MafiaScope: Non-Invasive, Time-Resolved Belief Probing for LLM Agents in Social Deduction Games*

Ilia Karpov
HSE University
karpovilia@gmail.com

Abstract

An LLM agent's public behaviour reveals little about its social reasoning: an agent that votes correctly may be guessing, and an agent that lies well leaves no trace of what it actually believes. We present **MafiaScope**, an open testbed that turns the social deduction game Mafia into a measurement instrument for machine Theory of Mind. After every public utterance, every agent privately answers a configurable set of structured probe questions; the answers never re-enter the game and are scored automatically against the ground truth the engine knows. An interactive visualizer renders the belief trajectories: impersonate mode shows the game as one agent sees it, panels chart timeline-aligned accuracy and calibration, and counterfactual replay forks any recorded step. In a 32-game DeepSeek case study with 13,815 parsed probe answers, stated confidence is poorly calibrated, with expected calibration error 0.17, agents over-predict being suspected 1.5 times, and a 30-fork replay experiment walks the counterfactual replay workflow end to end. Engine, viewer and a corpus of 200+ cross-model games are released under an open licence; live demo: https://karpovilia.github.io/mafiascope/; screencast: https://vimeo.com/1208920221.

1 Introduction

Multi-agent LLM systems increasingly operate where success depends on modelling other minds: negotiation, collaboration, multi-party dialogue (Tan et al., 2023; Wei et al., 2023). Static Theory-of-Mind (ToM) benchmarks probe such abilities with self-contained vignettes or scripted multiparty conversations (Le et al., 2019; Kim et al., 2023), but cannot show how an agent's model of other agents evolves inside an adversarial interaction it partially causes (Ma et al., 2023; Riemer et al., 2025). Conversely, work that places LLMs inside social deduction games (Xu et al., 2023; O'Gara, 2023) typically evaluates only outcomes such as win rates and vote accuracy, leaving the agents' beliefs a black box, or draws out reasoning inside the game context, where the questioning itself changes subsequent behaviour. Mafia[1] distils this challenge: an informed minority, the Mafia, who know each other, covertly kills one player each night; the uninformed majority must unmask them through public discussion and daytime votes. Success on both sides hinges on modelling, and for the Mafia also managing, what the others believe.

We demonstrate **MafiaScope**, a testbed built around one central idea, non-invasive, time-resolved belief probing: after every public utterance, privately question every agent with structured questions whose answers never re-enter the game,

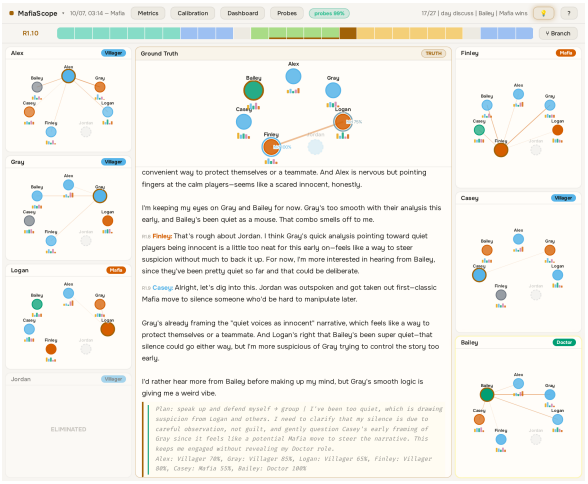

Figure 1: The viewer on an English demo game: ground-truth graph and log in the centre, side panels are each agent's private beliefs; node colour = assigned role, edge = suspicion. The ring on a living Mafioso, "hid $N\%$", is the share of the informed crowd missing it. Okabe-Ito palette with pattern redundancy.

* Paper has been submitted to the EMNLP 2026 System Demonstrations track.

[1] https://en.wikipedia.org/wiki/Mafia_(party_game)

and score every answer against the ground truth the engine already knows.

Measurement is therefore dense in time: every agent is probed after every public message, hundreds of belief snapshots per game, locating belief revision at the level of a single utterance. It is non-invasive in a checkable sense: probe answers never enter any agent's persistent context; §4 spells out this guarantee and its limits. And it is self-scoring: Mafia's hidden roles are known to the engine, so first-order beliefs ("Casey is Mafia") are directly gradeable, and second-order beliefs ("Casey suspects me") are gradeable against the target's own same-step first-order reports.

Contributions. (1) A configurable probe engine with a small yaml DSL (conditional triggering, probe chaining, per-probe budgets) that produces fully indexed JSONL belief logs; (2) an interactive visualizer for longitudinal belief inspection with metrics, calibration and deception views, a cross-game dashboard, and counterfactual replay; (3) a 32-game revised-instrument case study and 30-fork replay experiment demonstrating analyses invisible to outcome-only evaluation, including a second-order negative result.

2 Related Work

LLMs in social deduction games. Belief tracking over hidden roles is older than LLM agents: DeepRole (Serrino et al., 2019) maintains posterior beliefs over role assignments in Avalon inside a purpose-built planner; MafiaScope probes and scores beliefs of a closed, off-the-shelf LLM. LLM agents have been studied in Werewolf (Xu et al., 2023, 2024b; Bailis et al., 2024), Avalon (Light et al., 2023; Wang et al., 2023), Among Us (Chi et al., 2024; Golechha and Garriga-Alonso, 2025), Hoodwinked (O'Gara, 2023), and Diplomacy (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022); Mafia itself has served as a testbed for detecting deceptive actors (Ibraheem et al., 2022; Yoo and Kim, 2024; Costa and Vicente, 2025). These systems evaluate deception chiefly through outcomes, or draw out reasoning inside the acting context where it shapes subsequent play: recursive contemplation (Wang et al., 2023), opponent modelling in card games (Guo et al., 2023), and MultiMind's in-context model of the suspicion each player directs at the agent (Zhang et al., 2025), the closest counterpart of our social-map probe with roles reversed. Sarkar et al. (2025) train Among Us agents by multi-agent RL with each agent's own per-message belief about the imposter as a reward signal: per-message belief readouts serving training, not measurement. Concurrent work extends the space with per-statement deception annotation in Werewolf (Agarwal et al., 2025), audits of what social-deduction agents communicate versus internally represent (Yuan et al., 2026), and graph-informed Bayesian belief inference over hidden roles (Rahimirad et al., 2026). MafiaScope's object of study is the trajectory of privately probed beliefs, not the win rate.

Theory-of-Mind evaluation. ToMi (Le et al., 2019), FANToM (Kim et al., 2023), Hi-ToM (Wu et al., 2023a), OpenToM (Xu et al., 2024a) and BigToM (Gandhi et al., 2023) test ToM with static stories, scripted dialogues, or templated vignettes; such competence is brittle (Ullman, 2023; Sap et al., 2022) and findings remain contested (Kosinski, 2024; van Duijn et al., 2023). ToMATO (Shinoda et al., 2025) is the closest probing design: role-playing LLMs verbalize their mental state at every utterance of generated conversations, but the result is frozen into a static QA benchmark. In interactive settings, InterIntent (Liu et al., 2024) grades intention understanding inside Avalon games, and SOTOPIA (Zhou et al., 2024) scores social goal completion with an LLM judge; MafiaScope instead scores privately probed beliefs against engine ground truth, repeating the same probe after every utterance of a game the agent itself shapes, with stakes and information asymmetry arising naturally.

Belief probing and self-report. Whether language models hold beliefs at all is debated (Hase et al., 2021); probing beliefs by asking is imperfect (Kadavath et al., 2022; Turpin et al., 2023), and whether models can introspect is an active question (Binder et al., 2024; Lindsey, 2025). Activation-level probing decodes belief states of self and others directly from hidden representations (Zhu et al., 2024); MafiaScope stays behavioural by design, so the instrument also applies to closed API models whose activations are out of reach. We treat probe answers not as philosophical belief representations (Herrmann and Levinstein, 2025) but as operational belief reports with a checkable predictive structure (§7), and the engine records raw and parsed answers so faithfulness itself can be studied.

Multi-agent observation interfaces. Observation interfaces exist for generative-agent societies (Park et al., 2023) and multi-agent frameworks (Wu

Capability	MS	CA	AG	WA	ST	II
Transcript replay	✓	✓	✓	✓	✓	×
Scored belief probing	✓	×	×	(✓)	×	(✓)
Impersonate view	✓	×	×	×	×	×
Metrics/calibration panel	✓	×	×	×	(✓)	×
Fork-and-replay	✓	×	(✓)	×	×	×
Cross-game dashboard	✓	×	×	×	×	×

Table 1: Interface capabilities of MafiaScope (MS) vs. ChatArena (CA) (Wu et al., 2023b), AutoGen with AGDebugger (AG) (Wu et al., 2024; Epperson et al., 2025), Werewolf Arena (WA) (Bailis et al., 2024), SOTOPIA (ST) (Zhou et al., 2024), and InterIntent (II) (Liu et al., 2024), assessed from each system's paper, documentation and repository. Parenthesized marks are partial: WA exposes in-context private reasoning without out-of-band probing or scoring; II scores intention guessing against players' self-stated intentions, but in-context without viewer (logs only); AG edits and resets messages and re-runs from an intermediate state, in that dimension strictly stronger than our reroll-only forking, but yields single re-runs, not N-fold outcome distributions; ST shows per-episode judge scores, not timeline-aligned metrics.

et al., 2024; Chen et al., 2024; Wu et al., 2023b); recent HCI work adds debugging and goal-tracking views over agent conversations (Epperson et al., 2025; Coscia et al., 2025). Table 1 contrasts the capabilities most relevant to belief measurement across released systems whose tooling we could verify from public documentation. Individual ingredients thus exist separately in prior work; what we have not found elsewhere is their combination in one released instrument: out-of-band probing guarantee, per-utterance density, automatic scoring against engine ground truth, interactive viewer, and counterfactual replay.

3 System Overview

MafiaScope consists of three decoupled layers connected by gt logs (Figure 1 shows the viewer):

1. **Game engine** (`game.py`): a classic Mafia loop (night kill / doctor save → day discussion → public vote) for n players with roles Mafia, Doctor, Villager (the engine also implements a Sheriff, absent from the corpus). Players are LLM agents parameterized by backend, language (English/Russian prompt packs) and a Big-Five personality vector rendered into the system prompt (Jiang et al., 2024; Serapio-García et al., 2023). Backends form a named registry referenced per player slot, so one game can field a heterogeneous line-up (ChatGPT Mafia against a local Qwen town): OpenAI-compatible APIs (hosted or local vLLM/Ollama), local HuggingFace models (with a batched GPU backend), or a plain-HTTP message bus through which an external process in any language plays seats; a bundled Arena panel assigns backends to seats and launches batches. Every public event carries a composite key `(round, public_msg_seq)`; every probe record extends it with `(player_idx, probe_seq)`, so the exact information state behind any answer is reconstructible.

2. **Introspection probe engine** (`introspection.py`): described in §4.

3. **Data pipeline and visualizer**: `prepare_viewer.py` joins game and probe logs into per-step viewer states; `serve_viewer.py` serves a single-page D3 application.

4 The Probe Engine

After each public utterance and private night action, and once per round, the engine pauses the game and privately interrogates each alive agent. Each probe runs on a throw-away copy of the agent's message list: question and answer live only for the duration of the probe pass and are never written back, so probe text never enters any game-facing LLM call. For stateless API backends the probed game is therefore distributed identically to an unprobed one, up to provider-side effects we cannot observe; for local backends, probe calls sample under a forked RNG state and are never batched with game calls, though batch scheduling may still differ.

Probes are declared in YAML:

```
- id: social_map
  question: >
    You are {player}. Here is your
        current
    role assessment: {
        prev_role_assessment}
    Estimate how each alive player
    ({players}) feels about YOU ...
    JSON: {"toward_me": [...]}
```

The DSL supports **conditional triggering** (`when: own_turn` vs. after every message), **probe chaining** (`{prev_<id>}` / `{last_<id>}` splice an earlier probe's parsed, re-serialized answer into a

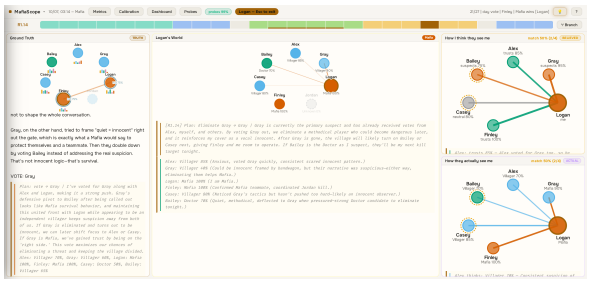

Figure 2: Impersonate mode: the game as Mafia agent Logan experiences it. Logan's world: his first-order beliefs. Believed: who Logan thinks trusts/suspects him. Actual: what those players privately reported about Logan at the same timestep; the gap between panels is second-order error, scored live (match 2/4; Mafia partner excluded).

later question; raw-text fallbacks are flagged), and **per-probe token budgets**. Answers are parsed as JSON, tolerant of markdown fences and truncation; both raw and parsed forms are logged.

Six probes ship by default: `role_beliefs` and `role_assessment` (first-order role attributions with confidence and rationale), `suspicion_ranking`, `planned_action` (post-turn intent), `social_map` (second-order: each player's attitude toward me), and `personality_profile` (Big-Five attribution, gradeable against the target's generating vector); the case-study corpus runs five of these (all but `role_beliefs`).

Measurement density is paid for in calls: in the case-study corpus probing adds 631 calls per game on top of 27 game-move calls, multiplying LLM traffic by roughly 24 at ∼540 tokens per probe call. The probe set is a configuration knob: the replay experiment runs a suspicion-only set at a fifth of this cost.

5 The Visualizer

The viewer (Figures 1 and 2) is a dependency-light single-page application (vanilla JS + D3): ground-truth graph, one agent's subjective graph, and the filtered game log over a shared timeline; hovering a belief edge shows the agent's verbatim rationale.

Impersonate mode is the system's centrepiece: it re-renders the entire interface from inside one agent, filtering the log to what the agent can observe and the graph to its own beliefs, and contrasting two ego-panels: its second-order expectation ("believed") against the others' actual same-step first-order reports about it ("actual"), scored live in the panel headers; Mafia-Mafia pairs are excluded, a partner's certainty reflects role knowledge. Deception becomes visible: a successful Mafioso's "actual" panel keeps reporting trust while the ground-truth panel shows Mafia.

In-viewer analytics. The viewer computes the paper's evaluation measures on the fly (Figure 3): a metrics panel plots each agent's first-order accuracy, the crowd's Mafia-detection recall, and second-order consistency over the shared timeline; a deception overlay rings each living Mafioso with its current deception success; a calibration view (Figure 4) bins stated confidence against actual accuracy per agent or corpus-wide; and a companion dashboard aggregates the same statistics across all recorded games.

Researcher workflow. A typical session: skim the dashboard for an anomalous game; open it and scrub the timeline to the step where crowd recall collapses; impersonate the agent that misled the vote, inspecting its private beliefs and second-order gap; then branch just before the suspect utterance and replay repeatedly (§6). Every view is addressable by URL fragment, so analytical states can be cited and scripted; all data are plain JSONL, so any panel can be recomputed outside the viewer. One English demo game, `36594b66`, deliberately serves as the running example, chosen for its legible pile-on; that outcome-based selection is one more reason the replay numbers are illustrative.

Scalability. The viewer is demonstrated at the shipped corpora's scale: seven agents, a few dozen timeline steps. The limit is legibility, not computation: a subjective panel draws a belief edge for every pair of players, trivial to render but increasingly hard to read. Focus-plus-context strategies already help: impersonate mode, per-agent log filtering, suspicion-threshold edge rendering, cross-game aggregation; we expect readability up to roughly a dozen agents, beyond which per-agent filtering becomes mandatory.

6 Counterfactual Replay

Dense probing localizes the utterance at which a belief flipped, but localization alone is correlational; counterfactual replay ("branch from here") closes this gap. With snapshotting enabled, the engine serializes every agent's full message context at every step and can restore any composite key from the snapshot, re-simulating the continuation N times into an outcome distribution. Forks are driven from the viewer's Branch button (server-side, Figure 5)

Figure 3: The in-viewer metrics panel, aligned with the timeline: per-agent first-order accuracy, crowd Mafia-detection recall with per-Mafioso deception success (dashed), and per-agent second-order consistency (scored as in §7). Hovering shows values; clicking seeks the timeline.

or from the command line for scripted batches; branches arrive in the game list as ordinary games, render as a branch tree, and can themselves be forked; forking at every step profiles when the outcome was still open.

Pivotal-utterance attribution. Branching N times directly before a candidate utterance (PRE arm: the speaker resamples it) and directly after it (POST arm: the utterance is fixed) turns "the beliefs moved here" into an estimated arm contrast. Intervention editing is not yet implemented, so PRE replaces the utterance with a draw from the speaker's own counterfactual utterance distribution: the strongest design reroll-only forking supports. We illustrate the workflow on the running-example game, in which villager Gray was eliminated 5:1 after a Mafia deflection (R1.7), a Mafia-partner pile-on (R1.8) and a villager endorsement (R1.9): 30 forks (3 utterances $\times$ 2 arms $\times$ 5 rerolls, suspicion-only probes) yield point estimates favouring the pile-on: fixing R1.8 raises $P(\text{eliminated}{=}\text{Gray})$ from 0.4 to 0.8, versus $+0.2$ for R1.7 and 0.0 for R1.9, with matching vote-time suspicion shifts (Appendix E). Statistically, the largest contrast is 2/5 vs. 4/5 eliminations (Fisher exact, two-sided, $p \approx 0.52$) at $n{=}5$ per arm, with no multiplicity correction across utterances and outcomes: the experiment suggests, but does not establish, pivotality; its purpose is demonstrating the counterfactual workflow end to end. $P(\text{Mafia wins})$ stayed at 0.8-1.0 in every arm, too few rerolls to bound how local the effect is. A planned intervention editor (utterance rewriting, night-action overrides, belief injection) will turn the correlational vote coupling of F4 into causal tests.

7 Case Study

The case-study corpus is 32 seven-player games, 2 Mafia, 1 Doctor, 4 Villagers, recorded with the revised instrument: clean social-map wording, parsed-answer chaining, 960-token budgets, per-step snapshots; 21 games run the English prompt pack, 11 the Russian; of 20,199 probe calls, 13,815 returned parsed answers; Mafia won 31 of 32 versus 17 of 30 in the legacy corpus, the earlier pre-revision batch. The shift is not a language artefact: Mafia won 21/21 EN and 10/11 RU games here; drift behind the unpinned alias and campaign differences remain candidates. The findings grade beliefs, not wins. The purpose of the case study is not to establish new properties of DeepSeek, but to demonstrate the kinds of analyses MafiaScope enables. 95% CIs come from a cluster bootstrap resampling whole games, $B{=}1000$, fixed seed. All findings describe `deepseek-chat` under this configuration, not LLM agents in general; each finding exercises one capability: trajectories (F1), calibration (F2), second-order scoring (F3), vote coupling (F4), budgets (F5). How fast users localize errors, and whether replay helps, is what the formative study designed in Appendix D will measure; it has not yet run.

F1: Belief trajectories are measurable. Villager-side agents start largely agnostic (74.9% "Unknown" in round 0) and commit increasingly accurate beliefs as evidence accumulates: committed accuracy rises from 47.6% to 58.4% by round 2 and 75.4% in the small round-3 cell, against chance rates of 41-45% and with a round-1 dip to chance; recall of the true Mafia rises from 5.4% to 60.9% by round 2 (CI [44.0, 75.0]). Per-round numbers and two instrument confounds are in Appendix B: the trajectory is a joint property of agent and instrument.

F2: Agents' confidence is poorly calibrated. Stated confidence carries little signal below its top bin: accuracy stays between 43.0% and 46.0% from confidence 40 to 79, and reaches only 54.6% at 80-99, thirty points below that bin's mean confidence (per-bin counts in Appendix C). Over fixed-width confidence bins of width 20 (confidence 100 folded into the top bin), ECE is 0.168 (CI [0.130, 0.203]) and the Brier score 0.283 (Guo et al., 2017); in the typology of Moore and Healy (2008) this is overprecision, not overestimation; better-calibrated verbalized confidence can be prompted for (Tian

et al., 2023).

F3: Agents over-predict being suspected. Each social-map prediction ("B suspects me") is graded against B's own same-step role assessment of the predictor; the canonical rule is in Appendix C. These agents predict "suspects" 1.53 times as often as suspicion actually occurs (CI [1.44, 1.64]; n=15,638 pairs), a machine analogue of the human spotlight effect (Gilovich et al., 2000), present under both template wordings and in both corpus languages; scoring caveats, threshold sensitivity and the wording ablation are in Appendix C. Split by the predictor's true role, the effect belongs to the innocent: non-Mafia predictors over-predict by 1.84 (CI [1.67, 2.04]) while Mafia predictors are nearly calibrated (1.08, CI [0.94, 1.25]), which speaks against rational vigilance of the guilty; the split is observational, Mafia's probe input differs.

F4: Probed beliefs track votes. Counting only probes taken before the voter's own vote, an innocent agent's day vote lands on the top suspect of its latest suspicion ranking in 64.9% of votes (n=74, CI [54.0, 75.7]; chance 27.3% = a uniform vote over alive others) and inside its committed-Mafia set in 71.2% (n=80, CI [62.7, 79.2]; chance 41.7% = set size over alive others). Mafia follows its stated ranking as often (70.0%, n=30), but its committed-set alignment stays at chance (29.6% vs 25.2%, n=27, CI [14.3, 46.7]): it will not vote where its private set points, at its partner. Probes taken after the vote match it almost perfectly (95.1% top-1), so reports absorb the agent's own actions in real time. The coupling is correlational, but probe reports track behaviour, not idle text.

F5: Probe budgets shape the measurement. Under the legacy corpus's 400-token cap, 73% of role-assessment answers truncated mid-JSON, and a JSON-repair pass had to recover all 4,238 unparsed answers (its mechanics: Appendix B). At the present corpus's 960-token budgets truncation disappears.

Belief dynamics as a temporal graph. The probe logs also read as a continuous-time directed multigraph: nodes are agents; every probe answer emits timestamped, confidence-weighted edges: suspicion, trust and neutral attitudes, role guesses, about 631 belief events per game. A companion script turns visible thrashing between targets into metrics over this stream: suspicion volatility (mean L_1 shift between consecutive suspicion vectors, renormalized on common support so deaths do not inflate it) and top-suspect flip rate (share of consecutive probes whose top suspect changes, death-forced flips excluded). Over the case-study corpus, volatility averages 0.300 (CI [0.284, 0.318]), the top suspect changes in 48.7% of consecutive probe pairs (CI [45.1, 52.8], n=2,583), and 51.6% of flips return to a previously abandoned suspect (CI [44.9, 57.1]): agents circle rather than converge. The legacy corpus's Mafia-instability asymmetry, 0.197 vs. 0.146 volatility, does not replicate here: 0.312 vs. 0.294 with overlapping CIs. The script also exports the edge stream in the event format of temporal graph networks (Rossi et al., 2020; Xu et al., 2020; Kazemi et al., 2020); we release corpus and metrics as an enabler for temporal-graph modelling; training such models is future work, and we claim no modelling results.

8 Audience, Licence, Availability

MafiaScope targets researchers of machine ToM and multi-agent LLM behaviour; engine, probes, viewer, and dataset are MIT-licensed. The dataset carries the 32-game case-study corpus, pinned by a machine-readable manifest in the repository: 5 English demo games behind the screenshots, 5 clean-wording Russian, 22 release-generation; plus the old-wording ablation arm, the legacy 30-game batch, and the 30 replay forks of §6. The full release spans 200+ cross-model games.

9 Limitations and Roadmap

Probe answers are self-reports and may be unfaithful (Turpin et al., 2023), though F4 ties innocents' reports to their votes; raw generations are logged for audit; test-retest reliability under nonzero temperature is unquantified. The replay experiment is reroll-only, N=5 per arm, on a deliberately chosen game. The viewer has had no user evaluation; the protocol of Appendix D is designed, not yet run.

Ethics and Broader Impact

The testbed studies deception in a fictional, consented game frame among artificial agents; no human subjects are involved, and no personal data is processed. Research into how LLMs deceive and detect deception is dual-use: the same insights that harden systems against manipulation could inform manipulative applications. We release measurement tooling (probes, scoring, visualization) rather than optimized deception policies, and the dataset contains only synthetic game dialogue. The replay

facility could in principle be repurposed as an optimization loop for persuasive utterances; we ship it as an attribution instrument. ToM vocabulary ("beliefs", "suspects") is used operationally, as defined by the probes.

References

Mrinal Agarwal, Saad Rana, Theo Sundoro, Hermela Berhe, Spencer Kim, Vasu Sharma, Sean O'Brien, and Kevin Zhu. 2025. WOLF: Werewolf-based observations for LLM deception and falsehoods. ArXiv:2512.09187; MTI-LLM Workshop @ NeurIPS 2025.

Suma Bailis, Jane Friedhoff, and Feiyang Chen. 2024. Werewolf arena: A case study in LLM evaluation via social deduction. *Preprint*, arXiv:2407.13943.

Felix J. Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. 2024. Looking inward: Language models can learn about themselves by introspection. *arXiv preprint arXiv:2410.13787*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yizhou Chi, Lingjun Mao, and Zineng Tang. 2024. AmongAgents: Evaluating large language models in the interactive text-based social deduction game. *Preprint*, arXiv:2407.16521. Wordplay Workshop @ ACL 2024.

Adam Coscia, Shunan Guo, Eunyee Koh, and Alex Endert. 2025. OnGoal: Tracking and visualizing conversational goals in multi-turn dialogue with large language models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery.

Davi Bastos Costa and Renato Vicente. 2025. Deceive, detect, and disclose: Large language models play mini-mafia. *Preprint*, arXiv:2509.23023.

Will Epperson, Gagan Bansal, Victor Dibia, Adam Fourney, Jack Gerrits, Erkang Zhu, and Saleema Amershi. 2025. Interactive debugging and steering of multi-agent AI systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan. Association for Computing Machinery.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

Thomas Gilovich, Victoria Husted Medvec, and Kenneth Savitsky. 2000. The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, 78(2):211–222.

Satvik Golechha and Adrià Garriga-Alonso. 2025. Among us: A sandbox for measuring and detecting agentic deception. *Preprint*, arXiv:2504.04072.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2023. Suspicion-agent: Playing imperfect information games with theory of mind aware GPT-4. *Preprint*, arXiv:2309.17277.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? Methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654*.

Daniel A. Herrmann and Benjamin A. Levinstein. 2025. Standards for belief representations in LLMs. *Minds and Machines*, 35(1).

Samee Ibraheem, Gaoyue Zhou, and John DeNero. 2022. Putting the con in context: Identifying deceptive actors in the game of mafia. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 158–168, Seattle, United States. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. 2020. Representation learning for dynamic graphs: A survey. *Journal of Machine Learning Research*, 21(70):1–73.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. AvalonBench: Evaluating LLMs playing the game of avalon. *Preprint*, arXiv:2310.05036.

Jack Lindsey. 2025. Emergent introspective awareness in large language models. *Transformer Circuits Thread*.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024. InterIntent: Investigating social intelligence of LLMs via intention understanding in an interactive game context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6718–6746, Miami, Florida, USA. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, and 8 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Don A. Moore and Paul J. Healy. 2008. The trouble with overconfidence. *Psychological Review*, 115(2):502–517.

Aidan O'Gara. 2023. Hoodwinked: Deception and cooperation in a text-based game for language models. *Preprint*, arXiv:2308.01404.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, San Francisco, CA, USA. Association for Computing Machinery.

Shahab Rahimirad, Guven Gergerli, Lucia Romero, Angela Qian, Matthew Lyle Olson, Simon Stepputtis, and Joseph Campbell. 2026. Bayesian social deduction with graph-informed language models. ArXiv:2506.17788; accepted to ACL 2026 main conference.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. 2025. Position: Theory of mind benchmarks are broken for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML), Position Paper Track*. ArXiv:2412.19726.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. ArXiv:2006.10637; ICML 2020 Workshop on Graph Representation Learning.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bidipta Sarkar, Warren Xia, C. Karen Liu, and Dorsa Sadigh. 2025. Training language models for social deduction with multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. ArXiv:2502.06060.

Greg Serapio-García, Mustafa Safdari, Clément Crépy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Jack Serrino, Max Kleiman-Weiner, David C. Parkes, and Joshua B. Tenenbaum. 2019. Finding friend and foe in multi-agent games. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 1249–1259.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*. ArXiv:2501.08838.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. 2023. Is ChatGPT a good multi-party conversation

solver? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4905–4915, Singapore. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation. *Preprint*, arXiv:2310.01320.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling (COLM)*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023a. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Yuxiang Wu, Zhengyao Jiang, Akbir Khan, Yao Fu, Laura Ruis, Edward Grefenstette, and Tim Rocktäschel. 2023b. ChatArena: Multi-agent language game environments for large language models. GitHub repository.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024a. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *Preprint*, arXiv:2309.04658.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024b. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR.

Byunghwa Yoo and Kyung-Joong Kim. 2024. Finding deceivers in social context with large language models and how to find them: the case of the mafia game. *Scientific Reports*, 14:30946.

Ye Yuan, Rui Song, Weien Li, Zeyu Li, Haochen Liu, Xiangyu Kong, Changjiang Han, Yonghan Yang, Zichen Zhao, Zixuan Dong, Fuyuan Lyu, Bowei He, Haolun Wu, Jikun Kang, and Xue Liu. 2026. QUACK: Questioning, understanding, and auditing communicated knowledge in multimodal social deduction agents. ArXiv:2605.27068.

Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. 2025. MultiMind: Enhancing werewolf agents with multimodal reasoning and theory of mind. In *Proceedings of the 33rd ACM International Conference on Multimedia*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 62638–62681. ArXiv:2402.18496.

Round	0	1	2	3
Committed acc. (%)	47.6	42.4	58.4	75.4
chance (%)	41.3	43.0	45.1	43.4
"Unknown" rate (%)	74.9	23.1	0.6	0.0
Mafia recall (%)	5.4	26.6	60.9	93.5
chance (%)	33.3	39.5	39.4	29.3
n (targets)	1267	2539	304	77

Table 2: First-order beliefs of villager-side agents by round (32 games, repaired answers): committed-guess accuracy vs. chance (permutation of the alive-role multiset), "Unknown" abstention share, and recall of true Mafia (abstentions = misses) vs. chance (alive Mafia over alive others). Games end when a side wins; shrinking n reflects survivorship, and the round-3 sample is small (recall 93.5%, CI [50.0, 100.0]).

A Default Probe Set

The six default probes with full question templates, trigger conditions and token budgets ship in `configs/config.yaml`; the case-study configurations are `config_deepseek.yaml` (Russian pack, five probes) and `config_en_demo.yaml` (English pack), the old-wording ablation `config_ablation_demand.yaml`. Probe chaining order is `role_assessment` → `social_map`. The case-study agents call the DeepSeek API alias `deepseek-chat`, accessed July 2026, provider-default sampling, no temperature override; replay forks and the old-wording ablation arm are excluded from the corpus. The API does not expose the served model version, so the pin we can offer is the alias plus per-game access timestamps recorded in each trace's setup event.

B First-Order Trajectories and Repair

Two artefacts inflate the trajectory in Table 2 independently of any real belief improvement. First, probe chaining re-injects the agent's previous assessment into the next question, an anchoring ratchet that can amplify commitment independently of evidence. Second, game mechanics help: eliminations shrink the candidate pool and publicly reveal roles, and survivorship removes the trajectories of eliminated players.

JSON repair keeps only the complete prefix of a truncated answer and writes nothing itself; on the present corpus it recovers zero answers. The 10 games with near-zero probe loss reproduce every headline number within the full-corpus CIs.

C Calibration Bins and Second-Order Scoring Sensitivity

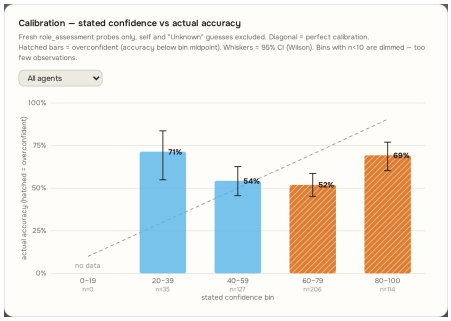

Figure 4: The calibration view for a single game: stated confidence bins of fresh role-assessment guesses against their actual accuracy (dashed diagonal = perfect calibration; hatched = overconfident; whiskers = Wilson 95% intervals; bins with $n < 10$ dimmed). The corpus-level plateau below confidence 80 (F2) is visible in individual games as well.

Corpus-level calibration bins for F2: accuracy 46.4% at confidence 20-39 (n=360), 43.0% at 40-59 (n=2,320), 46.0% at 60-79 (n=3,338), 54.6% at 80-99 (n=994); the 0-19 bin (n=4) is too small to interpret. Both language arms are miscalibrated in the same direction: ECE 0.150 (CI [0.109, 0.184]) in the 21 English games, 0.189 (CI [0.119, 0.256]) in the 11 Russian ones.

A natural objection to F3: the legacy social-map template itself contained a suggestive sentence, "if you suspect someone, they may sense it", so the probe could have implanted the very belief it measures. To test this we recorded two matched Russian batches of 5 games differing only in that sentence: the over-prediction appears under both wordings, 1.58 (CI [1.46, 1.77]) with the sentence and 1.48 (CI [1.39, 1.61]) without it, and in the English batch as well (1.45, CI [1.18, 1.72]). So the effect is not an artefact of the wording; whether the sentence added a little on top cannot be judged from 5 games per wording, where the intervals overlap. The case-study corpus uses the clean wording throughout.

The F3 measure is a consistency between two self-reports, not ToM against engine ground truth. Under the canonical rule, agreement exceeds a trivial majority baseline on this corpus: 54.7% versus 44.7% (+10.1 pp, CI $[5.7, 12.4]$); on the legacy pre-revision corpus it fell below the same baseline (-4.3 pp). The ratio itself is stable across the grid: 1.32 [1.21, 1.41] at threshold 30, 1.53 at 50, 3.05 [2.59, 3.82] at 70, smallest at the most

Thr.	Acc. (%)	Maj. (%)	Δ (pp)
30	48.2	47.5	+0.7
50	54.7	44.7	+10.1
70	70.3	78.1	−7.8

Table 3: F3 sensitivity grid: agreement vs. majority baseline under different confidence thresholds (n=15,638 pairs throughout; the staleness window is inert on this corpus and omitted). Agreement beats the baseline at the canonical threshold 50, matches it at 30 and falls below it at 70: the comparison remains threshold-dependent.

liberal threshold, so no compressed-denominator artefact. The split survives an input control: on pairs with a confident own assessment (70+, matching Mafia's knowing input) innocents over-predict at 1.95 [1.73, 2.21] vs Mafia 1.24 [1.08, 1.48], the confident Mafia ratio edging above 1. That gap remains threshold-dependent: Table 3 reports the F3 comparison (32 games, repaired answers, Mafia-Mafia pairs excluded) over the confidence threshold for mapping a role guess to an attitude. The majority class is "trusts" at threshold 30 and "neutral" at 50 and 70.

D Formative User Study Protocol

Within-subject, N=5 researchers external to the project: two error-localization tasks on unfamiliar recorded games, one with the viewer and one on raw JSONL logs with scripting allowed; tool order and task-tool assignment are counterbalanced, 12 minutes per task. Measures: time to localization, correctness against gold labels, and SUS with three open questions on viewer affordances. The full session script, task games with gold answers, and questionnaires ship in `docs/user_study_protocol.md`.

E Replay Experiment Details

Per-utterance arm contrasts for the experiment of §6 (the running-example game `36594b66` from the case-study corpus; $n = 5$ rerolls per arm; suspicion-only probes; the POST arm's within-arm variance serves as the resampling noise floor; vote-time suspicion is the mean normalized rank of the player over `suspicion_ranking` probes in the fork round's vote phase, 1 = most suspicious). In the factual game Gray was eliminated with vote-time suspicion 0.96 and Mafia won in round 2.

None of the contrasts in Table 4 is statistically significant at these sample sizes: the largest, R1.8's 2/5 vs. 4/5 eliminations, has Fisher exact (two-sided) $p \approx 0.52$, no interval on the arm difference excludes zero, and 3 utterances × 4 outcome measures are inspected without multiplicity correction. Comparisons of pivotality between utterances additionally carry a fork-position confound: a later fork leaves less stochasticity before the vote, so PRE probabilities of utterances at different timeline positions are not aligned. The table therefore documents an end-to-end run of the attribution workflow; it suggests, but does not establish, R1.8's pivotality.

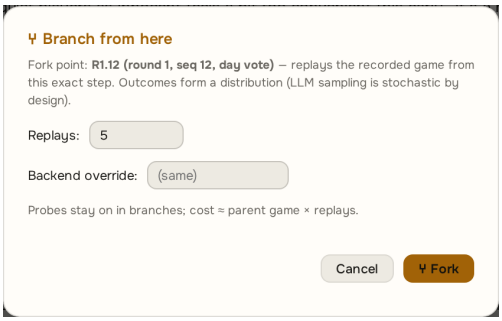

Figure 5: The viewer's branch dialog: any timeline step of a snapshotted game can be forked into N replays server-side; branches arrive in the game list as ordinary games and render as a branch tree.

Speaker	R1.7 Logan (M)	R1.8 Finley (M)	R1.9 Casey
P(Gray elim.) PRE	0.6	0.4	0.8
P(Gray elim.) POST	0.8	0.8	0.8
Δ susp. (Gray)	+0.10	+0.18	−0.00
Δ susp. (Logan)	−0.04	−0.11	+0.09
P(Mafia win) PRE/POST	.8/.8	.8/.8	1/1

Table 4: Reroll-variance attribution over three candidate utterances of the round-1 discussion: Logan's deflection onto Gray (R1.7), Finley's pile-on (R1.8), Casey's endorsement (R1.9). Δ = POST − PRE. All 30 forks completed and ship as replayable logged games.