Helmsman of the Masses? Evaluate the Opinion Leadership of Large Language Models in the Werewolf Game

Silin Du and Xiaowei Zhang

Presenter: Silin Du

Department of Management Science and Engineering
School of Economics and Management
Tsinghua University
dsl21@mails.tsinghua.edu.cn

October 8, 2024

Contents

Large language models (LLMs) have demonstrated a high level of expertise in comprehending and producing human-like natural languages.

- Social deductive games are suitable scenarios to study *social preference* of LLMs [Meng, 2024]. (e.g., Werewolf [Xu et al., 2023a, Xu et al., 2023b] and Avalon [Wang et al., 2023a])
- LLMs, such as GPT-4 [Achiam et al., 2023], exhibit strategic behaviors including cooperation, confrontation, deception, and persuasion in these games [Xu et al., 2023a, Xu et al., 2023b, Lan et al., 2023, Wang et al., 2023a].

Background II

Potential *opinion leadership* of LLMs has been overlooked and confounded [Kano et al., 2023].

Definition 1.1 (Opinion Leaders)

Opinion leaders are individuals who exert personal influence upon a certain number of other people in certain situations [Rogers and Cartano, 1962].

Definition 1.2 (Opinion Leadership)

Opinion leadership is the composite ability of opinion leaders to comprehensively employ the aforementioned strategies to influence the decisions of their followers and shape public opinion [Bamakan et al., 2019].

Practical implications: interaction design (IxD), decision optimization, and public regulation.

- In multi-agent systems, such as smart manufacturing, a few opinion leaders can significantly impact task efficiency and outcomes [Rapanos, 2023].
- As AI assistants and customer service agents, LLMs can shape user experience and business decisions.
- In social media and forums, the opinion leadership of LLMs risks influencing social discourse and public decisions.

Methods of Measuring Opinion Leadership

Self-assessment: using well-established scales

- Monomorphic scales [Childers, 1986, Flynn et al., 1996, Reynolds and Darden, 1971].
- Maven scales [Feick and Price, 1987, Boster et al., 2011].
- Polymorphic scales [Noelle-Neumann, 1983, Gnambs and Batinic, 2011].

Assessment by others: expensive and laborious [Weimann et al., 2007]; widely used in health-related contexts [Valente and Pumpuang, 2007]

Assessment algorithms: *social web context*; key figures with weighting → OL score

- [Probst et al., 2013]: Two main approaches
- Methods of *detecting* influential OLs [Singh et al., 2013, Ennaji et al., 2018, Bamakan et al., 2019]

Main criteria: contacts, activity, feedback, and citation/imitation [Jungnickel, 2018]

Challenges

- Opinion leaders are often concealed [Bamakan et al., 2019]
- Assessing the opinion leadership is a daunting task due to the complexity of decision-making tasks
- It's impractical to conduct large-scale real-task randomized controlled trials to evaluate the diverse effects on human and AI followers.

Fortunately, the Werewolf game provides a promising testing ground to address these challenges.

- The *Sheriff* role in the Werewolf game is jointly elected by other players and can summarize the statements and give decision-making suggestions.

Contribution

1. To the best of our knowledge, this is the first in-depth analysis of opinion leadership within LLMs. We clarify the opinion leadership of diverse LLMs in various contexts.
2. We introduce the setting of the *Sheriff* and implement a Werewolf game framework, which can seamlessly integrate diverse LLM-based agents and human players. Besides, We devise *two novel metrics* to evaluate the opinion leadership of different LLMs.
3. We conduct simulations and human evaluations to assess LLMs' opinion leadership, and we collect a Werewolf question-answering (*WWQA*) dataset for further analysis.

Contents

LLM-based agents

- reasoning and decision-making [Yao et al., 2023], interacting with the environment [Ahn et al., 2022, Cui et al., 2023]
- cognitive abilities [Shapira et al., 2023, Zhuang et al., 2023] → simulating believable human behavior [Park et al., 2023]

Multi-agent collaboration

- *Frameworks*: AgentVerse [Chen et al., 2023b], AutoGen [Wu et al., 2023], MetaGPT [Hong et al., 2023], ChatEval [Chan et al., 2023]
- *Scenarios*: courtroom simulations [Talebirad and Nadiri, 2023], game development [Hong et al., 2023], auctions [Chen et al., 2023a]
- *Gameplay*: Texas Hold'em poker [Gupta, 2023], complex video games [Wang et al., 2023b, Zhu et al., 2023]

Multi-player social deduction games with LLMs

- *Werewolf*
  - [Xu et al., 2023a]: a tuning-free framework; strategic behaviors (trust, confrontation, camouflage, leadership, etc.)
  - [Xu et al., 2023b]: integrate reinforcement learning to enhance the agents' decision-making prowess
- *Avalon* [Lan et al., 2023, Light et al., 2023, Wang et al., 2023a]

Opinion leadership in human society

- *Social learning* [Festinger, 1954]
  When faced with uncertainty, humans tend to learn from "opinion leaders" [Bala and Goyal, 1998]
- *Opinion leaders*
  Individuals with extensive knowledge in a specific subject area (experts) or individuals with extensive social connections (social connectors) [Goldenberg et al., 2006]
- The role of **trust** in opinion leadership
  - To exert influence, opinion leaders must be trustworthy [Nahapiet and Ghoshal, 1998]
  - The extent to which their opinions are adopted relies on their credibility [Dirks and Ferrin, 2001, Mayer et al., 1995]
  - Diminishes conflicts, alleviates the necessity for information verification [Currall and Judge, 1995], and enhances people's inclination and ability to embrace others' perspectives [Carley, 1991, Mayer et al., 1995]

Contents

We consider the game setting with 7 players, including 2 Werewolves, 3 Villagers, 1 Seer, and 1 Guard.

At night,

- *Werewolf*: recognize each other and choose one player to kill.
- *Seer*: chooses one player to check its hidden tole.
- *Guard*: protects one player including themselves from the Werewolves.

During the day

- *Announcement*: The moderator announced the result last night to all players.
- *Discussion*: Each player takes turns to make a statement in an order determined by a special role, the Sheriff.
- *Voting*: Each player votes to eliminate one player or chooses not to vote.

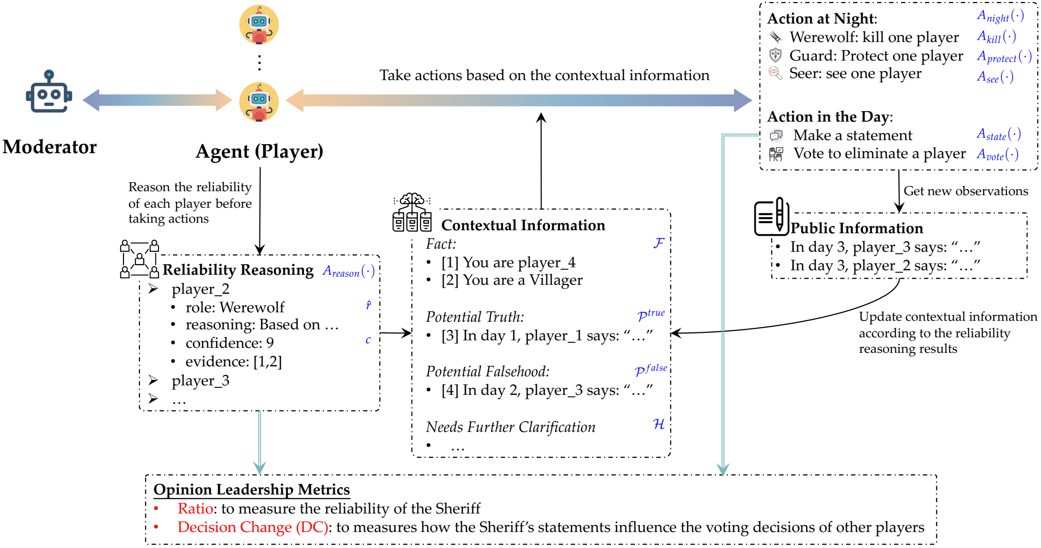


Fig. 3.1: Game framework

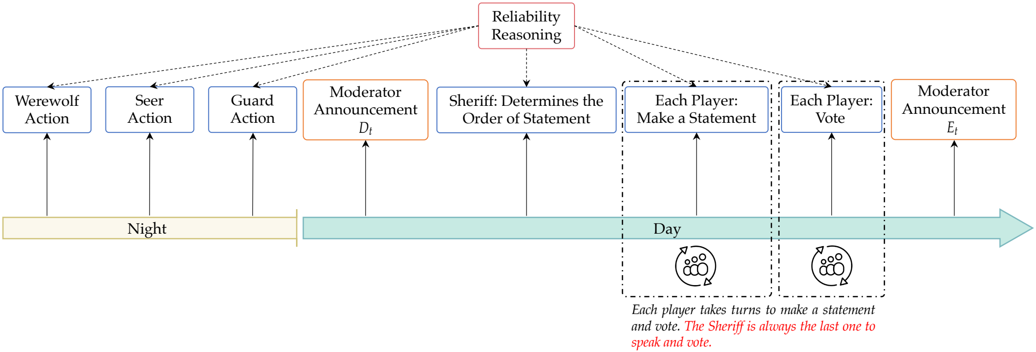


Fig. 3.2: The whole process during round t

LLM-based Player

- Players of different roles are required to perform a series of actions, including night actions (to kill, protect, or see) and day actions (to speak and vote).
- The LLM-based players perform actions through prompting.
  - The game rules and the assigned role
  - The contextual information
  - Task description
- We divide the contextual information for player X_i^r into the following two parts.
  - *Facts*: the role of X_i^r, the announcement made by the moderator, etc.
  - *Public Statements*: the public statement of other players.
- We implement an additional action, called *reliability reasoning*, before making night or day actions.
- In this step, we prompt the LLM to infer the *identities* of other players based on historical information and provide *confidence levels*.

Settings of the Sheriff

- All alive players will receive a special message after the election phase:

> *After discussion and a vote,* `player_l` *was selected as the Sheriff, who can determine the order of statements, summarize the discussion, and provide advice for voting at last.*

- The Sheriff can determine the order of statements. It first performs the reasoning step and then selects its left- or right-hand side player to speak first.
- The statement order will slightly influence the collection of public statements. But the whole process of making statements and voting is the same.

In different scenarios and research contexts, various definitions of opinion leaders have been proposed [Chowdhry and Newcomb, 1952, Katz, 2015, Lazarsfeld et al., 1968].

- Opinion leaders are generally more trustworthy
- Opinion leaders influence the views and even decisions of others.

Two evaluation metrics to measure the opinion leadership of LLM-based players.

- Ratio
- Decision Change (DC)

[Metric 1] Ratio: the reliability of the Sheriff

- The average mutual reliability of all players except the Sherif:

$$\bar{m}_1(t) = \frac{1}{(N_d(t)-1)(N_d(t)-2)} \sum_{i \in \texttt{Alive}_d(t), i \neq L(t)} \sum_{j \in \texttt{Alive}_d(t), j \neq L(t), j \neq i} m^v_{i,j,t}$$

where $N_d(t) = |\texttt{Alive}_d(t)|$ be the number of alive players on the day of round t.

- The average reliability of other players toward the Sheriff is

$$\bar{m}_2(t) = \frac{1}{N_d(t)-1} \sum_{i \in \texttt{Alive}_d(t), i \neq L(t)} m^v_{i,L(t),t}$$

- Then the Ratio is defined as:

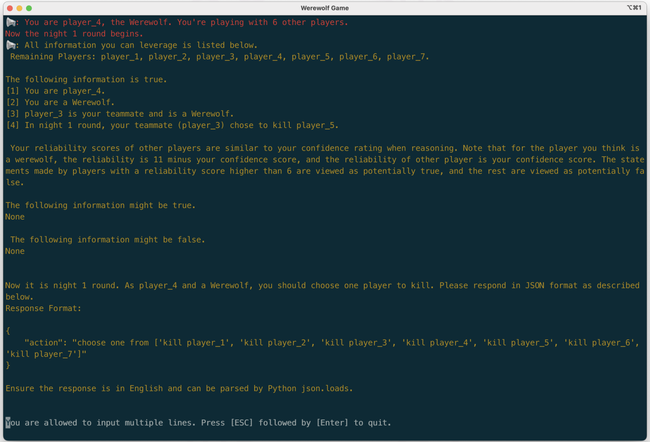
$$\text{Ratio} = \frac{1}{T} \sum_{t=1}^{T} \frac{\bar{m}_2(t)}{\bar{m}_1(t)}$$

Opinion Leadership Metrics IV

[Metric 2] Decision Change (DC): how Sheriff influence the voting decisions of other players

- At the end of the discussion on day t, all players are required to reason the reliability of the Sheriff and make a *pseudo-voting decision* $(A'_{vote,t}(X_i))$.
- Then the Sheriff makes a statement and all players move into the voting phase to yield $A_{vote,t}(X_i)$.
- The proportion of players that change their decision to be in line with the Sheriff.

$$\mathrm{DC} = \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{i\in \texttt{Alive}_d(t), i\neq L(t)} \mathbb{I}\left\{\left(A'_{vote,t}(X_i) \neq A_{vote,t}(X_{L(t)})\right) \text{ AND } \left(A_{vote,t}(X_i) = A_{vote,t}(X_{L(t)})\right)\right\}}{N_d(t)-1}$$



```
Werewolf Game
: You are player_4, the Werewolf. You're playing with 6 other players.
Now the night 1 round begins.
: All information you can leverage is listed below.
 Remaining Players: player_1, player_2, player_3, player_4, player_5, player_6, player_7.

The following information is true.
[1] You are player_4.
[2] You are a Werewolf.
[3] player_3 is your teammate and is a Werewolf.
[4] In night 1 round, your teammate (player_3) chose to kill player_5.

 Your reliability scores of other players are similar to your confidence rating when reasoning. Note that for the player you think is a werewolf, the reliability is 11 minus your confidence score, and the reliability of other player is your confidence score. The statements made by players with a reliability score higher than 6 are viewed as potentially true, and the rest are viewed as potentially false.

The following information might be true.
None

 The following information might be false.
None


Now it is night 1 round. As player_4 and a Werewolf, you should choose one player to kill. Please respond in JSON format as described below.
Response Format:

{
    "action": "choose one from ['kill player_1', 'kill player_2', 'kill player_3', 'kill player_4', 'kill player_5', 'kill player_6', 'kill player_7']"
}

Ensure the response is in English and can be parsed by Python json.loads.


You are allowed to input multiple lines. Press [ESC] followed by [Enter] to quit.
```

Fig. 3.3: Screenshot of the interface during human evaluation

Contents

WWQA Dataset

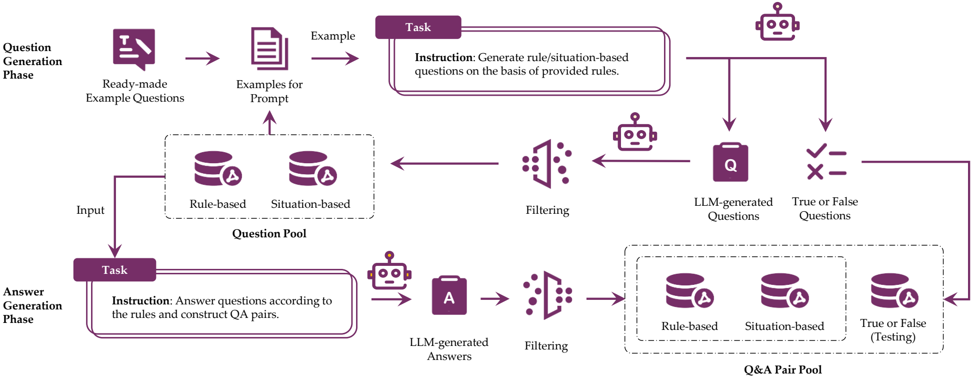


Fig. 4.1: Overview of the data generation process

Simulation I

- **Baseline model**: GLM-3 [Zeng et al., 2022]
- **Sheriff**: ChatGLM3-6B [Zeng et al., 2022], Mistral-7B [MistralAI, 2023], Baichuan2-13B [Yang et al., 2023], InternLM-20B [InternLM, 2023], Yi-34B [Young et al., 2024], GLM-3 [Zeng et al., 2022], GLM-4 [ZhipuAI, 2024], and GPT-4 [Achiam et al., 2023].

Table 4.1: Evaluation results on different LLMs

Metric \ Model	C3-6B	M-7B	B-13B	In-20B	Yi-34B	GLM-3	GLM-4	GPT-4
Binary QA								
Accuracy	0.582	0.756	0.750	0.794	0.792	0.760	0.846	0.850
F1	0.565	0.753	0.749	0.789	0.794	0.761	0.846	0.851
Opinion Leadership								
Ratio	0.863	0.820	0.922	0.884	0.882	1.054	1.167	1.093
DC	0.088	0.151	0.118	0.068	0.037	0.126	0.113	0.107

Table 4.2: Evaluation results on fine-tuned LLMs

Metric \ Model	C3-6B	C3(FT)	M-7B	M(FT)	B-13B	B(FT)	In-20B	In(FT)
Binary QA								
Accuracy	0.582	0.564	0.756	0.766	0.750	0.768	0.794	0.850
F1	0.565	0.563	0.753	0.754	0.760	0.769	0.789	0.841
Opinion Leadership								
Ratio	0.863	0.847	0.820	0.779	0.922	1.002	0.884	0.948
DC	0.088	0.047	0.151	0.034	0.118	0.076	0.068	0.110

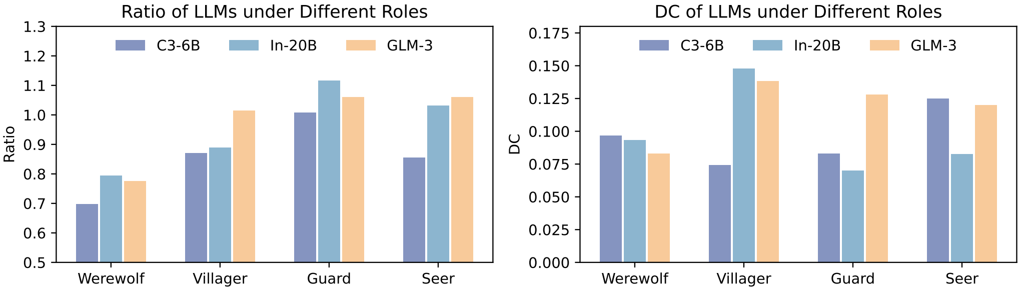


Fig. 4.2: Opinion leadership of LLMs under different roles

Human Evaluation

- We invite several college students proficient in English and familiar with the Werewolf game to join in the human evaluation experiment.
- Each human player will play 5 games with 6 LLM-based players.

Table 4.3: Human evaluation results

Methods \ Metrics	Ratio	DC	Correlation
Simulation	1.054	0.126	-
Human Evaluation	1.341	0.083	0.233

Contents

- A *framework* of the Werewolf game that incorporates a Sheriff role, catering to both LLMs and human players.
- Two *metrics* to measure the Sheriff's opinion leadership: the first metric measures their reliability, and the second assesses their influence on other players' decisions.
- A *WWQA* dataset is collected to enhance LLMs' grasp of the game rules by fine-tuning. Initial attempts indicate that it's non-trivial to promote the opinion leadership of LLMs.
- Through extensive simulations, we evaluate LLMs of different scales and find that only a few LLMs show a certain degree of opinion leadership.
- Human evaluation experiments suggest that LLMs can gain human trust but struggle to influence human decisions.

Thank You for Listening!

Helmsman of the Masses? Evaluate the Opinion Leadership of Large Language Models in the Werewolf Game

Silin Du and Xiaowei Zhang

Presenter: Silin Du

Department of Management Science and Engineering, Tsinghua University

`dsl21@mails.tsinghua.edu.cn`

October 8, 2024

– Paper –

– Code –

– Contact –

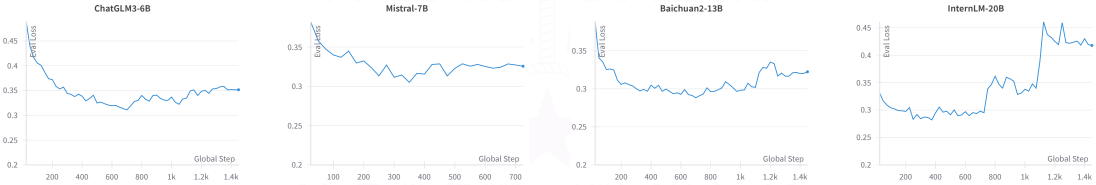
Table 5.1: Examples of Hallucination

A hallucination that one can speak during the night. player_4: I want to address the concerns raised about my silence last night. As the Seer, I observed player_5 and found out they are a Villager. This information could be useful in identifying the Werewolves. However, I remained silent to avoid providing any information that could potentially put me in danger. I urge everyone to carefully consider the actions and statements of player_6.
A hallucination that some player has already made a statement. Moderator: The Sheriff (player_7) chose player_6 to make a statement first in day 1. player_6: I'm not quite sure who to suspect yet. It's still early in the game and we have limited information. I think player_5's silence is a bit suspicious, but it could just be a strategy.
Time confusion. player_3 said nothing. player_4: player_3 kept silent yesterday, which could be a sign of guilt or fear.
Identity confusion. player_7: I encourage everyone to remain cautious and consider all possibilities during the voting phase. player_4's statement in day 2 implicating player_7 and me as Werewolves is concerning.

Homogeneous simulation, where all players are implemented using the same LLM. This setup allows the simulation to proceed strictly following the game setting.

Table 5.2: Homogeneous simulation results

Metric \ Model	C3-6B	B-13B	GLM-3	GLM-4
Ratio	1.021	1.010	1.068	1.152
DC	0.122	0.108	0.118	0.098



(a) C3-6B (b) M-7B (c) B-13B (d) In-20B

Fig. 5.2: Evaluation loss of different LLMs during fine-tuning

Details of Human Evaluation

We invite 8 college students, including 3 graduate students, from a top-tier university in China to participate in the human evaluation. All students have passed the CET-6 English exam and have experience with Werewolf games either online or offline. Our experiment mainly consists of the following three steps.

1. *Training.* We first explain the purpose of the experiment and then confirm with each participant that they understand the settings and rules of the Werewolf game. We extract 5 records from the WWQA dataset to test the participants, ensuring they could participate in the game in English. We also demonstrate how to input content following the format instruction.
2. *Experiment.* Each participant participates in the game by typing text, as shown in Figure 4.1. Each participant plays 5 iterations with GLM-3. Before each iteration, they should input an integer to initialize the roles of all players. Human players do not participate in the election of the Sheriff and will not be selected as the Sheriff. An iteration of the game ends either according to the rules or immediately if the human player is eliminated.
3. *Interview.* After the experiment, we conduct a simple interview with the participants, asking them to evaluate the LLM's performance in the Werewolf game, including whether the LLM can follow the game rules, whether its logic is rigorous, and whether it's easy to distinguish between LLM-based players and the human players.

Examples of WWQA Dataset

Rule-based QA Pairs	
question:	What happens to the player with the most votes in the voting phase?
answer:	The player with the most votes in the voting phase is eliminated from the game.
question:	How do the Werewolves choose a player to kill?
answer:	The Werewolves choose a player to kill by voting and selecting the player who receives the most votes.
Situation-based QA Pairs	
question:	During the discussion phase, if players 1, 3, and 5 suspects player_2, but player_2 is the Seer and reveals their role, how might this affect the voting outcome?
answer:	During the discussion phase, if players 1, 3, and 5 suspect player_2, but player_2 reveals their role as the Seer, it might affect the voting outcome. The revelation of player_2 being the Seer may lead to a shift in suspicion towards other players, potentially changing the voting decisions.
question:	If player_3 receives the most votes in the voting phase and player_3 is the last remaining Werewolf, will the Werewolves win the game?
answer:	If player_3 receives the most votes in the voting phase and player_3 is the last remaining Werewolf, the Werewolves will not win the game. The Werewolves win the game if the number of remaining Werewolves is equal to the number of remaining Seer, Guard, and Villagers. In this scenario, if player_3 is the last remaining Werewolf, there would still be Seer, Guard, and Villagers left, and the Werewolves cannot win.
Binary QA Pairs	
question:	Is the game won by the Werewolves if all Villagers are eliminated?
answer:	No
question:	Do the Werewolves know each other?
answer:	Yes

Table 5.3: Examples of QA pair data

Reference I

[Achiam et al., 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023).
Gpt-4 technical report.
arXiv Preprint arXiv:2303.08774.

[Ahn et al., 2022] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. (2022).
Do as i can, not as i say: Grounding language in robotic affordances.
arXiv Preprint arXiv:2204.01691.

[Bala and Goyal, 1998] Bala, V. and Goyal, S. (1998).
Learning from neighbours.
The Review of Economic Studies, 65(3):595–621.

[Bamakan et al., 2019] Bamakan, S. M. H., Nurgaliev, I., and Qu, Q. (2019).
Opinion leader detection: A methodological review.
Expert Systems with Applications, 115:200–222.

[Boster et al., 2011] Boster, F. J., Kotowski, M. R., Andrews, K. R., and Serota, K. (2011).
Identifying influence: Development and validation of the connectivity, persuasiveness, and maven scales.
Journal of Communication, 61(1):178–196.

[Carley, 1991] Carley, K. (1991).
A theory of group stability.
American Sociological Review, pages 331–354.

[Chan et al., 2023] Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. (2023).
Chateval: Towards better llm-based evaluators through multi-agent debate.
arXiv Preprint arXiv:2308.07201.

Reference II

[Chen et al., 2023a] Chen, J., Yuan, S., Ye, R., Majumder, B. P., and Richardson, K. (2023a).
Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena.
arXiv Preprint arXiv:2310.05746.

[Chen et al., 2023b] Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Qian, C., Chan, C.-M., Qin, Y., Lu, Y., Xie, R., et al. (2023b).
Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents.
arXiv Preprint arXiv:2308.10848.

[Childers, 1986] Childers, T. L. (1986).
Assessment of the psychometric properties of an opinion leadership scale.
Journal of marketing research, 23(2):184–188.

[Chowdhry and Newcomb, 1952] Chowdhry, K. and Newcomb, T. M. (1952).
The relative abilities of leaders and non-leaders to estimate opinions of their own groups.
The Journal of Abnormal and Social Psychology, 47(1):51.

[Cui et al., 2023] Cui, C., Ma, Y., Cao, X., Ye, W., and Wang, Z. (2023).
Receive, reason, and react: Drive as you say with large language models in autonomous vehicles.
arXiv Preprint arXiv:2310.08034.

[Currall and Judge, 1995] Currall, S. C. and Judge, T. A. (1995).
Measuring trust between organizational boundary role persons.
Organizational Behavior and Human Decision Processes, 64(2):151–170.

[Dirks and Ferrin, 2001] Dirks, K. T. and Ferrin, D. L. (2001).
The role of trust in organizational settings.
Organization Science, 12(4):450–467.

Reference III

[Ennaji et al., 2018] Ennaji, F. Z., El Fazziki, A., El Alaoui El Abdallaoui, H., Benslimane, D., and Sadgal, M. (2018).
Opinion leaders' prediction for monitoring the product reputation.
International Journal of Web Information Systems, 14(4):524–544.

[Feick and Price, 1987] Feick, L. F. and Price, L. L. (1987).
The market maven: A diffuser of marketplace information.
Journal of marketing, 51(1):83–97.

[Festinger, 1954] Festinger, L. (1954).
A theory of social comparison processes.
Human Relations, 7(2):117–140.

[Flynn et al., 1996] Flynn, L. R., Goldsmith, R. E., and Eastman, J. K. (1996).
Opinion leaders and opinion seekers: Two new measurement scales.
Journal of the academy of marketing science, 24:137–147.

[Gnambs and Batinic, 2011] Gnambs, T. and Batinic, B. (2011).
Convergent and discriminant validity of opinion leadership.
Journal of Individual Differences.

[Goldenberg et al., 2006] Goldenberg, J., Lehmann, D. R., Shidlovski, D., and Barak, M. M. (2006).
The role of expert versus social opinion leaders in new product adoption.
Marketing Science Institute Report, 6(4):67–84.

[Gupta, 2023] Gupta, A. (2023).
Are chatgpt and gpt-4 good poker players?–a pre-flop analysis.
arXiv Preprint arXiv:2308.12466.

Reference IV

[Hong et al., 2023] Hong, S., Zheng, X., Chen, J., Cheng, Y., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., et al. (2023).
Metagpt: Meta programming for multi-agent collaborative framework.
arXiv Preprint arXiv:2308.00352.

[InternLM, 2023] InternLM (2023).
Internlm: A multilingual language model with progressively enhanced capabilities.
`https://github.com/InternLM/InternLM`.

[Jungnickel, 2018] Jungnickel, K. (2018).
New methods of measuring opinion leadership: A systematic, interdisciplinary literature analysis.
International Journal of Communication, 12:23.

[Kano et al., 2023] Kano, Y., Watanabe, N., Kagaminuma, K., Aranha, C., Lee, J., Hauer, B., Shibata, H., Miki, S., Nakamura, Y., Okubo, T., et al. (2023).
Aiwolfdial 2023: Summary of natural language division of 5th international aiwolf contest.
In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 84–100.

[Katz, 2015] Katz, E. (2015).
Where are opinion leaders leading us.
International Journal of Communication, 9(1):1023–1028.

[Lan et al., 2023] Lan, Y., Hu, Z., Wang, L., Wang, Y., Ye, D., Zhao, P., Lim, E.-P., Xiong, H., and Wang, H. (2023).
Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay.
arXiv Preprint arXiv:2310.14985.

[Lazarsfeld et al., 1968] Lazarsfeld, P. F., Berelson, B., and Gaudet, H. (1968).
The people's choice: How the voter makes up his mind in a presidential campaign.
Columbia University Press.

Reference V

[Light et al., 2023] Light, J., Cai, M., Shen, S., and Hu, Z. (2023).
Avalonbench: Evaluating llms playing the game of avalon.
arXiv Preprint arXiv:2310.05036.

[Mayer et al., 1995] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995).
An integrative model of organizational trust.
Academy of Management Review, 20(3):709–734.

[Meng, 2024] Meng, J. (2024).
Ai emerges as the frontier in behavioral science.
Proceedings of the National Academy of Sciences, 121(10):e2401336121.

[MistralAI, 2023] MistralAI (2023).
Mistral 7b: The best 7b model to date, apache 2.0.
`https://mistral.ai/news/announcing-mistral-7b/`.

[Nahapiet and Ghoshal, 1998] Nahapiet, J. and Ghoshal, S. (1998).
Social capital, intellectual capital, and the organizational advantage.
Academy of Management Review, 23(2):242–266.

[Noelle-Neumann, 1983] Noelle-Neumann, E. (1983).
The effect of media on media effects research.
Journal of communication, 33(3):157–165.

[Park et al., 2023] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023).
Generative agents: Interactive simulacra of human behavior.
arXiv Preprint arXiv:2304.03442.

Reference VI

[Probst et al., 2013] Probst, F., Grosswiele, L., and Pfleger, R. (2013).
Who will lead and who will follow: Identifying influential users in online social networks-a critical review and future research directions.
Business & Information Systems Engineering, 5(3):179–193.

[Rapanos, 2023] Rapanos, T. (2023).
What makes an opinion leader: Expertise vs popularity.
Games and Economic Behavior, 138:355–372.

[Reynolds and Darden, 1971] Reynolds, F. D. and Darden, W. R. (1971).
Mutually adaptive effects of interpersonal communication.
Journal of Marketing Research, 8(4):449–454.

[Rogers and Cartano, 1962] Rogers, E. M. and Cartano, D. G. (1962).
Methods of measuring opinion leadership.
Public Opinion Quarterly, pages 435–441.

[Shapira et al., 2023] Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. (2023).
Clever hans or neural theory of mind? stress testing social reasoning in large language models.
arXiv Preprint arXiv:2305.14763.

[Singh et al., 2013] Singh, S., Mishra, N., and Sharma, S. (2013).
Survey of various techniques for determining influential users in social networks.
In *2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, pages 398–403. IEEE.

[Talebirad and Nadiri, 2023] Talebirad, Y. and Nadiri, A. (2023).
Multi-agent collaboration: Harnessing the power of intelligent llm agents.
arXiv Preprint arXiv:2306.03314.

[Valente and Pumpuang, 2007] Valente, T. W. and Pumpuang, P. (2007).
Identifying opinion leaders to promote behavior change.
Health education & behavior, 34(6):881–896.

[Wang et al., 2023a] Wang, S., Liu, C., Zheng, Z., Qi, S., Chen, S., Yang, Q., Zhao, A., Wang, C., Song, S., and Huang, G. (2023a).
Avalon's game of thoughts: Battle against deception through recursive contemplation.
arXiv Preprint arXiv:2310.01320.

[Wang et al., 2023b] Wang, Z., Cai, S., Liu, A., Ma, X., and Liang, Y. (2023b).
Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents.
arXiv Preprint arXiv:2302.01560.

[Weimann et al., 2007] Weimann, G., Tustin, D. H., Van Vuuren, D., and Joubert, J. (2007).
Looking for opinion leaders: Traditional vs. modern measures in traditional societies.
International Journal of Public Opinion Research, 19(2):173–190.

[Wu et al., 2023] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. (2023).
Autogen: Enabling next-gen llm applications via multi-agent conversation framework.
arXiv Preprint arXiv:2308.08155.

[Xu et al., 2023a] Xu, Y., Wang, S., Li, P., Luo, F., Wang, X., Liu, W., and Liu, Y. (2023a).
Exploring large language models for communication games: An empirical study on werewolf.
arXiv Preprint arXiv:2309.04658.

[Xu et al., 2023b] Xu, Z., Yu, C., Fang, F., Wang, Y., and Wu, Y. (2023b).
Language agents with reinforcement learning for strategic play in the werewolf game.
arXiv Preprint arXiv:2310.18940.

[Yang et al., 2023] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al. (2023).
Baichuan 2: Open large-scale language models.
arXiv Preprint arXiv:2309.10305.

[Yao et al., 2023] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023).
React: Synergizing reasoning and acting in language models.
arXiv Preprint arXiv:2210.03629.

[Young et al., 2024] Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al. (2024).
Yi: Open foundation models by 01. ai.
arXiv Preprint arXiv:2403.04652.

[Zeng et al., 2022] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al. (2022).
Glm-130b: An open bilingual pre-trained model.
arXiv Preprint arXiv:2210.02414.

[ZhipuAI, 2024] ZhipuAI (2024).
Glm-4 api document.
https://open.bigmodel.cn/dev/api#glm-4.

[Zhu et al., 2023] Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., et al. (2023).
Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory.
arXiv Preprint arXiv:2305.17144.

[Zhuang et al., 2023] Zhuang, Y., Liu, Q., Ning, Y., Huang, W., Lv, R., Huang, Z., Zhao, G., Zhang, Z., Mao, Q., Wang, S., et al. (2023).
Efficiently measuring the cognitive ability of llms: An adaptive testing perspective.
arXiv Preprint arXiv:2306.10512.