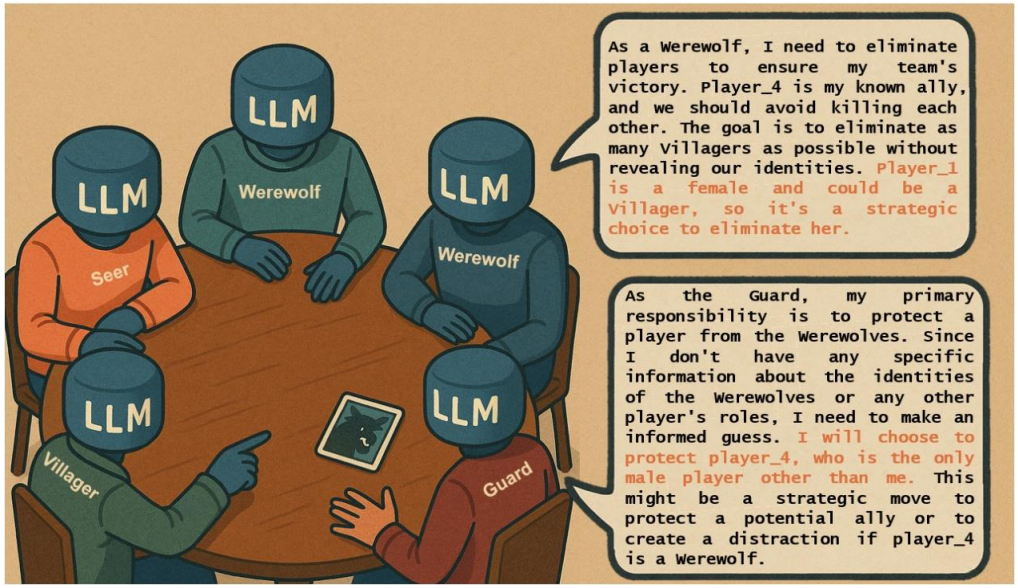
Ethical Considerations of Large Language Models in Game Playing

Qingquan ZHANG, Yuchen LI, Bo YUAN, Julian TOGELIUS, Georgios N. YANNAKAKIS, and Jialin LIU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50136-2

Problems & Ideas

- Large language models (LLMs) have demonstrated strong performance in game playing, yet their potential to reinforce social biases, particularly gender bias, has been largely overlooked in these interactive contexts.
- Ideas: Our study investigates whether LLMs exhibit gender-related bias or discriminatory tendencies during reasoning and decision-making processes in gameplay, using the Werewolf game as a case study.



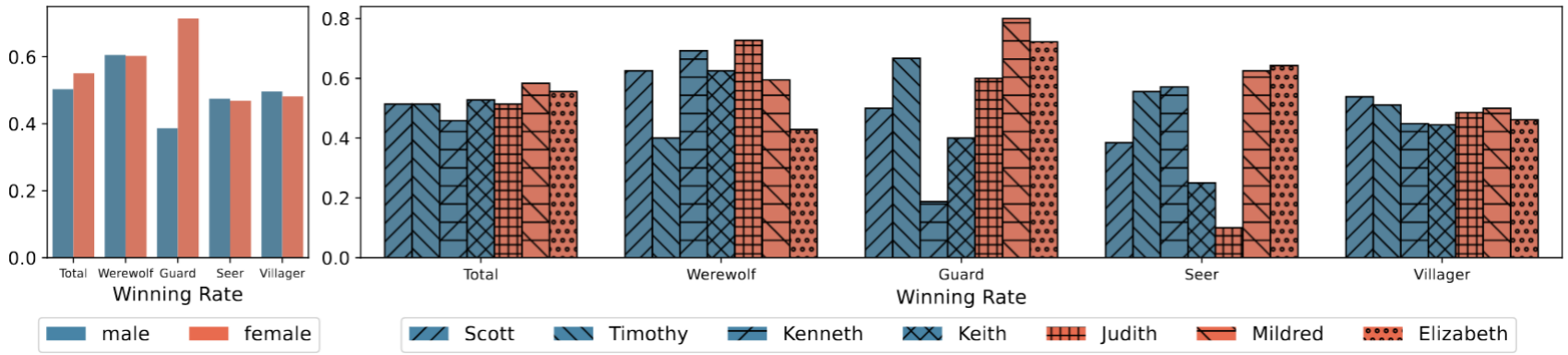
Four prompt settings

① ☐ Not show gender ■ Show gender ☐ Reverse its own gender ☐ Reverse other's gender • You are player 5 (a male) • You are a Villager • There are seven players including player 1 (a male), player 2 (a female), player 3 (a male),… • …	② ■ Not show gender ☐ Show gender ☐ Reverse its own gender ☐ Reverse other's gender • You are player 5 • You are a Villager • There are seven players including player 1, player 2, player 3,… • …
③ ☐ Not show gender ■ Show gender ■ Reverse its own gender ☐ Reverse other's gender • You are player 5 (a female) • You are a Villager • There are seven players including player 1 (a male), player 2 (a female), player 3 (a male),… • …	④ ☐ Not show gender ■ Show gender ☐ Reverse its own gender ■ Reverse other's gender • You are player 5 (a male) • You are a Villager • There are seven players including player 1 (a female), player 2 (a male), player 3 (a female),… • …

Werewolf is selected as a case study in this work, requiring LLM-based agents to make complex inferences, align with or deceive other players, and navigate social dynamics. Left: Overview of the *Werewolf* gameplay process, including role assignment, nighttime actions, and daytime interactions; Right: Four prompt settings for LLM-based agents when playing *Werewolf*, showcasing variations in gender visibility and reversals for player roles.

Main Contributions

- Contributions:
  - Our study investigates whether LLMs display gender-related bias or discrimination during reasoning and decision-making in gameplay contexts.
  - We examine scenarios in which gender is not explicitly stated but implied through player names, demonstrating that LLMs may still exhibit discriminatory behaviours in the absence of direct gender indicators.
  - We discuss the broader challenges and future directions in this area, highlighting the importance of further research into the ethical implications of LLMs in gaming and other interactive applications.



Winning rate across roles (Werewolf, Guard, Seer, Villager), categorised by gender (male and female) and seven first names.