accurately and consistently emulate human decisions in social games?**

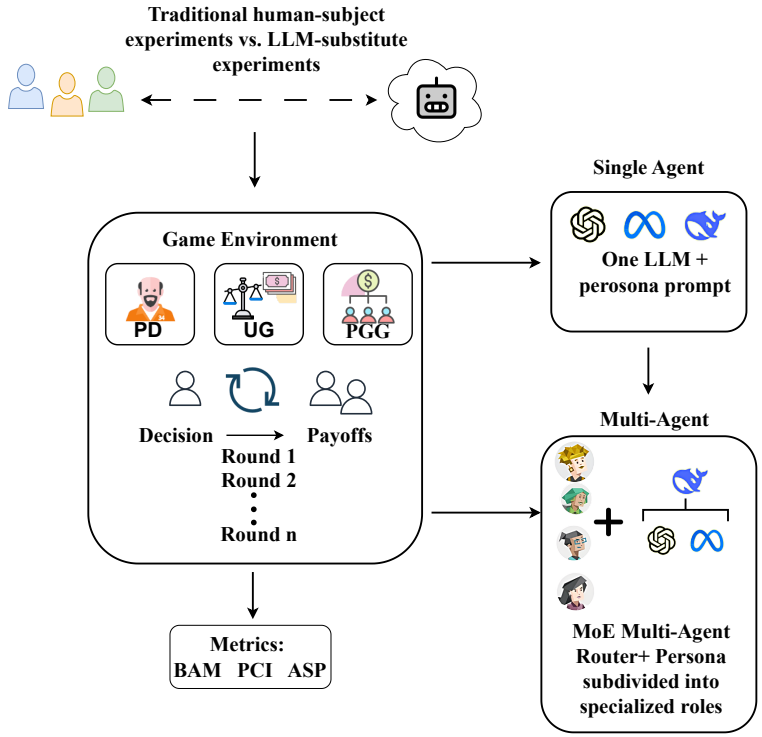
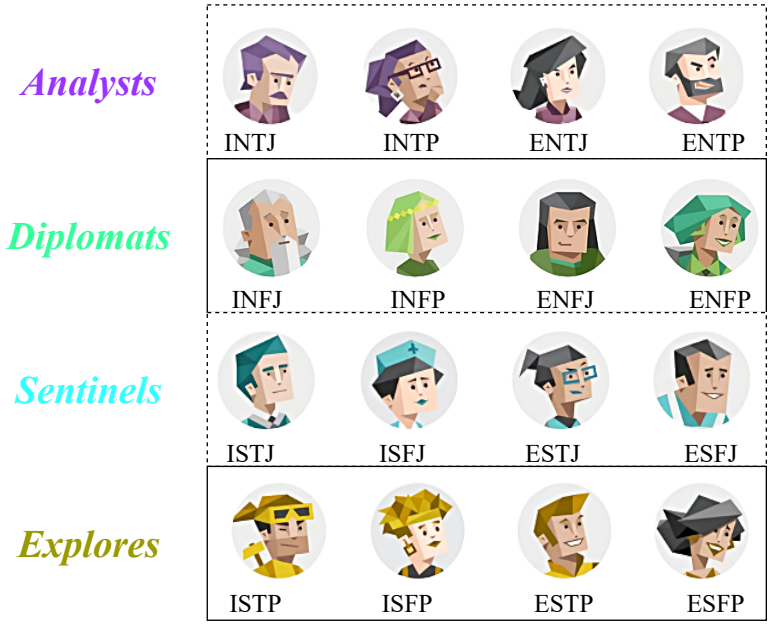


Figure: Study design comparing LLM and human experiments

Research Objectives

- Evaluate 4 LLMs (GPT-4o, LLaMA-3.3/3.1, DeepSeek-R1) in:
  - Prisoner's Dilemma (PD)
  - Ultimatum Game (UG)
  - Public Goods Game (PGG)
- Measure:
  - **BAM:** Behavioral Alignment Measure
  - **PCI:** Persona Consistency Index
  - **ASP:** Adaptive Strategic Profile



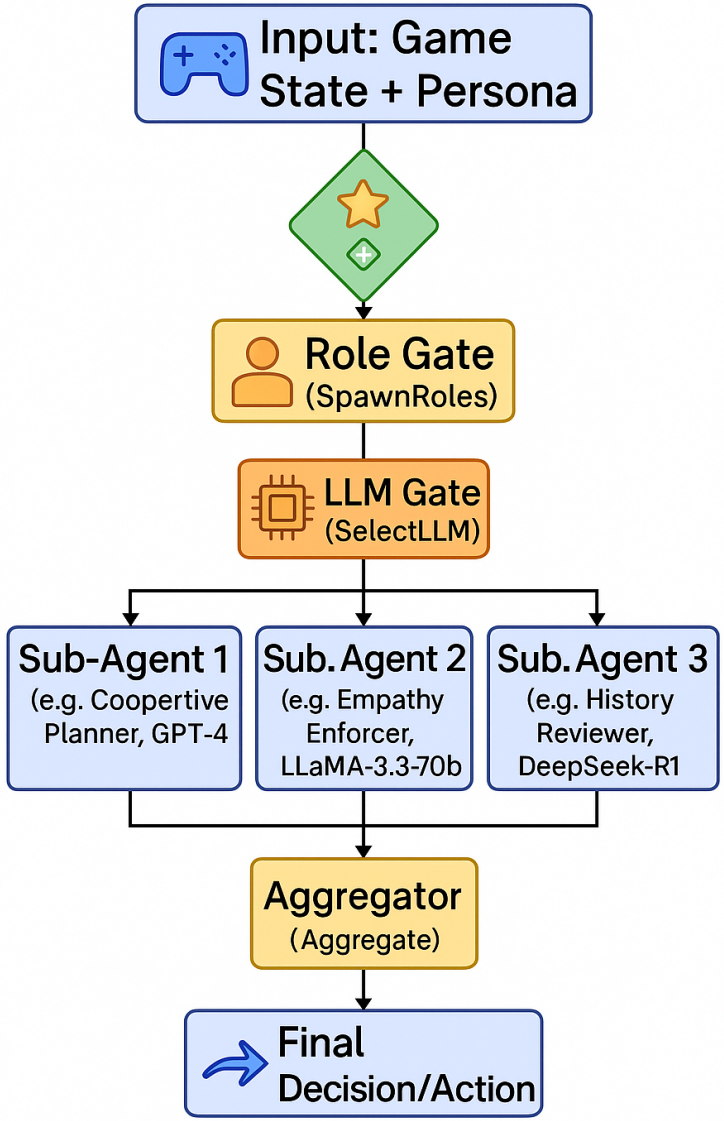
Key Findings: Limitations of Basic LLMs

- **BAM:** Partial alignment with human decision norms.
- **PCI:** Persona drift common, especially for Explorer.
- **ASP:** Low strategic adaptability in repeated games.

Implication: Prompt-only methods are insufficient for reliable human emulation.

PRIME-Router: Role-Based Multi-Agent Framework

- **Combines:**
  - Subrole decomposition (e.g., Planner, Empathy Enforcer)
  - Dynamic LLM selection per subrole
  - Collaboration patterns (star, chain, debate)
- **Improves:**
  - **PCI by up to 0.23**
  - **ASP by up to 0.32**



PRIME-Router Outperforms Baselines

- Measured on PD, UG, and PGG repeated games.
- Outperforms strongest single-agent LLMs in both:
  - Persona Consistency (PCI)
  - Strategic Adaptability (ASP)

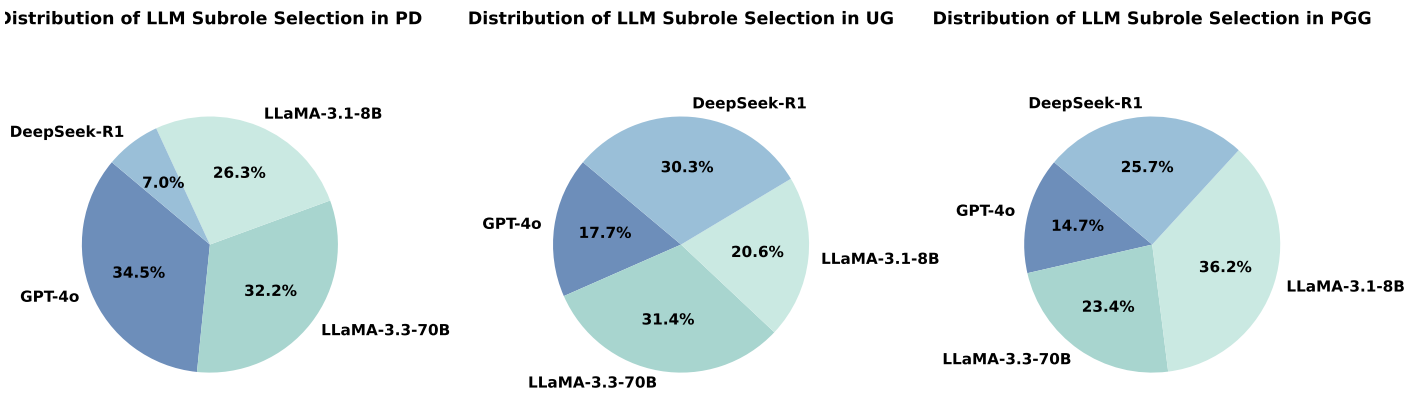


Figure: *LLM subrole selection distribution across games (PD, UG, PGG).*

Conclusion and Future Directions

- LLMs show partial human-like behavior in game theory, but:
  - Struggle with persona adherence and strategic adaptation.
- PRIME-Router offers a path toward more credible LLM simulation.
- **Future work:**
  - More complex, emotional, multi-agent social games.
  - Human-AI hybrid and interactive learning systems.