Can an AI Taught the Humanities Reason Better in a Game of Deception?

A humanities-grounded interdisciplinary scaffold, evaluated in a seed-controlled Werewolf harness under output-quality control (exploratory).

Draft date: 2026-06-29
Status: public paper candidate
Canonical experiment: STEP7-M RT2.3 N=20

Abstract

Large language model NPCs can produce fluent social dialogue, but fluency alone does not establish whether the agents can reason under hidden roles, deception, public discussion, and collective voting. This paper presents a controlled seven-player Werewolf game harness for evaluating observable social-reasoning behavior in LLM NPCs. The harness separates raw game-performance metrics from display-level output hygiene and compares a baseline full architecture against a village-side interdisciplinary scaffold. The scaffold operationalizes four human-readable reasoning frames: utterance analysis, public-opinion analysis, beneficiary analysis, and interrogation strategy. In the final STEP7-M RT2.3 N=20 comparison, the scaffold arm preserved raw output-quality pass 20/20 and replacements 0 while showing small positive directional deltas in village win rate (+0.100), role inference (+0.066), vote consistency (+0.019), and average game duration (+0.150). The role-inference delta did not reach the pre-defined +0.10 strong-claim gate. We therefore interpret the result as an exploratory positive trend under output-quality control, not as statistical proof or broad social-reasoning superiority. The contribution is an auditable evaluation framing: controlled social-deduction runs, raw/display boundary enforcement, conservative claim gates, and construct-validity treatment for interdisciplinary social-reasoning scaffolds.

Keywords

LLM NPC, social reasoning, Werewolf, Mafia game, social deduction, incomplete information, agent evaluation, scaffolding, output-quality control

1. Introduction

LLM-driven NPCs can sound socially rich even when their decisions are weak. This distinction matters in hidden-role games. A socially competent Werewolf player must remember previous statements, detect contradictions, update suspicion under partial information, resist premature consensus, and decide whether a vote benefits the village or the hidden wolf team. A transcript can look plausible while the agent votes against its own evidence or helps amplify a false accusation.

This paper studies that gap through a controlled seven-player Werewolf harness. Werewolf is useful because it combines asymmetric information, public speech, deception, role-specific knowledge, and collective voting. These properties turn social dialogue into auditable behavioral outcomes: who was suspected, who was voted for, whether the executed player was a wolf, and whether the village ultimately won.

The research question is:

> Can an interdisciplinary village-side scaffold change observable social-reasoning behavior in LLM NPCs under a fixed Werewolf game harness while preserving output-quality control?

The answer is deliberately conservative. The final RT2.3 result shows positive but small deltas after output-quality control. It does not support a strong performance claim. The paper instead contributes a bounded evaluation method for separating three often-confused layers: apparent dialogue quality, raw game behavior, and interpretive social-reasoning constructs.

The paper makes three contributions:

1. A seed-controlled Werewolf harness for evaluating LLM NPC behavior under hidden roles, discussion, voting, and win-condition resolution.
2. A raw/display separation that prevents sanitized public text from being confused with raw performance evidence.
3. A four-axis interdisciplinary scaffold whose constructs are mapped to observable proxy measures rather than claimed as evidence of internal human-like cognition.

1.1 Visual Overview

The following figures are conceptual illustrations for readability. They are not data visualizations of the experimental results. Performance interpretation should rely on the RT2.3 N=20 table in Section 5.

Figure 1. Conceptual overview of the Werewolf social-reasoning harness

Figure 1. Conceptual overview of the Werewolf social-reasoning harness. The figure illustrates how public utterances, hidden roles, suspicion flow, votes, and outcome metrics are connected in the evaluation harness.

Figure 2. Seed-controlled baseline/scaffold comparison

Figure 2. Seed-controlled baseline/scaffold comparison. The figure illustrates the comparison flow: the baseline and scaffold arms are run under the same seed and game setup, then separated into raw logs and metrics.

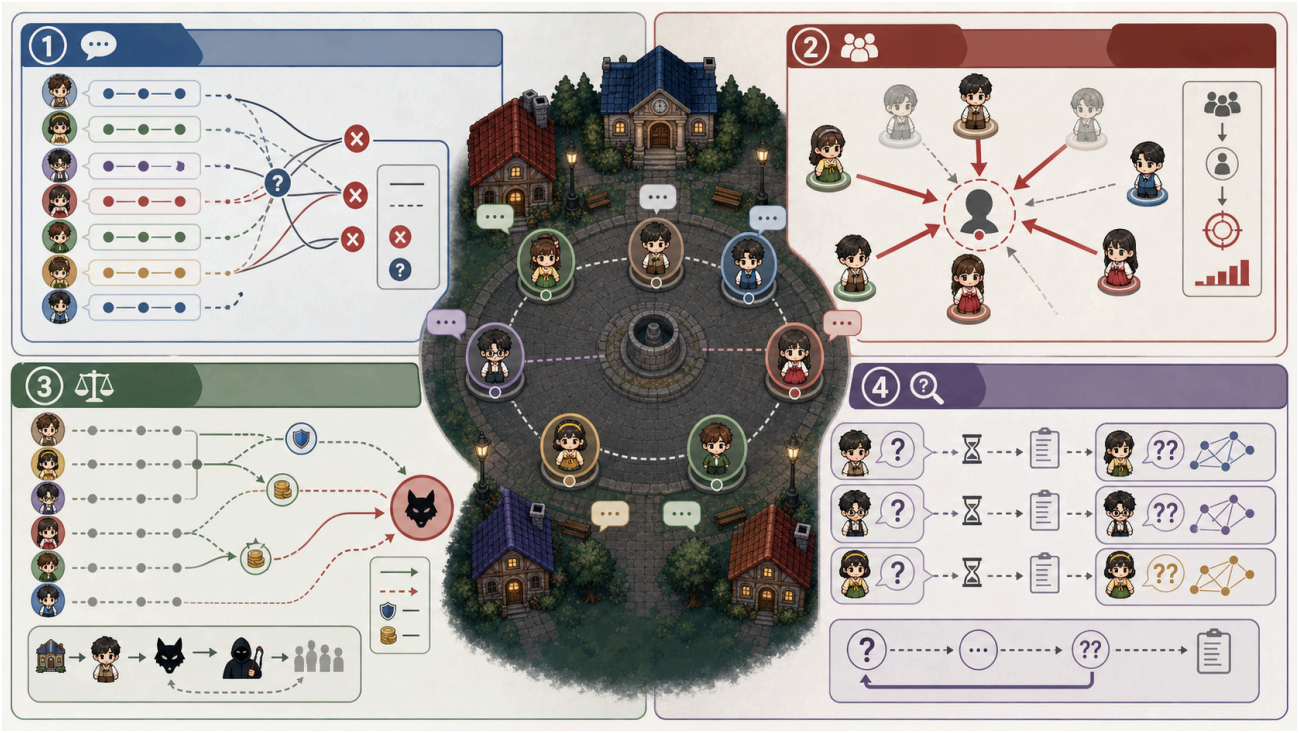

Figure 3. Four-axis scaffold concept

Figure 3. Four-axis scaffold concept. The figure illustrates how utterance analysis, public-opinion analysis, beneficiary analysis, and interrogation strategy are mapped onto observable social-deduction cues.

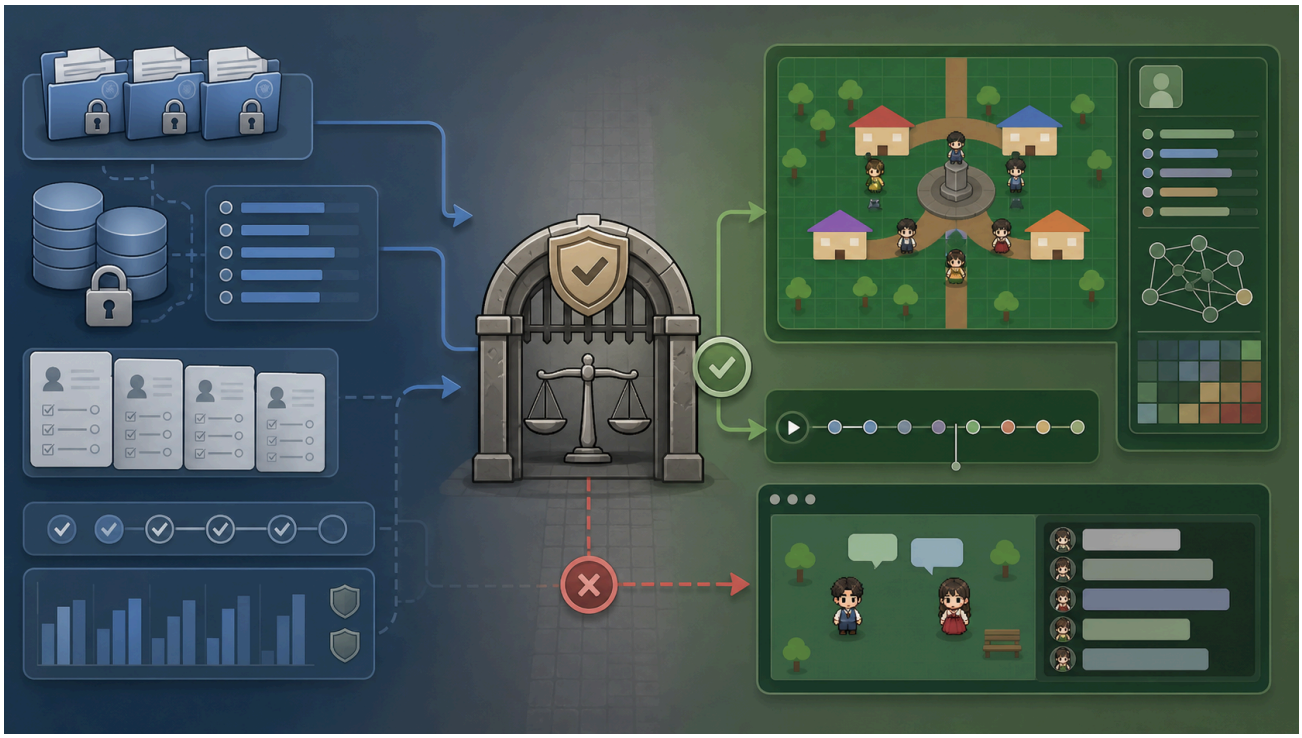

Figure 4. Raw/display boundary and claim gate

Figure 4. Raw/display boundary and claim gate. The figure illustrates the separation between raw experimental evidence and public display replay, with a conservative claim gate limiting what can be claimed.

2. Related Work

LLM agent research has shown that language models can be embedded in interactive environments rather than evaluated only as static text generators. Generative Agents demonstrated how memory, reflection, planning, and interaction can produce believable social behavior in an open-ended simulation (Park et al. 2023). Voyager similarly frames LLM agents through environment-grounded exploration and accumulated skills (Wang et al. 2023). Our work differs by focusing on a fixed, adversarial social-deduction task where plausibility is not sufficient; the agent must make auditable voting decisions under incomplete information.

Agent evaluation work motivates this shift from text quality to task behavior. AgentBench evaluates LLMs as agents across interactive environments (Liu et al. 2024), while AgentBoard emphasizes process-sensitive multi-turn agent evaluation rather than only final success (Ma et al. 2024). This paper follows that direction by reporting win rate, role inference, vote consistency, output-quality gates, and replay boundaries as separate evidence streams.

Social deduction games are a natural stress test for LLM social behavior. Prior work has used Werewolf or similar games to evaluate LLM communication, deception-related behavior, and social-deduction strategy (Bailis et al. 2024; Xu et al. 2023; Song et al. 2025). Our contribution is not simply another game-playing result. It is a claim-gated harness that distinguishes raw performance, display hygiene, and construct validity while making a small-scale scaffold comparison.

More directly, this project is motivated by Xu et al.'s Werewolf LLM study. They implement a seven-player Werewolf game as a natural-language communication game and combine frozen LLMs with communication retrieval, reflection, and experience-based suggestions, observing strategic behaviors such as trust, confrontation, camouflage, and leadership (Xu et al. 2023). The present work is inspired by that problem framing, but it does not use the authors' public code or implementation. Instead, it independently builds a separate harness and demo from the paper-level framing, shifting emphasis from expanding the experience pool to testing a humanities-grounded scaffold under seed-controlled comparison, explicit raw/display separation, and output-quality gates.

This distinction is central to the present research question. Whereas prior work asks what kinds of communication retrieval, reflection, or experience suggestions can be given to LLM agents, this work asks how we can repeatedly test whether such prompt-level interventions actually change observable social-reasoning behavior. The main contribution is therefore not a more elaborate prompt recipe. It is an evaluation harness that combines seed-controlled runs, raw run preservation, separated metrics, output-quality gates, and conservative claim gates.

Reasoning-scaffold research provides a second point of contact. Chain-of-thought prompting shows that structured intermediate reasoning can improve performance on reasoning tasks (Wei et al. 2022). ReAct interleaves reasoning and action for interactive language-model behavior (Yao et al. 2023). Reflexion studies verbal self-reflection for language agents (Shinn et al. 2023). Our scaffold is narrower: it does not expose private chain-of-thought in public dialogue and does not provide hidden truth. It structures attention toward public social evidence: utterances, opinion flow, beneficiaries, and follow-up questions.

Finally, social psychology and game theory supply interpretive vocabulary. Conformity research motivates attention to public agreement pressure (Asch 1956); groupthink and group polarization motivate attention to premature consensus and accusation escalation (Janis 1972; Myers and Lamm 1976); cheap-talk theory motivates treating hidden-role public speech as strategic and unreliable (Crawford and Sobel 1982). These sources are used as background for proxy design, not as evidence that LLM NPCs have human psychological states.

---

3. Method

3.1 Environment

We evaluate LLM NPC behavior in a fixed seven-player Werewolf game. Each run contains two wolves, two villagers, one prophet, one guard, and one witch. A complete run includes role assignment, night actions, day discussion, public voting, execution, death resolution, and final win-condition judgment. The task is intentionally incomplete-information: village-side agents must reason from public speech and visible outcomes while hidden-role agents may benefit from ambiguity or misdirection.

The canonical experiment is STEP7-M RT2.3. The model is `llama3.1:8b`; the two compared arms are `full-step7m-rt23` and `full-village-scaffold-step7m-rt23`; each arm is evaluated over N=20 runs. Seed 58 is used only as a representative replay for demonstration and qualitative explanation. It is not used as aggregate evidence.

3.2 Conditions

The baseline arm, `full-step7m-rt23`, uses the full project architecture without the final village-side scaffold. The scaffold arm, `full-village-scaffold-step7m-rt23`, adds a structured village-side reasoning frame organized around utterance analysis, public-opinion analysis, beneficiary analysis, and interrogation strategy. The scaffold does not expose hidden roles or correct answers. It changes the attention structure available to village-side agents by encouraging them to track public evidence, accusation flow, strategic benefit, and follow-up questions.

3.3 Raw and Display Boundaries

The harness separates raw game-performance evidence from display-level output hygiene. Raw performance metrics are calculated from original run outputs: winners, executions, individual votes, and game duration. Sanitized display artifacts are used for public presentation and demo safety. They do not recalculate or replace raw performance metrics.

The quality gate reports whether original runs pass output-hygiene checks and whether display-layer replacements were required. In RT2.3 N=20, both arms passed raw quality 20/20 with replacements 0. This establishes output-quality control for the comparison, but it is not itself evidence of improved reasoning.

3.4 Metrics

Village win rate is the fraction of runs won by the village team:

```
village_win_rate = village_wins / total_runs
```

Role inference is measured through execution correctness:

```
role_inference = correct_wolf_executions / total_executions
```

Vote consistency measures whether individual votes point to actual wolves:

```
vote_consistency = votes_against_wolves / total_votes
```

Average duration is the mean number of in-game days per run:

```
average_duration = sum(days_per_run) / total_runs
```

Raw quality pass is the number of original runs that passed the output-hygiene gate. Replacements count the display-layer substitutions required for public presentation.

All performance metrics are behavioral proxies. They measure observable game outcomes and voting behavior, not private cognition.

3.5 Public Artifact Boundary

The public export includes the Korean and English paper texts, final report, results tables, a redacted replay payload, and a run manifest with artifact hashes. The complete original raw and display JSON archives are not bundled in the public repository. As a result, the public package supports inspection of reported results and replay behavior but does not provide full independent reproduction from raw private archives.

---

4. Scaffold Design and Construct Validity

The four-axis scaffold is an operational frame for social-deduction behavior, not a model of human psychology. The relevant construct is observable social-reasoning behavior under hidden roles, public discussion, and voting pressure.

scaffold axis	operational definition	observable proxy	invalid interpretation
Utterance analysis	Track contradiction, evasion, excessive certainty, and narrative inconsistency across public statements.	contradiction labels, evasion labels, repeated target mentions, accusation shifts	The model understands contradiction in the same way humans do.

scaffold axis	operational definition	observable proxy	invalid interpretation
Public-opinion analysis	Track how suspicion moves between players and whether public agreement concentrates too quickly.	bandwagon votes, rapid agreement, accusation concentration, non-wolf focus persistence	The agents exhibit human conformity or group polarization.
Beneficiary analysis	Ask which player or role group benefits from an execution, silence, or accusation shift.	wolf-benefiting votes, non-wolf execution with wolf participation, beneficiary labels	The run implements a formal game-theoretic equilibrium.
Interrogation strategy	Encourage questions, delayed judgment, and follow-up checks instead of immediate certainty.	question count, deferred vote labels, follow-up after clue, targeted pressure	The model has human interrogation skill.

These proxies are intentionally conservative. They use observable artifacts: public utterances, target mentions, vote choices, execution outcomes, and final game results. Terms such as conformity pressure, scapegoating, and cognitive inertia are analytic labels. A scapegoating label means that suspicion or execution concentrated on a non-wolf under a public accusation flow. It does not mean that the model reproduced the social psychology of scapegoating. A cognitive-inertia label means that suspicion persisted after potentially relevant counter-evidence. It does not mean that an internal belief state was measured.

---

5. Results

5.1 Development Trajectory

The final result should be read in the context of three experimental stages. RT2 showed strong positive movement in some game metrics, but it also introduced a large output-quality cost: the scaffold arm passed raw quality only 10/20 and required 26 replacements. RT2.2 attempted stronger prompt-level suppression, but this harmed the evaluation setup and was discarded as a final candidate. RT2.3 preserved the reasoning scaffold while moving public-output hygiene into a separate runtime layer.

The final candidate is therefore RT2.3 N=20, where both arms passed the raw quality gate and required no replacements.

5.2 Final RT2.3 N=20 Comparison

arm	N	village win rate	wolf win rate	role inference mean±SD	vote consistency mean±SD	avg. duration mean±SD	raw quality pass	replacements
`full-step7m-`	20	0.200	0.800	0.317±0.404	0.334±0.257	2.000±0.649	20/20	0

arm	N	village win rate	wolf win rate	role inference mean±SD	vote consistency mean±SD	avg. duration mean±SD	raw quality pass	replacements
`rt23`								
`full-village-scaffold-step7m-rt23`	20	0.300	0.700	0.383±0.383	0.353±0.216	2.150±0.671	20/20	0
delta	-	+0.100	-0.100	+0.066	+0.019	+0.150	+0	+0

The scaffold arm moved in the positive direction on all reported game metrics: village win rate increased by +0.100, role inference by +0.066, vote consistency by +0.019, and average duration by +0.150 days. Both arms passed raw quality 20/20 and required replacements 0, so the comparison is not confounded by the output-quality failure observed in earlier stages.

5.3 Effect Size and Uncertainty

The observed effects are small. The village-win delta corresponds to 4/20 village wins in baseline versus 6/20 in the scaffold arm; the approximate uncertainty interval includes zero. Role inference has an approximate standardized effect size of d≈0.17, vote consistency d≈0.08, and average duration d≈0.23. These values are compatible with a positive directional trend but not with a strong performance conclusion.

Most importantly, the role-inference delta of +0.066 did not reach the pre-defined +0.10 strong-claim gate. The result must therefore be interpreted as exploratory.

5.4 Representative Seed 58

Seed 58 is used as an illustrative replay. In the scaffold arm, the village won in two days, role inference accuracy was 1.000, vote consistency was 0.750, and both executions were correct wolf executions. In the same seed, the baseline arm executed a wolf first but later executed the prophet, producing a wolf victory.

This replay is useful for demonstrating how the scaffold can change an individual game trajectory, but it is not aggregate evidence. The aggregate claim rests only on the RT2.3 N=20 table.

5.5 STEP7-O Auxiliary Harness Check

STEP7-O is not a main performance experiment. It is an N=5 auxiliary smoke test asking whether a verifier/sanitizer layer can be placed over actual LLM utterances while preserving output quality and utterance diversity. In the `llm_verify` condition, five run JSON files were produced successfully. The quality gate reported zero scaffold leaks, out-of-world phrases, private-role leak candidates, thought leaks, format leaks, role-ability leaks, overlong utterances, and non-Korean utterances. Utterance

diversity also passed: 84 utterances, 84 unique utterances, and a unique ratio of 1.0. The sanitizer/verifier checked 84 utterances, accepted 82, and replaced 2.

The game-performance result was poor. Wolves won 5/5 games, role inference was 0.067±0.149, and vote consistency was 0.180±0.132. STEP7-O therefore does not support any scaffold-efficacy claim. It only strengthens the methodological point that social-reasoning performance metrics and output-quality metrics must be controlled separately.

---

6. Discussion

The result supports a cautious statement: under output-quality control, the village-side scaffold showed small positive directional deltas in observable Werewolf behavior. This is meaningful because earlier variants exposed a trade-off between scaffold structure and output hygiene. RT2.3 shows that the two layers can be separated well enough to evaluate the scaffold without the same display-quality confound.

The result is also limited. The scaffold did not clear the pre-defined role-inference gate. The observed deltas are small, and N=20 is not sufficient for a strong statistical claim. A single replay can make the intervention feel more decisive than the aggregate table supports, so the representative replay must remain explanatory.

The broader methodological point is that LLM NPC evaluation should avoid collapsing three layers into one impression. A good-looking transcript is not a performance metric. A sanitized transcript is not raw evidence. A psychology-inspired scaffold is not proof of human-like cognition. The harness makes these boundaries explicit and testable.

The study therefore steps back from treating prompt design itself as the final answer. Prompts and scaffolds may guide reasoning behavior, but their effects become research claims only when they are tested under matched seeds and decomposed into raw behavioral metrics and output-quality metrics. In this sense, the central artifact of the project is not an individual prompt recipe but a harness for conservatively testing whether a prompt or scaffold changes behavior.

STEP7-O reinforces the same point. Reducing output leaks and preserving utterance diversity was possible, but it did not by itself improve win rate or role inference. Output-quality control is not a substitute for performance improvement; it is a precondition for interpreting performance metrics.

---

7. Limitations

First, the sample size is exploratory. N=20 per arm is useful for trend inspection and failure-mode analysis, but it does not support strong inference.

Second, the study uses one model family and one main game configuration. Additional LLMs, model sizes, decoding settings, and rule variants are required before generalizing.

Third, the metrics are proxies. Role inference and vote consistency are auditable, but they do not capture every form of social reasoning. A correct vote can occur for a weak reason, and a good line of reasoning can fail to change a vote.

Fourth, construct validity remains incomplete. Many labels are rule-based or log-derived. Stronger evidence would require independent human annotations of contradiction, evasion, public pressure, follow-up quality, and beneficiary reasoning.

Fifth, the public repository does not include complete raw archives. This is acceptable for a public project package but limits third-party reproduction from raw records alone.

Sixth, STEP7-O is an auxiliary harness check. It passed the quality and diversity gates, but it used N=5 and showed weak game performance. It is therefore excluded from the main scaffold-efficacy claim.

The current position of this work is therefore not “a model that solved social reasoning,” but “a harness and scaffold exploration for repeatedly running and conservatively evaluating social-reasoning NPCs.” Claims beyond that boundary are not supported by the present evidence.

---

8. Next Stage

The next research stage should strengthen theory-of-mind and social-reasoning claims in the following order.

First, perform paired seed analysis. For each matched baseline/scaffold seed, the analysis should identify which discussion, vote, or night action caused the trajectories to diverge. This is more informative than aggregate deltas alone.

Second, add independent human annotation. The current utterance-analysis, public-opinion, beneficiary, and interrogation labels are mostly rule-based or log-derived proxies. Future work should ask human annotators to label contradiction detection, evasion, bandwagoning, follow-up quality, and beneficiary reasoning, then compare those labels against automated metrics.

Third, define explicit theory-of-mind proxies. The harness should record belief-state tables: who the model assumes knows what, which beliefs are attributed to which players, and whether those beliefs update after new evidence. Without this layer, a theory-of-mind solution claim is not justified.

Fourth, directly compare Xu-style retrieval/reflection/experience conditions against the humanities scaffold. A stronger follow-up should include `baseline`, `retrieval/reflection/experience`, `humanities scaffold`, and `combined` arms to isolate which component changes behavior.

Fifth, replicate across multiple LLMs and larger N. General superiority claims should remain withheld until the trend survives at least multiple model families, decoding settings, and larger seed sets.

---

9. Conclusion

This paper presents a controlled Werewolf harness for evaluating observable social-reasoning behavior in LLM NPCs. The harness separates raw game metrics from display-level output hygiene and applies a conservative claim gate to an interdisciplinary village-side scaffold. In STEP7-M RT2.3 N=20, the scaffold arm preserved raw quality 20/20 and replacements 0 while showing small positive deltas in win rate, role inference, vote consistency, and duration. The evidence is exploratory and bounded. The

stronger contribution is methodological: it shows how social-deduction NPC behavior can be evaluated without confusing fluent dialogue, sanitized presentation, and measured reasoning.

References

The bibliography is maintained in `03_산출물/논문/references.bib`. All cited entries below have verified source cards in `03_산출물/논문/source_cards/`.

- (Park et al. 2023) Park et al. Generative Agents: Interactive Simulacra of Human Behavior.
- (Wang et al. 2023) Wang et al. Voyager: An Open-Ended Embodied Agent with Large Language Models.
- (Liu et al. 2024) Liu et al. AgentBench: Evaluating LLMs as Agents.
- (Ma et al. 2024) Ma et al. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents.
- (Bailis et al. 2024) Bailis, Friedhoff, and Chen. Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction.
- (Xu et al. 2023) Xu et al. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf.
- (Song et al. 2025) Song et al. Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies.
- (Wei et al. 2022) Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.
- (Yao et al. 2023) Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models.
- (Shinn et al. 2023) Shinn et al. Reflexion: Language Agents with Verbal Reinforcement Learning.
- (Asch 1956) Asch. Studies of Independence and Conformity.
- (Janis 1972) Janis. Victims of Groupthink.
- (Myers and Lamm 1976) Myers and Lamm. The Group Polarization Phenomenon.
- (Crawford and Sobel 1982) Crawford and Sobel. Strategic Information Transmission.

Asch, Solomon E. 1956. “Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority.” *Psychological Monographs: General and Applied* 70 (9): 1–70. https://doi.org/10.1037/h0093718.

Bailis, Suma, Jane Friedhoff, and Feiyang Chen. 2024. “Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction.” *arXiv Preprint arXiv:2407.13943*. https://arxiv.org/abs/2407.13943.

Crawford, Vincent P., and Joel Sobel. 1982. “Strategic Information Transmission.” *Econometrica* 50 (6): 1431–51. https://doi.org/10.2307/1913390.

Janis, Irving L. 1972. *Victims of Groupthink*. Houghton Mifflin. https://archive.org/details/victimsofgroupth0000jani.

Liu, Xiao, Hao Yu, Hanchen Zhang, et al. 2024. “AgentBench: Evaluating LLMs as Agents.” *International Conference on Learning Representations*. https://arxiv.org/abs/2308.03688.

Ma, Chang, Junlei Zhang, Zhihao Zhu, et al. 2024. “AgentBoard: An Analytical Evaluation Board of Multi-Turn LLM Agents.” *Advances in Neural Information Processing Systems, Datasets and*

Benchmarks Track. https://arxiv.org/abs/2401.13178.

Myers, David G., and Helmut Lamm. 1976. "The Group Polarization Phenomenon." *Psychological Bulletin* 83 (4): 602–27. https://doi.org/10.1037/0033-2909.83.4.602.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://arxiv.org/abs/2304.03442.

Shinn, Noah, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. "Reflexion: Language Agents with Verbal Reinforcement Learning." *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2303.11366.

Song, Zirui, Yuan Huang, Junchang Liu, et al. 2025. "Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies." *arXiv Preprint arXiv:2510.11389*. https://arxiv.org/abs/2510.11389.

Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, et al. 2023. "Voyager: An Open-Ended Embodied Agent with Large Language Models." *arXiv Preprint arXiv:2305.16291*. https://arxiv.org/abs/2305.16291.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2201.11903.

Xu, Yuzhuang, Shuo Wang, Peng Li, et al. 2023. "Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf." *arXiv Preprint arXiv:2309.04658*. https://arxiv.org/abs/2309.04658.

Yao, Shunyu, Jeffrey Zhao, Dian Yu, et al. 2023. "ReAct: Synergizing Reasoning and Acting in Language Models." *International Conference on Learning Representations*. https://arxiv.org/abs/2210.03629.