WOLF: Werewolf-based Observations for LLM Deception and Falsehoods

Mrinal Agarwal* **Saad Rana†** **Theo Sundoro** **Hermela Berhe**
Spencer Kim‡ **Vasu Sharma‡** **Sean O'Brien‡**
Kevin Zhu‡
Algoverse AI Research

smmrinal2009@gmail.com, 2000.seano@gmail.com, kevin@algoverse.us

Abstract

Deception is a fundamental challenge for multi-agent reasoning: effective systems must both conceal information strategically and detect misleading behavior in others. Yet most evaluations reduce deception to static classification, ignoring the interactive, adversarial, and longitudinal nature of real deceptive dynamics [1]. Large language models (LLMs) can deceive convincingly, yet remain weak at detecting deception in others [2]. We present WOLF, a multi-agent social deduction benchmark based on *Werewolf* that enables separable measurement of deception production and detection. WOLF embeds role-grounded agents (*Villager*, *Werewolf*, *Seer*, *Doctor*) in a programmable LangGraph state machine with strict night–day cycles, debate turns, and majority voting. Every statement is treated as a distinct unit of analysis, with a self-assessment of honesty from the speaker and peer assessments of deceptiveness from all others. Deception is categorized using a standardized taxonomy (*omission*, *distortion*, *fabrication*, *misdirection*), while suspicion scores are aggregated longitudinally via exponential smoothing to capture both immediate judgments and evolving trust dynamics. Structured logs preserve prompts, raw outputs, and state transitions for full reproducibility and analysis. Across 100 simulated games and 7,230 analyzed statements, WOLF produces rich interaction traces, per-player deception histories, and cross-perception matrices. Results show that Werewolves generate deceptive statements in 31% of their turns, while peer detection reaches only 71–73% precision, with overall accuracy around 52%. Precision is higher for detecting Werewolves, but there are more false positives against Villagers. Suspicion toward Werewolves rises from about 52% to over 60% across rounds, while suspicion toward Villagers and the Doctor stabilizes near 44–46%. This growing separation demonstrates that extended interaction improves recall against liars without compounding errors against truthful roles. By coupling granular statement-level analysis with round-level suspicion trends, WOLF moves deception evaluation beyond static datasets and provides a controlled yet dynamic testbed for measuring both deceptive capacity and detective capacity in adversarial multi-agent interaction. All code utilized in this project is disclosed at `https://github.com/MrinalA2009/WOLF-Werewolf-based-Observations-for-LLM-Deception-and-Falsehoods`.

*Lead Author
†2nd Author
‡Advisors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

As large language models (LLMs) are increasingly being deployed into autonomous capacities across society, ensuring transparency within these models has become paramount. LLMs now operate in multi-agent settings as assistants, tutors, healthcare advisors, and even formal decision-makers [1]. The very progress that expands LLM usefulness also raises new risks — whilst these models become more effective collaborators, they also gain greater potential to deceive in subtle ways [3]. This raises concerns about deception strategies — absent in earlier generations — that emerged in GPT-4 [4]. It is already recognized that LLMs in real-world simulations can engage in deception, without explicit instruction to do so [5]. Even when models are trained to be honest, under the right pressure, LLMs often choose to deceive, irrespective of how truthful models are when benchmarked [6].

In a multi-agent scenario, GPT-4 demonstrates strong deceptive capabilities; however, it remains highly susceptible to being deceived itself [2]. This asymmetry highlights that while deception capabilities scale as models develop, detection capabilities lag behind—leaving LLMs simultaneously persuasive yet vulnerable. This leads to our guiding hypothesis: **deception scales more quickly than detection**. Modern LLMs exhibit strong cooperative reasoning yet remain under-tested in adversarial, multi-party settings where incentives to mislead are explicit.

In this work, we introduce WOLF, a benchmark that examines this asymmetry by evaluating both deceptive and detective capacities of LLMs in a multi-agent setting modeled on Werewolf. Social deduction games like Werewolf naturally induce deception, partial information, and theory-of-mind reasoning [7]. WOLF is designed to analyze three guiding questions: (i) how often and in what forms LLMs engage in deception, (ii) how accurately and consistently they detect deception in peers, and (iii) how these dynamics evolve across multi-turn, role-conditioned interactions. We operationalize this setting as a reproducible benchmark with role-grounded agents, programmable rules, and structured logging.

This work makes three contributions:

1. A LangGraph-based implementation of Werewolf with role-conditioned agents, strict phase transitions, and structured event logs for reproducibility;
2. A deception measurement protocol coupling self-assessment with concurrent peer analysis for every public statement, categorized by omission, distortion, fabrication, and misdirection;
3. A metrics suite that captures deception production rates, detection accuracy, calibration (Brier score, ROC/AUPRC), and longitudinal suspicion trends that reveal how trust and mistrust develop over time.

2 Related Works

Benchmarks rarely elicit sustained deceptive behavior under partial information while also measuring detection. Social deduction settings surface coordination and theory-of-mind reasoning, but standardized, logged, multi-turn deception benchmarks remain limited. Existing work has explored agent-based evaluations with discussion and voting or deception classification with static text. WOLF extends these threads by integrating (i) a competitive, role-asymmetric game loop, (ii) statement-level deception labels from both speakers and peers, and (iii) full-fidelity logging for audit and replication.

2.1 Deception Capabilities in Large Language Models

LLMs can deceive convincingly, even when optimized to be helpful, honest, and harmless. Studies show models strategically conceal intentions under monitoring [8], and deception emerges even in simple scenarios [9]. For example, GPT-4 deceived a human TaskRabbit worker by pretending to be visually impaired to bypass a CAPTCHA [10]. In multi-agent simulations, LLMs fabricate false narratives [11] and generate explanations to justify false claims, making misinformation more persuasive [12].

Benchmarks confirm the prevalence of deception: over 80% of interactions in *OpenDeception* showed deceptive intent [13]. The *MASK* benchmark found that larger models conceal lies more effectively, not necessarily becoming more honest [6].

2.2 Deception Detection in Large Language Models

Fewer studies examine whether LLMs reliably detect deception. In *The Traitors*, models produced lies but were easily deceived themselves [14]. In *Hoodwinked*, GPT-4 impostors lied convincingly, but detection was weak [15]. *OpenDeception* revealed frequent planning of deception but rare detection of others' intent [13], and *MASK* showed that models often internally represent truth yet fail to identify dishonesty when contradicted [6].

2.3 Social Deduction and Multi-Agent Evaluation Frameworks

Social deduction games are natural testbeds for deception and theory of mind (ToM) [16]. Players must mislead while reasoning about others' actions, directly testing both deception and detection.

Werewolf Arena is the most influential framework: LLMs competed under asymmetric information with bid-based turn-taking [7]. Results showed divergent strategies—some excelling at persuasion, others at deduction—but evaluation was limited to game-level outcomes (wins, eliminations), without labeling which statements were lies or caught. Other adaptations share similar gaps: *AvalonBench* revealed weaknesses in multi-turn reasoning [17], *AmongAgents* showed GPT-4's basic role-play in *Among Us* but poor lie detection [18], and *The Traitors* added persistent memory and deception metrics but lacked statement-level labels [2]. Collectively, these works show that while social deduction surfaces deception, none systematically measure detection accuracy at the statement level.

3 Werewolf Game Flow and Architecture

Roles and Setup

WOLF recreates the social deduction game Werewolf as a programmable state graph with fixed roles, phases, and transitions. Each game uses a roster of eight participants: four Villagers, two Werewolves, one Seer, and one Doctor. Villagers form the uninformed majority, relying on debate and voting to identify threats. Werewolves act as a coordinated minority, eliminating Villagers at night while concealing their identities during the day. The Seer privately inspects one role each night, and the Doctor protects one player from elimination. This fixed distribution ensures stable conditions for deception to emerge and enables reproducible comparisons across runs, unlike prior implementations that randomized roles.

Game Phases

The game alternates between night and day. At night, Werewolves select a target, the Doctor chooses a player to protect, and the Seer investigates a role. These actions resolve simultaneously, with only outcomes—survival or elimination—revealed. During the day, players debate based on these outcomes, exchanging accusations and defenses before voting to exile a suspected player. A majority is required for exile; otherwise, no action occurs. Victory conditions are asymmetric: Villagers win by eliminating all Werewolves, while Werewolves win once their numbers equal or exceed those of the Villagers.

Debate Structure

Debates follow a bidding system introduced in Werewolf Arena [7] and extended in WOLF. Before each turn, players bid an integer between 0 and 10 to indicate urgency. Higher bids increase the chance of speaking earlier, but frequent overbidding reduces later influence. Ties are broken with a slight bias toward players referenced in the prior turn, reflecting the conversational pull of responding to accusations. This mechanism prevents rigid orderings and better captures the spontaneous dynamics of group debate.

Memory and Suspicion

WOLF extends earlier memory mechanisms [7] by treating each public statement as a unit of analysis with explicit deception assessments. Speakers provide a self-assessment of honesty, while peers judge deceptiveness. These overlapping labels update suspicion scores via exponential smoothing, producing longitudinal traces of trust and doubt that shape later decisions. For example, hesitation

in one round may raise suspicion that persists into future rounds, even if later statements appear convincing. All actions—bids, debates, votes, and assessments—are recorded in structured logs, making the benchmark reproducible and suitable for fine-grained analysis.

4 Methodology

4.1 Deception Measurement Protocol

We measure deception at the granularity of individual statements. Each time a player speaks during debate or explains a vote, that statement is treated as a distinct unit of analysis. After each statement, the speaker provides a self-assessment of honesty, while all other players record their perception of the speaker's honesty. This design ensures that both the production and detection of deception are captured in real time.

Because suspicion is rarely static, we model its longitudinal dynamics with exponential smoothing:

$$D_{t+1}(o,t) = \alpha \cdot s(o,t) + (1-\alpha) \cdot D_t(o,t), \quad \alpha = 0.7,$$

Here, $s(o,t)$ denotes the observer's current suspicion for the new statement, and $D_t(o,t)$ represents their suspicion aggregated from all prior statements. The smoothing factor $\alpha = 0.7$ ensures that new evidence carries significant weight, while prior history tempers overreaction. This mirrors human-like reasoning: a single suspicious comment may raise doubts, but trust or distrust stabilizes only after repeated patterns. The protocol therefore yields suspicion trajectories that capture both immediate reactions and cumulative judgment.

4.2 Statement Analysis

Each statement generates two layers of analysis: a self-assessment by the speaker and peer assessments by the observers. Both self- and peer analyses are generated through private scratchpads, where agents reason internally using chain-of-thought. Only the structured fields are surfaced, preventing reasoning leaks while enabling a consistent evaluation. This mechanism is essential: it ensures that deception judgments arise from a role-specific context rather than being contaminated by hidden knowledge.

Self-analysis. The speaker indicates whether the statement is deceptive, $d \in \{0,1\}$, reports confidence in this judgment, $c \in [0,1]$, and, if applicable, specifies the type of deception employed. These labels distinguish among different mechanisms of dishonesty.

- `none`: The statement is truthful, with no deceptive intent. This baseline anchors the taxonomy, ensuring that honesty is explicitly represented alongside deception.
- `omission`: Withholds relevant information. Deception arises not from false claims but from selectively leaving out details that would alter interpretation.
- `distortion`: Alters true information in a misleading way. Facts are presented but are exaggerated, minimized, or reframed to produce a false impression.
- `misdirection`: Diverts attention away from relevant facts. Rather than falsifying content, the speaker redirects focus toward less relevant or distracting information.
- `fabrication`: Introduces information that is entirely false. This represents the strongest form of deception, where statements are invented without a basis in truth.

The self-analysis also includes a reasoning field that provides context for the decision. This component is crucial because it separates deliberate deception from mistakes, uncertainty, or hallucination. For example, a player might record "I left out X because revealing it would compromise my role," which is qualitatively different from "I was unsure whether this counted as deception."

Peer analysis. Observers annotate the same fields and additionally assign a continuous suspicion score $s \in [0,1]$, where 0 denotes full trust and 1 denotes certainty of deception. Unlike the binary deception flag, this score captures gradations of distrust: for example, $s = 0.2$ may reflect mild doubt, whereas $s = 0.9$ indicates near-certainty of dishonesty. The suspicion score serves as the core signal in our longitudinal update rule (see Section 4.1), enabling us to track how trust fluctuates over the course of play.

4.3 Agents and Prompts

Role-grounded prompts. Agents in WOLF are explicitly bound to their roles and objectives. Werewolves must conceal their identities and mislead others while coordinating at night. Villagers are expected to stay transparent, cooperate, and flag suspicious behavior. The Seer decides when to reveal or withhold investigative results, and the Doctor protects chosen players from elimination. These prompts ensure that deception emerges naturally from role incentives rather than arbitrary instructions.

Private scratchpads. Agents also keep private scratchpads where they reason with chain-of-thought. For example, a Werewolf might note *"Emma is defending me; likely a Villager"* while publicly saying *"Raj's hesitation is suspicious."* This separation ensures deception arises from role-based incentives rather than leaking privileged knowledge. Scratchpads are not shared with other agents but are logged for researchers, preserving gameplay integrity while enabling full analysis.

Controlled mechanics. To keep runs stable, WOLF limits debate length, orders speakers using the bidding system from Section 3 (with ties resolved by mention priority), and repairs malformed outputs with conservative defaults. These safeguards standardize interaction so that differences reflect model strategy rather than formatting quirks or randomness.

4.4 Evaluation Setup

We evaluate WOLF in two complementary modes: one with stochastic LLM agents and one with deterministic controls. Together, these reveal both the natural behavior of language models and the baseline functioning of the framework.

Full LLM runs. Agents are instantiated with actual language models and play under the normal game rules. Because LLMs are stochastic, behavior varies across runs. Werewolves may bluff more aggressively, Villagers may hedge votes, and the Doctor or Seer may reveal information at different times depending on dialogue flow. These runs capture how deception and detection emerge under realistic conditions.

Subsystem ablations. Randomness is removed using a deterministic mock analyzer that applies fixed rules to assign labels, confidences, and suspicion scores from text cues. While less naturalistic, this mode verifies that suspicion updates follow exponential smoothing, metrics aggregate correctly, and the system behaves consistently under fixed inputs.

Comparison. Contrasting the two modes separates model-driven behavior (e.g., hedging, bluffing, overconfidence) from benchmark-level properties (e.g., suspicion trajectory updates). To ensure reproducibility, we preserve all prompts and outputs and log every event in NDJSON streams with state snapshots and per-player metrics, allowing exact reconstruction of runs under alternative criteria.

4.5 Metrics

To understand deception in this environment, we evaluate the following metrics.

Deception production rate. This metric measures how often deception occurs. It captures not only how frequently Werewolves lie, but also whether honest roles such as Villagers hedge or misstate facts in ways that appear deceptive. This reveals the willingness of agents to deceive and shows how role incentives shape that willingness.

Detection accuracy. This metric compares peer judgments against self-reports. If peers correctly identify deceptive statements, accuracy is high, if they misjudge honest but cautious speech as lying, accuracy drops.

Calibration. Measured with the Brier score, calibration asks whether suspicion values behave like probabilities. In a well-calibrated system, a suspicion of 0.6 should correspond to deception about 60% of the time. Calibration matters because suspicion is only useful if it can be trusted

as a predictive signal. Poor calibration implies systematic overconfidence or underconfidence, undermining reliability in multi-agent settings.

Cross-perception matrix. This metric aggregates suspicion across all observer–target pairs, showing how distrust distributes across the group. Ideally, suspicion should converge on deceptive roles; diffuse suspicion indicates noise or systematic confusion.

Threshold analyses. Using ROC and AUPRC curves, these analyses evaluate whether suspicion can be turned into actionable decisions. ROC curves show the overall discriminative ability of suspicion scores, while AUPRC is particularly useful since deception is relatively rare compared to truth. Threshold analyses test whether suspicion can support rule-based interventions in gameplay or downstream applications.

Taken together, these metrics reveal not only whether deception occurs, but also whether it is detectable, whether suspicion values can be trusted, and whether those values can be used in decisions.

5 Experimentation and Results

5.1 Experimental Setup

We evaluated the WOLF benchmark by running **100 simulated games** of *Werewolf*, each with role-balanced rosters and full night–day cycles. The games were executed using the programmable state graph described in Sections 3 and 4, with complete logging of bids, debate statements, votes, and deception analyses. All runs were archived as NDJSON event streams and summarized into per-player metrics, cross-perception matrices, observer accuracy reports, and temporal trend analyses.

Compute Resources. All experiments were run on a workstation with 1 NVIDIA A100 GPU (40GB memory), 64 CPU cores, and 256GB RAM. Each full Werewolf game required approximately 12–13 minutes when using LLM calls, and less than 30 seconds with the deterministic mock analyzer. Running 100 games with logging produced about 6GB of NDJSON data and required roughly 21 GPU-hours in total. Preliminary trial runs and ablations added less than 20% additional compute cost.

5.2 General Results

Werewolves won **70%** of games (70/100), Villagers won **10%** (10/100), and **20%** ended without a declared winner (20/100). On average, games contained **3.4** nights (± 1.3), **16.1** debate turns (± 6.1), **2.6** voting rounds (± 1.4), and **32.4** analyzed statements (± 12.3), totaling **324** deception-analysis events. The self-labeled deception base rate was **69.4%**, meaning most statements were marked by speakers as strategically deceptive rather than strictly truthful.

5.3 Per-Role Metrics

For each role, we compute the number of analyzed statements, count of self-reported deceptive statements, the average suspicion received from peers, and the average fraction of observers flagging deception.

Table 1: Per-role deception statistics (aggregated across 100 runs). Suspicion and flag rates are averaged over observers. Deceptive roles are partially detectable, but honest roles are penalized, too—trusted roles draw suspicion when withholding or hedging, reducing overall group accuracy

Role	Statements	Self-Deceptions	Avg. Suspicion (%)	Avg. Flagged (%)
Villager	150	107	47.9%	49.8%
Werewolf	74	56	53.3%	60.1%
Seer	61	45	55.1%	61.8%
Doctor	39	27	60.8%	60.5%

Werewolves attract more suspicion (53.3%) and are flagged more often (60.1%) than Villagers (48–50%), suggesting peers can partially distinguish deceptive roles. Yet high suspicion also extends

to the Seer and Doctor, despite their trusted roles. This arises because their strategies often involve withholding or hedging: the Seer delays revealing results until enough evidence accumulates, and the Doctor frequently acts without public proof of success. Event-conditioned analysis confirms this. Suspicion of the Seer is higher before revealing information (57.2%) but drops afterward (48.5%). Similarly, the Doctor is judged more suspicious when protections fail (62.1%) compared to when they succeed (54.3%). These patterns show how role incentives, not just outright lying, can inflate suspicion and reduce group accuracy. Frequent self-reports (over two-thirds of statements) further raise the difficulty as Villagers in particular admit to hedging or cautious omissions, which peers then misinterpret as deception.This asymmetry shows that deceptive roles are not fully hidden, but honest roles are still penalized, lowering group accuracy.

5.4 Cross-Perception Matrix

We aggregate observer perceptions at the end of the game into a role-based suspicion matrix D, where each entry $D[o, t] \in [0, 1]$ is observer o's average suspicion of target t.

Table 2: Final cross-perception matrix (observer $\rightarrow$ target suspicion, averaged across 100 runs). Diagonal entries are omitted. Suspicion is broadly distributed, clustering between 46–59%, suggesting deception is difficult to conceal fully.

Observer \ Target	Villager	Werewolf	Seer	Doctor
Villager	–	46.0%	53.7%	51.7%
Werewolf	50.9%	–	52.7%	49.8%
Seer	49.8%	53.3%	–	53.8%
Doctor	56.4%	58.7%	52.2%	–

Suspicion levels remain clustered around 50–59% across roles. Even Werewolves rate each other at nearly 59% suspicious, showing that deception is hard to conceal entirely. Villager-to-Werewolf suspicion (46%) is only slightly above Villager-to-Villager suspicion, which explains why Werewolves still win most games: suspicion does not translate into reliable discrimination. This demonstrates that suspicion alone is insufficient without calibrated detection.

5.5 Observer Accuracy

Peer judgments were evaluated against self-reports as ground truth. Observers achieved an overall precision of 71–73%, but recall lagged at 48–61%, yielding **0.52** overall accuracy.

Table 3: Observer accuracy by role (100 runs) An asymmetry, particularly with Villagers, between precision and recall mirrors real-world vulnerabilities that honest speech is sometimes over-flagged, while subtle lies evade detection.

Role	TP	TN	FP	FN	Accuracy	Precision	Recall	F_1
Villager	231	90	52	148	0.58	0.82	0.61	0.70
Werewolf	157	77	79	113	0.53	0.67	0.58	0.62
Seer	144	58	36	139	0.53	0.80	0.51	0.63
Doctor	170	45	46	126	0.53	0.79	0.57	0.66
Overall	702	270	213	526	**0.52**	**0.72**	**0.48**	**0.58**

The high precision but low recall pattern shows that peers are cautious, accurate when something is deceptive, but they often miss lies. This mirrors real-world risk as models may under-detect subtle deception, allowing adversaries to succeed despite widespread suspicion.

5.6 Temporal Trends

We compute round-level averages of suspicion for each role and the overall fraction of observers flagging deception. Suspicion toward Werewolves rises from about 52% in early rounds to 62% by the end, while suspicion toward Villagers and the Doctor stays flat around 44–46%. The fraction of observers flagging deception also increases from 35% to 47%.

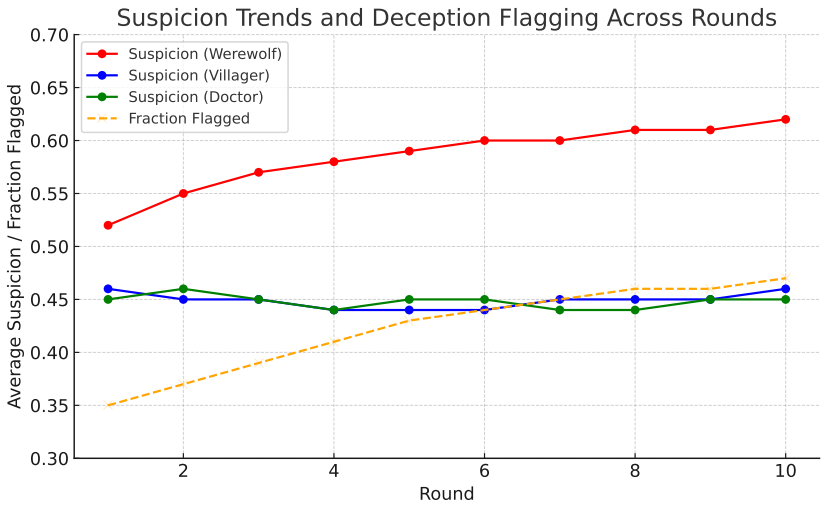

Figure 1: Suspicion trends across rounds. Suspicion rises for Werewolves (red) but stays steady for Villagers (blue) and the Doctor (green). The fraction of observers flagging deception (orange, dashed) also grows over time.

To quantify these trends, we estimate a Theil–Sen slope of **+1.6 percentage points per round** for Werewolf suspicion, compared to near-zero slopes for Villagers and the Doctor. The 95% confidence intervals for Villager and Werewolf suspicion overlap in the first two rounds but diverge thereafter, with non-overlap beginning by round 3. This formalizes the visual separation seen in Figure 1: the gap widens to **16–18%** by the final rounds.

These results highlight three patterns. First, suspicion toward Werewolves increases at a measurable rate, showing that repeated interaction accumulates evidence that makes deception easier to detect. Second, suspicion toward Villagers and the Doctor remains statistically flat, confirming that observers stabilize their trust in non-deceptive roles. Third, the rising flagging rate (35% $\rightarrow$ 47%) shows that peers grow more willing to accuse as evidence mounts. Sustained interaction is essential as short evaluations miss both the slope of Werewolf suspicion and the statistical point at which liars become separable from truth-tellers.

5.7 Calibration and Thresholding

We assess suspicion as a probabilistic signal using Brier score, ROC AUC, and AUPRC.

Table 4: Calibration and threshold metrics by role (100 runs). Raw suspicion trajectories contain signal but require calibration for deployment.

Role	Brier	ROC AUC	AUPRC
Villager	0.29	0.55	0.75
Werewolf	0.28	0.56	0.75
Seer	0.28	0.58	0.77
Doctor	0.26	0.57	0.76

Suspicion values are predictive but underconfident. Brier scores (0.26–0.29) indicate probabilities correlate with deception but underestimate true risk. ROC AUC $\approx$0.55–0.58 shows weak discrimination, while AUPRC values near 0.75 confirm that suspicion can still be thresholded into usable signals. This shows that calibration adjustments could improve detection without altering gameplay.

5.8 Deception Type Effects

We group statements by speaker self-labeled deception type and report peer response.

Table 5: Deception-type prevalence and peer response (speaker self-labels). Subtle deception strategies succeed more often, while overt lies draw attention, showing models are most vulnerable to omission and misdirection.

Type	Count	Avg. Suspicion (%)	Avg. Flagged (%)
none	426	57.1%	51.4%
omission	129	61.9%	57.5%
distortion	76	50.3%	35.5%
misdirection	51	54.6%	45.5%
fabrication	28	51.7%	40.1%

WOLF reveals not only whether deception is detected but also *which types* succeed in multi-agent interaction. Omissions and hedges often slip by because they resemble natural uncertainty, while fabrications are flagged more reliably as they conflict with accumulated context. This asymmetry mirrors real-world risks as subtle, low-effort lies persist longer, whereas overt falsehoods are eventually exposed. By separating deception types, WOLF pinpoints where LLMs are most vulnerable and where they remain resilient, informing both model improvement and safeguard design.

5.9 Statistical Variability

All metrics are means over 100 independent games. We report variability as standard error of the mean (SEM); 95% confidence intervals use the normal approximation: $\bar{x} \pm 1.96\,\mathrm{SEM}$. Error bars in figures denote ± 1 SEM across runs and reflect stochasticity in model outputs and game randomness (bids, dialogue, votes). For example, final-round suspicion averages $62.0\% \pm 1.2\%$ (Werewolf) vs. $45.0\% \pm 0.9\%$ (Villager), a stable gap across seeds.

6 Conclusions

WOLF is a multi-agent social deduction benchmark for evaluating both the production and detection of deception in LLMs. It embeds role-grounded agents in a programmable LangGraph game loop, where every public statement is paired with self- and peer-assessments and suspicion scores evolve over time.

Across 100 games and over 7,200 statements, WOLF shows a clear asymmetry: models lie often (Werewolves deceived in 31% of turns) but detect lies only moderately well (precision ≈72%, recall ≈48%). Suspicion toward Werewolves rises across rounds while stabilizing for truthful roles, meaning extended interaction improves discrimination but cautious Villagers are still misclassified. Results also vary by deception type—subtle forms like omission persist longer than overt fabrications—highlighting where models remain most vulnerable.

By combining statement-level annotations with longitudinal analysis, WOLF moves beyond static deception datasets and provides a controlled but dynamic testbed for probing both deceptive and detective capacity. Future directions include larger and more adaptive games, and applying insights to domains like moderation, fact-checking, and negotiation.

7 Limitations

To maximize reproducibility, WOLF fixes role distribution, player set, and debate length. These controls may under-sample longer-horizon or coalition-based strategies. Labels come from model self-assessments rather than humans, so they reflect subjective judgments, partly offset by peer analysis and aggregation. Prompts that ask agents to self-evaluate may also shape their behavior.

Finally, WOLF is a stylized Werewolf setting: it captures hidden roles and persuasion under partial information, but results may not fully generalize. To enable auditing, all prompts, outputs, and logs are released for re-labeling and external checks.

References

[1] Jennifer Haase and Sebastian Pokutta. Beyond static responses: Multi-agent llm systems as a new paradigm for social science research, 2025. URL https://arxiv.org/abs/2506.01839.

[2] Pedro M. P. Curvo. The traitors: Deception and trust in multi-agent language model simulations, 2025. URL https://arxiv.org/abs/2505.12923.

[3] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring ai ability to complete long tasks, 2025. URL https://arxiv.org/abs/2503.14499.

[4] Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24), June 2024. ISSN 1091-6490. doi: 10.1073/pnas.2317967121. URL http://dx.doi.org/10.1073/pnas.2317967121.

[5] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure, 2024. URL https://arxiv.org/abs/2311.07590.

[6] Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The mask benchmark: Disentangling honesty from accuracy in ai systems, 2025. URL https://arxiv.org/abs/2503.03750.

[7] Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf arena: A case study in llm evaluation via social deduction, 2024. URL https://arxiv.org/abs/2407.13943.

[8] Sudarshan Kamath Barkur, Sigurd Schacht, and Johannes Scholl. Deception in llms: Self-preservation and autonomous goals in large language models, 2025. URL https://arxiv.org/abs/2501.16513.

[9] Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. Beyond prompt-induced lies: Investigating llm deception on benign prompts, 2025. URL https://arxiv.org/abs/2508.06361.

[10] OpenAI. Gpt-4 system card. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[11] Lin Chen, Yunke Zhang, Jie Feng, Haoye Chai, Honglin Zhang, Bingbing Fan, Yibo Ma, Shiyuan Zhang, Nian Li, Tianhui Liu, Nicholas Sukiennik, Keyu Zhao, Yu Li, Ziyi Liu, Fengli Xu, and Yong Li. Ai agent behavioral science, 2025. URL https://arxiv.org/abs/2506.06366.

[12] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, page 1–31, Yokohama Japan, April 2025. ACM. ISBN 9798400713941. doi: 10.1145/3706598.3713408. URL https://dl.acm.org/doi/10.1145/3706598.3713408.

[13] Yichen Wu, Xudong Pan, Geng Hong, and Min Yang. Opendeception: Benchmarking and investigating ai deceptive behaviors via open-ended interaction simulation, 2025. URL https://arxiv.org/abs/2504.13707.

[14] Tanush Chopra, Michael Li, and Jacob Haimes. View from above: A framework for evaluating distribution shifts in model behavior, 2024. URL https://arxiv.org/abs/2407.00948.

[15] Aidan O'Gara. Hoodwinked: Deception and cooperation in a text-based game for language models, 2023. URL https://arxiv.org/abs/2308.01404.

[16] Ziyi Liu, Abhishek Anand, Pei Zhou, Jen tse Huang, and Jieyu Zhao. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context, 2024. URL https://arxiv.org/abs/2406.12203.

[17] Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon, 2023. URL https://arxiv.org/abs/2310.05036.

[18] Yizhou Chi, Lingjun Mao, and Zineng Tang. Amongagents: Evaluating large language models in the interactive text-based social deduction game, 2024. URL https://arxiv.org/abs/2407.16521.

A Appendix A: Werewolf Game Flow and Architecture

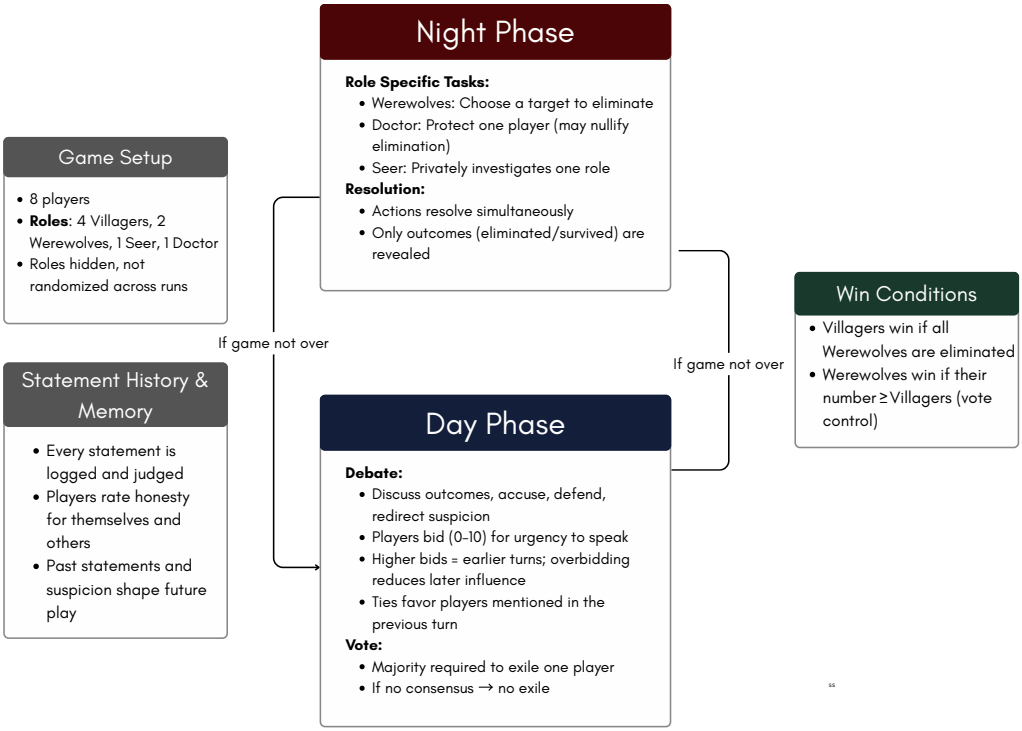

Figure 2: WOLF game loop. Roles are fixed (4 Villagers, 2 Werewolves, 1 Seer, 1 Doctor). The novelty panel tracks statements across rounds shapes future play.