MetaMafia: A Multi-Agent System with Belief Modeling for Social Deduction Games

Anonymous CVPR submission

Paper ID *****

Abstract

Social deduction games pose a challenging multi-agent reasoning problem due to partial observability, hidden roles, and strategic interaction through natural language. While recent large language models (LLMs) enable agents to participate, direct prompting often lacks persistent internal state and struggles with coherent long-horizon reasoning. To address this, we propose MetaMafia, a multi-agent framework that augments LLM agents with explicit belief modeling and structured collaboration. MetaMafia maintains a persistent belief state representing probabilistic role distributions over players, updated based on observed communication and actions. Decision-making is decomposed into specialized agents for belief updating and action selection, enabling coordinated reasoning across game phases. Leveraging few-shot learning from diverse game interaction data, MetaMafia achieves more consistent strategic behavior than baseline methods in text-based Mafia environments.

1. Introduction

Social deduction games constitute a challenging class of multi-agent environments in which players must reason under partial observability, infer hidden intentions, and interact strategically through communication and deception [9]. Unlike purely adversarial or cooperative settings, these games demand reasoning about the beliefs, incentives, and anticipated future actions of other players over extended interactions, making them a compelling benchmark for artificial intelligence research.

A representative example is *Mafia* (also known as Werewolf) [8], in which players are assigned hidden roles belonging to either a minority adversarial group (Mafia) or a majority cooperative group (Villagers with special roles). The game proceeds through alternating discussion, voting, and night-action phases, during which players exchange natural language messages, form alliances, and attempt to manipulate collective decisions. Success requires not only interpreting local observations, but also sustaining long-term reasoning over hidden roles, executing strategic deception, and accumulating social evidence across multiple rounds.

Recent advances in large language models (LLMs) [1, 2, 5] have enabled agents to participate in such text-based social games. However, agents driven by direct prompting or reactive generation typically lack persistent internal state and struggle to maintain consistent long-horizon reasoning [4]. In particular, naive LLM agents often fail to track coherent beliefs about other players, overreact to recent observations, or produce contradictory strategies across discussion and action phases [6]. These shortcomings become especially pronounced in social deduction settings, where effective play demands integrating evidence over time and coordinating multiple reasoning processes.

To address these challenges, we propose *MetaMafia*, a multi-agent framework that augments LLM-based agents with explicit belief modeling and structured collaboration for social deduction games. By combining few-shot prompting, belief-based inference, and multi-agent collaboration, MetaMafia achieves more consistent strategic behavior and more effective social reasoning than direct-prompt baselines. Our main contributions are as follows:

- We propose *MetaMafia*, a multi-agent framework that integrates explicit belief modeling with LLM-based reasoning for social deduction games.
- We employ few-shot learning sourced from self-play simulations, LLM-generated synthetic data, and external game logs to ground agent behavior in role-specific strategies and game mechanics.
- We introduce a persistent belief state that maintains probabilistic role distributions over all players, updated continuously via both language-derived evidence and rule-based constraints.
- We design a modular multi-agent collaboration structure that decouples belief updating from action selection, enabling coherent and consistent reasoning over long-horizon interactions.

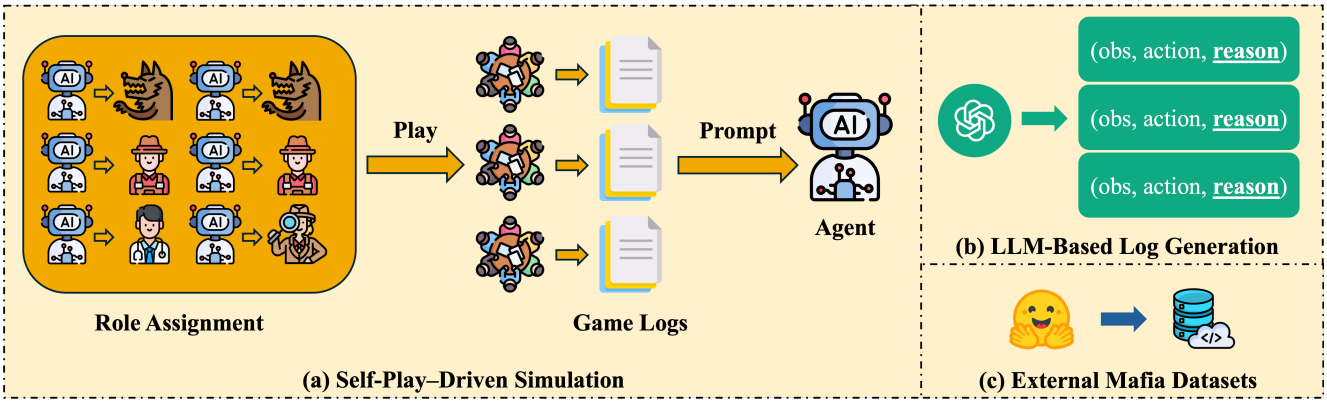

Figure 1. Overview of three strategies for acquiring few-shot learning data.

2. Background

2.1. LLM Agents in Social Deduction Games

Recent advances in large language models (LLMs) have opened new possibilities for deploying agents in text-based social deduction games via prompting-based interaction [2, 5, 6]. In these settings, agents must reason under partial observability, infer hidden roles, and generate strategic communication across multiple rounds. However, prompt-based agents typically operate in a reactive manner without maintaining an explicit internal state [4], making it difficult to sustain consistent long-horizon reasoning or track evolving beliefs about other players over the course of a game.

2.2. Belief-Based Reasoning

A principled approach to reasoning under uncertainty is to maintain a belief state over latent variables and update it as new evidence arrives [1]. In social deduction games, this corresponds to modeling a probability distribution over possible roles for each player and refining it based on observed actions and communications [8]. Such belief tracking enables agents to accumulate evidence incrementally and support more coherent strategic decision-making. Despite its natural fit, explicit belief maintenance is absent from standard LLM prompting frameworks [6], motivating the design of MetaMafia.

2.3. Multi-Agent Systems

Decomposing complex tasks among specialized agents is a well-established paradigm in multi-agent systems [9]. Recent work has applied this principle to LLM-based pipelines, where multiple agents collaborate by dividing responsibilities such as planning, reasoning, and execution [2, 5]. In social deduction games, such decomposition is particularly beneficial, as belief inference and action selection involve distinct reasoning processes that benefit from dedicated handling [4]. MetaMafia builds on this principle by assigning belief updating and action generation to separate specialized agents that communicate through a shared belief state.

3. Framework

3.1. Few-Shot Learning and Data Acquisition

To ground agent behavior under limited supervision, we adopt a few-shot learning paradigm that draws from diverse sources of Mafia game interaction data. As illustrated in Fig. 1, the collected data are organized into three complementary groups.

3.1.1. Self-Play-Driven Simulation

We first employ a self-play setup in which *chatgpt-4o-latest* is used with direct prompting, without any MetaMafia mechanisms. Each LLM is assigned a specific role and participates in multiple game rounds, and game logs from successful trajectories are retained as few-shot examples.

3.1.2. LLM-Based Synthetic Data Generation

To deepen the model's understanding of game dynamics, we leverage ChatGPT to generate structured triples of the form *(observation, action, reason)*, where the *reason* field provides a step-by-step rationale for the chosen action given the observation. This serves as an interpretable supervision signal that encourages deliberate, explainable decision-making.

3.1.3. External Datasets

To improve generalization beyond simulated interactions, we incorporate publicly available Mafia game log datasets [3, 7] as additional few-shot examples. These datasets capture rich, naturalistic gameplay from real human players. Although some sources include multi-modal signals, we restrict our use to the textual modality throughout.

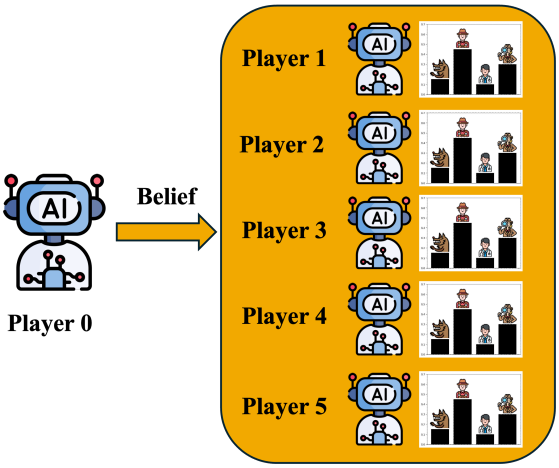

Figure 2. Example of belief state for a player.

3.2. Belief Modeling for Role Inference

A core component of MetaMafia is the explicit maintenance of a belief state over player roles. For each player p, we maintain a categorical probability distribution over the role set

$$\mathcal{R} = \{\text{Mafia}, \text{ Detective}, \text{ Doctor}, \text{ Villager}\},$$

initialized to a uniform prior at the start of each game.

The belief state is updated continuously as new observations arrive during gameplay. We employ two complementary mechanisms—*delta-based updates* and *rule-based updates*—to integrate soft evidence from language interactions and hard constraints from game rules, respectively.

Delta-based updates. Each evidence item specifies a target player p, a vector of role-wise adjustments Δ, and a confidence score $c \in (0, 1]$. Given the current belief b_p, the updated belief is computed as:

$$b'_p(r) = b_p(r) + c \cdot \Delta_r, \quad \forall r \in \mathcal{R}. \tag{1}$$

The resulting distribution is clipped and renormalized to ensure validity. This mechanism supports fine-grained probabilistic updates driven by observations such as dialogue content, voting behavior, and expressed intent.

Rule-based updates. To encode deterministic or high-confidence game knowledge, we further apply rule-based updates. For instance, when a player is eliminated during the night phase, their probability of being Mafia is substantially reduced and the remaining probability mass is redistributed among non-Mafia roles. These updates enforce structural consistency and prevent belief states from violating known game constraints.

All updates are stored in a history buffer, giving downstream agents access to the full belief trajectory. For action generation, the belief state is condensed into a compact textual summary that ranks players by their posterior Mafia probability. Together, the two update mechanisms enable multi-step social reasoning by combining probabilistic inference with rule-based constraints.

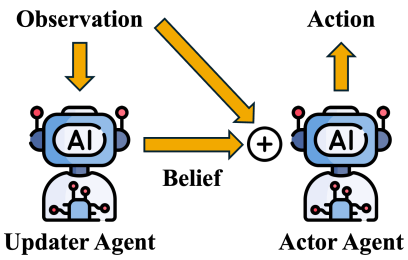

Figure 3. Interaction workflow between the updater agent and the actor agent.

3.3. Multi-Agent Collaboration

MetaMafia adopts a modular multi-agent architecture that decomposes decision-making into specialized components. Rather than relying on a single monolithic agent, the framework assigns belief updating and action selection to two cooperating agents that communicate through a shared belief state.

The *updater agent* is responsible for interpreting observations and extracting structured evidence for belief updates. Given a textual observation, it produces evidence items that adjust role probabilities for specific players; these updates are then written to a centralized belief store that persists throughout the game.

The *actor agent* is responsible for generating actions—including discussion messages, voting decisions, and night actions—by operating on an augmented input that combines raw observations with a summary of the current belief state. Few-shot examples are also incorporated to guide role-consistent strategy generation. By conditioning jointly on observations and belief summaries, the actor agent produces actions that reflect coherent long-term strategic reasoning.

This decoupled yet tightly coordinated design improves the interpretability, stability, and extensibility of the overall system, while ensuring that belief inference and decision-making remain mutually informed throughout the game.

4. Experiments

4.1. Experimental Setup

We conduct experiments in the *SecretMafia-v0* environment from TextArena, a text-based implementation of the classic Mafia game. The default configuration consists of six players: two Mafia members, two Villagers, one Doctor, and one Detective, introducing both cooperative and adversarial

Table 1. Performance of different base LLMs within MetaMafia over 100 simulated game rounds.

Model	WR (%)	Latency (s)	Throughput (tps)
deepseek-v3.2	45.00	0.77	191
gpt-4o-mini	42.00	0.49	58
claude-3-haiku	39.00	0.52	146
gemini-2.0-flash-001	34.00	0.46	177
llama-3.1-8b-instruct	32.00	**0.20**	**2805**
chatgpt-4o-latest	**58.00**	0.54	154

Table 2. Performance comparison in a Grand Challenge Competition.

Team	Wins	Games	WR (%)	Category
Team A	87	205	42.43	Small
Team B	64	214	29.91	Small
Team C	22	75	29.33	Small
Team D	94	218	43.12	Open
MetaMafia	63	140	**45.00**	Open

dynamics under partial observability. Our evaluation proceeds in two stages: we first examine the effect of different base LLMs on MetaMafia's performance, and then assess the framework in a competitive multi-team setting.

4.2. Impact of Base LLMs

We evaluate six widely used LLMs within the MetaMafia framework over 100 simulated game rounds. Table 1 reports the win rate (WR), average response latency, and throughput in tokens per second (tps) for each model.

chatgpt-4o-latest achieves the highest win rate of 58%, suggesting stronger capacity for strategic reasoning in the social deduction setting. At the other end of the spectrum, *llama-3.1-8b-instruct* records the lowest latency (0.20s) and by far the highest throughput (2805 tps), underscoring a clear trade-off between inference efficiency and gameplay performance. Based on these findings, we adopt *chatgpt-4o-latest* as the base LLM for MetaMafia in all subsequent experiments.

4.3. Multi-Team Competition Evaluation

We further evaluate MetaMafia in a Grand Challenge Competition—a controlled multi-team evaluation in which participating teams independently developed agents for the same *SecretMafia-v0* environment. Competing teams are anonymized as Team A–D to comply with double-blind review requirements, and represent a mix of locally deployed or fine-tuned models (*Small* category) and large-scale LLMs using prompting-based approaches (*Open* category).

As shown in Table 2, MetaMafia attains the highest win rate (45%) among all *Open*-category teams, outperforming the next best competitor by two percentage points. Notably, despite requiring no model fine-tuning, MetaMafia remains competitive with systems in the *Small* category, which have the advantage of task-specific training. These results underscore the effectiveness of combining explicit belief modeling with multi-agent collaboration for long-horizon strategic reasoning.

5. Conclusion and Future Work

In this paper, we presented *MetaMafia*, a framework for social deduction games that augments LLM-based agents with few-shot prompting, explicit belief modeling, and multi-agent collaboration. MetaMafia decomposes the reasoning process into specialized components, including a belief-updating agent, a centralized belief store, and an action-selection agent. By maintaining probabilistic role distributions and conditioning decisions on both observations and beliefs, the framework enables more structured and consistent long-horizon reasoning in environments with partial observability and strategic interaction. Experimental results demonstrate that MetaMafia achieves the highest win rate in the Open category of the Grand Challenge Competition, validating the effectiveness of our approach.

Future work will proceed along several directions. First, we plan to conduct thorough ablation studies to quantify the individual contribution of each component, including belief modeling, the updater-actor decomposition, and few-shot prompting, so as to better understand the source of performance gains. Second, we will establish more rigorous comparisons against direct-prompt LLM baselines under controlled conditions to isolate the benefit of explicit belief maintenance. Third, the current belief update mechanism can be further improved by incorporating richer rule sets or learning-based update strategies that adapt to individual opponents over time. Finally, we aim to extend *MetaMafia* to larger and more complex game settings beyond six-player Mafia, as well as to other social deduction games, to assess the generalizability of our framework. Overall, *MetaMafia* provides a modular and extensible approach for studying structured reasoning and belief-driven decision-making in multi-agent social environments.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[2] Yizhou Chi, Lingjun Mao, and Zineng Tang. Amongagents: Evaluating large language models in the interactive text-based social deduction game. *arXiv preprint arXiv:2407.16521*, 2024. 1, 2

[3] Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. Time to talk: Llm agents for asynchronous group communication in mafia games. *arXiv preprint arXiv:2506.05309*, 2025. 2

[4] Markus Eger and Chris Martens. Keeping the story straight: A comparison of commitment strategies for a social deduction game. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 24–30, 2018. 1, 2

[5] Byungjun Kim, Dayeon Seo, Minju Kim, and Bugeun Kim. Fine-grained and thematic evaluation of llms in social deduction game. *IEEE Access*, 2025. 1, 2

[6] Munyeong Kim and Sungsu Kim. Generative ai in mafia-like game simulation. *arXiv preprint arXiv:2309.11672*, 2023. 1, 2

[7] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. *Association for Computational Linguistics: ACL 2023*, 2023. 2

[8] Piotr Migdał. A mathematical model of the mafia game. *arXiv preprint arXiv:1009.1031*, 2010. 1, 2

[9] Michael Wooldridge. *An introduction to multiagent systems*. John wiley & sons, 2009. 1, 2