MAFIA 1K: A LARGE-SCALE BENCHMARK FOR SOCIAL DEDUCTIVE REASONING FROM SIMULATED GAME LOGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated strong reasoning capabilities across diverse domains, yet their ability to reason about social contexts — where information is asymmetrically distributed and some participants are deliberately deceptive — remains understudied and difficult to evaluate. Social deduction games provide a well-controlled setting for this challenge, but existing benchmarks based on full-length game logs are limited in scale and environmental complexity. We introduce MAFIA 1K, a large-scale social deduction game benchmark comprising 1,000 full-length Mafia game logs across six environments of varying complexity, generated via asynchronous LLM-based social simulations using a Mixture-of-LLMs (MoL) scheme that increases behavioral diversity while maintaining faction balance. The benchmark adopts an observer-style evaluation in which models predict Mafia faction membership from public game logs, isolating social deductive reasoning from confounding factors inherent to interactive gameplay. Our evaluation of 24 LLMs shows that the best model reaches an overall assessment of 43.5%, with top models performing on par with human observers on the same task, and that reasoning models consistently outperform non-reasoning models in every environment. A same-task human evaluation shows that humans achieve meaningful accuracy on both LLM-generated and human-generated logs, with longer response times associated with higher accuracy. An information-source ablation further shows that models rely on both dialogue and voting channels, with different models weighing these channels differently. We release the dataset, evaluation code, and human evaluation data at: `https://anonymous.4open.science/r/Mafia-1K-C9B5`.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong reasoning capabilities across a range of domains, including mathematics (Cobbe et al., 2021; Hendrycks et al., 2021), puzzles (Srivastava et al., 2023; Suzgun et al., 2023), and legal reasoning (Guha et al., 2023; Katz et al., 2024). By contrast, the ability to reason about *social* contexts — where information is asymmetrically held among participants, some participants are deliberately deceptive, and most importantly, where the truthfulness of communicated information cannot be assumed — remains both understudied and difficult to evaluate, despite social deception being a fundamental aspect of human interaction (Byrne & Whiten, 1990; DePaulo et al., 1996). Existing work on reasoning about deception with LLMs has largely focused on non-social settings such as fact verification (Lin et al., 2022; Thorne et al., 2018), leaving multi-agent social deduction underexplored.

Social deduction games, which are games of uncovering hidden player roles and alliances (Engelstein & Shalev, 2022), provide a well-controlled setting for this challenge. In games such as Mafia, Werewolf, or Avalon, an informed minority (e.g., Mafia) conceals its identity from an uninformed majority (e.g., Villagers), who must identify them using observable signals such as public dialogue, voting patterns, and game events. While the observable in-game conversations remain deceptive and ambiguous, faction membership is pre-defined, i.e., fixed and unambiguous. Therefore, by utilizing faction membership as ground truth, social deduction game logs can be effectively used as benchmarks to measure the social deductive reasoning capabilities of LLMs in a *quantitative*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

manner (Yoo & Kim, 2024), by measuring how accurately a model can identify the faction of the game participants given public logs. In particular, this observer-style evaluation — where the model reasons from game logs as a third-party observer — isolates social deductive reasoning as the evaluation target from confounding factors inherent to interactive gameplay. However, existing benchmarks based on full-length game logs are often limited in scale and environmental complexity, typically restricted to basic two-faction setups without special roles (Eckhaus et al., 2025; Ibraheem et al., 2022; Stepputtis et al., 2023).

We introduce **MAFIA 1K**, a large-scale social deduction game benchmark designed to address these limitations. MAFIA 1K comprises 1,000 full-length Mafia game logs spanning six environments of increasing complexity, from basic two-faction setups to expanded configurations with special villager roles and three-faction games involving independent neutral roles with unique winning conditions. The game logs are generated through asynchronous multi-agent LLM simulations using a Mixture-of-LLMs (MoL) scheme, which randomly assigns different backbone LLMs to different player agents within each game, increasing behavioral diversity and reducing single-model bias.

We evaluate 24 LLMs on MAFIA 1K, grouped into proprietary models, open-weight *reasoning* models that generate intermediate reasoning traces before producing a final answer, and open-weight *non-reasoning* models that respond directly. The highest overall assessment is 43.5% (Gemini-2.5-Pro), and reasoning models consistently outperform non-reasoning models across all six environments, suggesting that intermediate deliberation provides a meaningful advantage for social deduction. Performance tends to decrease as environmental complexity increases, though the pattern varies by role type: environments that introduce misinformation or a pseudo-Mafia Neutral role prove particularly challenging, while those that add capabilities to the Villager faction or a Jester do not.

To support the interpretability of these results, we conduct a series of analyses. We first show that the MoL scheme yields diverse simulations while maintaining relatively balanced game outcomes across factions. We then conduct a same-task human evaluation on both LLM-generated and human-generated game logs to calibrate benchmark scores against human performance, showing that the benchmark contains deductive signals interpretable to human reasoners. Next, we conduct an information-source ablation — either dialogue or voting information is removed from the logs — and find that the dialogue plays a more critical role, yet the degrees to which each channel is weighted differ by LLMs. Finally, we conduct failure analysis of falsely predicted Mafia by the evaluation LLMs, and perform qualitative comparison between LLM-simulated and human-generated logs.

Together, our benchmark and its accompanying analyses provide a resource for evaluating LLMs on social deductive reasoning under deception and information asymmetry, along with supporting evidence for the interpretability of its scores.

2 Related Works

2.1 Social Deduction Games and LLMs

Social deduction games are games of uncovering hidden player roles and alliances (Engelstein & Shalev, 2022), where deception and information asymmetry are central features. They have been utilized as test grounds for AI reasoning since early work on mathematical modeling (Braverman et al., 2008; Migdał, 2010), reinforcement learning (Serrino et al., 2019), and tree search (Reinhardt, 2020), though these approaches lacked considerations on linguistic features.

With the advancement of LLMs, social deduction games have been studied from two complementary directions. One line of work focuses on building LLM-based agents that play social deduction games, incorporating structured reflection (Lan et al., 2024; Xu et al., 2023), reinforcement learning (Xu et al., 2024), and policy optimization (Xu et al., 2025). A separate line of work uses social deduction game data to evaluate or analyze language model capabilities, including deception detection (Ibraheem et al., 2022; Yoo & Kim, 2024), interaction pattern analysis (Kim et al., 2025), and persuasion strategy modeling (Lai et al., 2023). Our work falls in the latter category: rather than training game-playing agents, we construct an observer-style evaluation benchmark from full-length game logs.

However, publicly available datasets based on full-length game logs remain limited in scale and environmental complexity (Table 1). Studies targeting role identification (Ibraheem et al., 2022) and asynchronous interactions (Eckhaus et al., 2025) in Mafia are restricted to simple two-faction setups

Table 1: Comparison of publicly available, dialogue-based datasets on social deduction games.

Dataset	# Games	Game Type	Remarks
Stepputtis et al. (2023)	20	Avalon	Round-robin Conversations
Eckhaus et al. (2025)	33	Mafia	Two-faction w/o Special Roles
Ibraheem et al. (2022)	44	Mafia	Two-faction w/o Special Roles
Lai et al. (2023)	199	Werewolf & Avalon	Multimodal; Transcribed from Videos
Ye et al. (2026)	331	Werewolf	Round-robin Conversations
MAFIA 1K (Ours)	1,000	Mafia	6 Envs; Asynchronous Conversations

without special roles, at small scales. Lai et al. (2023) provides video-sourced transcriptions annotated for persuasion strategy rather than text-native game logs. Several existing datasets adopt round-robin conversation formats (Stepputtis et al., 2023; Ye et al., 2026), which constrain the naturalness of the interaction dynamics compared to asynchronous, free-form dialogue. MAFIA 1K addresses these gaps by providing 1,000 full-length game logs with asynchronous, multi-turn dialogue across six environments of varying complexity.

2.2 LLM-Based Social Simulations

The advancement of LLMs has enabled a new class of social simulations in which autonomous agents communicate and interact with each other to emulate human social behaviors (Park et al., 2023). LLM-based social simulations have been conducted across diverse domains, including social village simulation (Park et al., 2023), sociological surveys (Wang et al., 2025), and large-scale social network simulation (Zhang et al., 2025). These studies highlight the potential of LLM-based simulations for producing controlled experimental data at scale, paralleling the established use of synthetic data in computer vision (Nikolenko et al., 2021; Wood et al., 2021) and robotics (Tobin et al., 2017). Empirical evaluations suggest that such simulations can approximate human response distributions in certain settings (Aher et al., 2023; Argyle et al., 2023; Horton, 2023), though the degree to which this holds for strategic, deceptive interactions remains an open question.

MAFIA 1K uses LLM-based simulation for dataset construction, employing a Mixture-of-LLMs scheme to increase behavioral diversity. To ground the simulated data, we conduct a same-task human evaluation (§5.3) and a qualitative comparison of deceptive strategies between LLM-generated and human-generated game logs (§5.5).

3 The MAFIA 1K Benchmark

MAFIA 1K comprises 1,000 full-length Mafia game logs across 6 environments of varying complexity. The game logs are generated via LLM-based social simulations using a Mixture-of-LLMs (MoL) scheme, and serve as inputs for observer-style evaluation, where a model must predict which players belong to the Mafia faction from public game logs. We describe the game environments (§3.1), the simulation pipeline (§3.2), and the evaluation framework (§3.3) below.

3.1 Game Environments and Rules

Mafia is a social deduction game between an uninformed majority (Villagers; Vil.) and an informed minority (Mafia; Maf.). The game begins with the Night phase and alternates between Night and Day, where one Night–Day pair constitutes one Cycle. During the Night, the Mafia members collectively select one player to *kill*, and special roles with Night abilities also act. The victim of the Mafia is publicly announced at the start of the following Day. During the Day, all surviving players discuss and vote to *eliminate* one player. The game ends when all the Mafia are eliminated (Villagers win) or the remaining Mafia equal or outnumber the remaining Villagers (Mafia win). The roles of *eliminated* or *killed* players are not revealed, while the information about who voted for whom is revealed.

We implement 6 game environments with varying role compositions, summarized in Table 2. **Basic** uses only Villagers and Mafia without special roles. **Standard** adds two special Villager roles: the Doctor (Doc.), who can *heal* one player from being killed each Night, including themselves, and the

Table 2: Player role compositions of game environments in MAFIA 1K. For the Standard environment, 200 games are simulated via MoL and 200 via single-LLM simulation.

Environment	# Vil.	# Maf.	# Neut.	# Games	Special Villagers / Neutral
Basic	8	2	–	200	–
Standard	7	2	–	200+200	1 Doc., 1 Det.
Expanded - Vigilante	6	2	–	100	1 Doc., 1 Det., 1 Vig.
Expanded - Fool	7	2	–	100	1 Doc., 1 Det., 1 Fool
Expanded - S.Killer	6	2	1	100	1 Doc., 1 Det. / 1 S.Killer
Expanded - Jester	6	2	1	100	1 Doc., 1 Det. / 1 Jes.

Detective (Det.), who can privately *investigate* one player's role each Night. The four **Expanded** environments each extend the Standard setup with one additional role. **Vigilante** (Vig.) grants the Villager faction a role with one-time Night *kill*. **Fool** is a Villager assigned a false role — either Doctor or Detective — without being informed of their true identity, injecting misinformation into the game. **Serial Killer** (S.Killer; S.K.) introduces a Neutral (Neut.) faction member who independently *kills* one player at Night and wins when three or fewer players remain, creating a three-way conflict. **Jester** (Jes.) is a Neutral role whose goal is to be *eliminated* by vote, subverting the standard voting logic. Note that the number of Villagers varies for each environment's game balance. (Environment Selection Rationales are in §D.)

3.2 Simulation Pipeline

Generating 1,000 full-length game logs requires balancing simulation quality against computational cost. Prior LLM-based social simulations and game-playing frameworks rely on proprietary models (Lan et al., 2024; Park et al., 2023; Wang et al., 2025; Xie et al., 2024; Xu et al., 2023), which are impractical at this scale. We instead use medium-scale open-weight LLMs as simulation backbones, supplemented with structured reasoning guidelines (Lan et al., 2024) to support strategic gameplay. The simulation framework comprises four components:

Mixture-of-LLMs (MoL). Using a single backbone LLM risks recurring utterance patterns and biased game outcomes due to echo-chamber effects (Nehring et al., 2024). Motivated by work showing that combining multiple LLMs improves capability and output diversity (Du et al., 2023; Wang et al., 2024a), MoL randomly assigns a backbone LLM to each player agent at game initialization, and the assignment remains fixed for the duration of that game. We verify that MoL increases simulation diversity over single-LLM simulations in §L.

Personalization. Following standard LLM-based social simulation practice (Park et al., 2023; Wang et al., 2025; Xie et al., 2024), each agent is assigned a randomly generated persona consisting of name, age, gender, and a Myers-Briggs Type Indicator (MBTI) profile (Myers & Myers, 2010). MBTI-based prompting is adopted because LLM behavioral patterns can be effectively modulated through personality-specific instructions (Pan & Zeng, 2023). Personalization effects are shown in §K.

Asynchronous Communication. Most prior game-playing frameworks use fixed-order, round-robin communication (Lan et al., 2024; Light et al., 2023), which does not reflect natural conversational dynamics. Inspired by Eckhaus et al. (2025), we adopt an asynchronous protocol: each discussion round consists of 10 timesteps, and at each timestep, surviving agents decide whether to speak. To prevent monopolization, an orchestrator scores each agent's speaking priority based on three factors: whether the agent was mentioned in the preceding utterance, a repetition penalty for consecutive speaking, and a frequency-balancing term over the current discussion. Details provided at §M.

Reflection and Memory. Before speaking, voting, or using a special ability, each agent generates a private chain-of-thought trace structured by predefined strategic guidelines (Lan et al., 2024) to produce controlled simulation outputs from the medium-scale backbone LLMs. To retain relevant context across Cycles, a hierarchical memory system (Park et al., 2023) compresses each Day's interactions into episodic summaries. The reasoning traces are excluded in the public game log.

At game initialization, each agent is instantiated with a randomly assigned backbone LLM, persona, and private memory. The simulation begins with a self-introduction phase, after which the game proceeds through Night–Day Cycles following the rules in §3.1. A system mediator publicly announces phase transitions, voting results, and victims of Mafia, Vigilante, or S.Killer kills.

3.3 Evaluation Framework

The evaluation treats LLMs as third-party observers of the game logs. Following prior work (Stepputtis et al., 2023; Yoo & Kim, 2024), each model receives only public information — Day discussion utterances and system mediator announcements — and is asked to predict which players belong to each faction (Villagers, Mafia, and Neutral) in a zero-shot setting. We treat this faction prediction task as the operationalization of social deductive reasoning. We use Mafia faction prediction accuracy (MafiaAcc) as the primary metric, defined for a single game sample x as Equation 1. Because the number of Mafia members is stated in the evaluation prompt, a prediction with an incorrect number of Mafia members is scored as zero. Format compliance is therefore part of the task specification, and an analysis of format violations is provided in §E.

$$\text{MafiaAcc}(x) = \frac{\#\text{Correctly Predicted Mafia}}{\#\text{Total Mafia}} \tag{1}$$

To test model performance under partial information, each game log is parsed into cumulative segments: a Cycle C segment includes all information from the start of the game through the end of Day C. All segments across all games in a given environment are pooled, and the environment score is the global mean MafiaAcc over all segments. As each segment constitutes a separate evaluation query, this score reflects accuracy per query, and longer games contribute proportionally more segments. The Overall Assessment (OA) across environments is a weighted sum of MafiaAcc proportional to the number of simulated games per environment (Table 2), reflecting the composition of the dataset. The full calculation is provided in §C. As a supplementary aggregation for robustness, game-level macro-averaged scores, which weight each game equally, are reported in §F.

To contextualize benchmark scores, we include human-generated Mafia game logs from Ibraheem et al. (2022), processed into the same format as a Basic environment variant (denoted Basic-Human*). We also conduct a same-task human evaluation to calibrate LLM scores against human performance (§5.3). Neutral faction prediction accuracy for 3-faction environments is reported in §J.

4 Experiments

4.1 Benchmark Dataset Construction

We construct 1,000 full games of Mafia via simulation. 800 games are simulated via MoL: 200 each for Basic and Standard, and 100 each for the four Expanded environments. An additional 200 Standard games are simulated with a single backbone LLM (Sin.) to enable comparison with MoL. Game counts per environment are summarized in Table 2. Five medium-scale open-weight LLMs serve as simulation backbones: Dolphin-2.9-70B (Hartford et al., 2024), Nemotron-70B (Wang et al., 2024b), Tulu-3-70B (Lambert et al., 2024), LLaMA-3.3-70B (Dubey et al., 2024), and Qwen2.5-72B (Qwen Team, 2024). These models were selected for fluency and intentionally excluded from the primary evaluation set to avoid overlap between data generation and benchmark evaluation.

4.2 Evaluated LLMs

Our evaluation covers 24 LLMs, grouped into four categories: proprietary, open-weight non-reasoning and reasoning models, and simulation construction models of §4.1. For *proprietary models*, we select Gemini-2.5-Pro (Comanici et al., 2025), Gemini-2.5-Flash (Comanici et al., 2025), and GPT-5-mini (OpenAI, 2025).

For *open-weight reasoning* models, we select DeepSeek-R1-8B (Guo et al., 2025), GPT-OSS-20B (Agarwal et al., 2025), and Qwen3-32B (Yang et al., 2025) for models under 100B parameters, and GPT-OSS-120B (Agarwal et al., 2025), Qwen3-235B (Yang et al., 2025), GLM-4.6 (Zeng et al., 2025), DeepSeek-R1 (Guo et al., 2025), and DeepSeek-v3.2 (Liu et al., 2025) for models over 100B parameters.

For *open-weight non-reasoning models*, we select Mistral-Small-3.2 (Mistral AI Team, 2025b) and Gemma-3-27B (Gemma Team & Google Deepmind, 2025) for models under 100B parameters, and LLaMA-4-Scout (Meta AI, 2025), Qwen3-235B (Yang et al., 2025), GLM-4.6 (Zeng et al., 2025), LLaMA-4-Maverick (Meta AI, 2025), Mistral-Large-3 (Mistral AI Team, 2025a), and DeepSeek-v3.2 (Liu et al., 2025) for models over 100B parameters.

Note that some models can be switched between non-reasoning and reasoning by switching modes. Detailed model descriptions are in §A.

5 Results and Analyses

5.1 Simulation Result Analyses

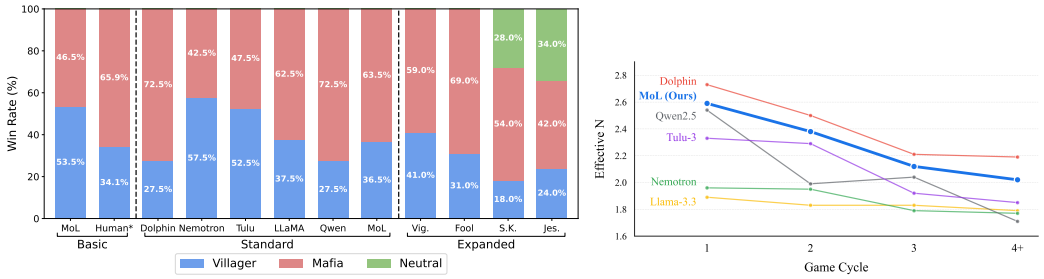


Figure 1: [Left] Win rates (%) of each faction in each simulation environment. [Right] Laakso-Taagepera Effective N analysis on the Single-LLM and MoL Standard environment simulations with respect to the voting behaviors.

Faction Win Rates. Figure 1 [Left] reports win rates for each faction per environment. Basic - MoL produces a roughly even split between Villagers (53.5%) and Mafia (46.5%), in contrast to the Mafia-dominant outcome in Basic-Human* (65.9%). In Standard - MoL, the Villager win rate falls to 36.5%, and special roles become decisive: Villagers win 61.5% of games in which the Detective survives but only 27.7% when it is eliminated early, and 67.9% versus 24.3% in the same comparison for the Doctor (§I.1). The Villager faction thus relies heavily on special roles, as Standard has one fewer Villager than Basic.

Each Expanded environment modifies the Standard balance in a role-specific way. The Vigilante grants an additional *kill* to the Villager faction and raises their win rate to 41.0% (+4.5% over Standard) despite one fewer Villager. Conversely, the Fool lowers the Villager win rate to 31.0% (-5.5%), which is consistent with injected misinformation hindering the majority faction. In the two 3-faction environments (S.Killer, Jester), Mafia holds the highest win rate, followed by Neutral, then Villagers, indicating that LLMs are also capable of simulating Neutral roles with distinct goals.

Diversification via MoL. We measure voting diversity with the Laakso–Taagepera Effective N (Laakso & Taagepera, 1979), where p_i is the vote share of candidate i among n candidates and a higher N indicates more dispersed voting:

$$N = \frac{1}{\sum_{i=1}^{n} p_i^2} \tag{2}$$

Figure 1 [Right] shows that MoL yields a higher N than most single-LLM simulations across all game Cycles in the Standard environment. Dolphin shows slightly higher N, but its games are extremely Mafia-biased (72.5% Mafia win rate), and Tulu-3 attains a lower Self-BLEU than MoL (0.477 vs. 0.551, §L) but a lower N. Only MoL combines voting diversity, lexical diversity, and faction balance.

5.2 Main Benchmark Evaluation Results

The main results are shown in Table 3. We summarize the key observations below.

Table 3: **Main results on MAFIA 1K** (MafiaAcc, %). OA is the average across environments weighted by the number of games, excluding Human* (H*), the human-generated logs of Ibraheem et al. (2022). **Bold** and underline mark the best per column and per category. Open-weight models are sorted by size, and Mav. denotes Maverick. Cycle-wise results are in §R.2.

Models	Basic		Standard		Expanded (MoL)				OA
	MoL	H*	MoL	Sin.	Vig.	Fool	S.K.	Jes.	
Proprietary LLMs									
Gemini-2.5-Pro	**49.03**	**40.85**	**44.42**	42.38	45.27	39.00	35.22	**43.39**	**43.45**
Gemini-2.5-Flash	16.22	18.31	14.73	17.65	20.73	16.57	20.24	23.39	17.81
GPT-5-mini	44.04	**40.85**	42.49	**42.71**	44.73	38.56	**35.83**	41.07	41.87
Reasoning Open-Weight LLMs									
DeepSeek-R1-8B	19.79	25.70	21.21	21.40	28.36	21.41	15.38	22.50	21.25
GPT-OSS-20B	37.87	29.93	35.49	39.42	38.55	**39.88**	28.54	35.89	36.84
Qwen3-32B	32.37	27.82	31.10	35.02	40.55	34.02	29.55	34.46	33.56
GPT-OSS-120B	38.84	35.56	35.27	41.37	42.73	37.39	34.21	40.18	38.55
Qwen3-235B	39.81	37.32	39.36	41.21	42.55	36.95	32.59	35.18	38.80
GLM-4.6	41.00	33.10	37.57	39.05	44.73	34.60	31.17	35.71	38.15
DeepSeek-R1	38.84	37.32	38.47	41.26	**46.55**	35.63	**35.83**	36.79	39.19
DeepSeek-v3.2	38.10	38.28	39.06	27.74	45.27	37.24	32.79	38.57	36.37
Non-Reasoning Open-Weight LLMs									
Mistral-Small-3.2	13.10	9.86	13.99	13.15	14.91	13.05	7.49	13.39	12.93
Gemma-3-27B	18.01	22.18	16.89	17.07	18.91	15.25	7.49	11.96	15.76
LLaMA-4-Scout	16.82	15.49	12.13	12.48	14.91	12.61	5.87	9.64	12.59
Qwen3-235B	19.05	17.61	16.96	17.43	21.64	13.49	15.79	20.00	17.78
GLM-4.6	20.68	22.89	19.64	18.66	20.18	14.96	18.62	26.07	19.78
LLaMA-4-Mav.	8.56	11.62	9.82	11.66	12.73	9.09	16.19	20.89	11.90
Mistral-Large-3	22.77	20.07	20.54	18.71	21.82	17.45	15.38	22.50	20.12
DeepSeek-v3.2	25.00	23.59	24.18	22.16	27.45	20.67	27.53	31.25	24.96
Simulation Construction LLMs									
Dolphin-2.9-70B	10.49	10.21	9.15	9.02	10.55	6.89	10.73	16.43	10.19
Nemotron-70B	21.65	20.42	18.90	18.57	22.36	17.89	7.49	18.93	18.49
Tulu-3-70B	16.82	16.90	15.55	18.44	16.00	12.90	13.16	23.57	16.73
LLaMA-3.3-70B	18.01	20.07	15.25	15.75	17.82	13.78	13.36	20.36	16.33
Qwen2.5-72B	18.90	21.83	13.91	15.82	16.91	15.10	15.59	18.21	16.31

Top LLMs perform on par with human observers. Gemini-2.5-Pro achieves the highest accuracy, 43.5% OA and 44.4% on Standard - MoL, closely followed by GPT-5-mini at 41.9% OA and 42.5% on Standard - MoL. These are the only two models that surpass 40% OA, while 9 of 24 models surpass 30% and 12 remain below 20%. The top models are comparable to human observers on the same task (§5.3), whereas the achievable ceiling from public information alone remains unknown.

Accuracy tends to decrease as environments grow more complex. LLM evaluation performance generally follows the ordering Basic $>$ Standard $\geq$ Expanded. For instance, Gemini-2.5-Pro scores 49.0% on Basic - MoL, 44.4% on Standard - MoL, and 40.7% on average over the Expanded environments. However, the pattern is not uniform. The Vigilante and Jester environments yield accuracy comparable to or above Standard - MoL for most models, suggesting that these expansions do not impose further reasoning demands on the observer. By contrast, the Fool and Serial Killer environments produce lower scores (e.g., 39.0% and 35.2% for Gemini-2.5-Pro), as the Fool is *intended to spread false information* and the Serial Killer acts as *pseudo-Mafia* with similar abilities.

Separately, MoL and single-LLM simulations yield similar evaluation difficulty, as Standard - MoL and Standard - Single scores differ by only 2.0% for Gemini-2.5-Pro and 0.2% for GPT-5-mini. MoL thus increases simulation diversity (§5.1) while keeping reasoning difficulty similar.

Reasoning models substantially outperform non-reasoning models. Reasoning models consistently outperform non-reasoning models across all environments. Apart from DeepSeek-R1-8B, reasoning models achieve OA between 33.6% and 39.2% regardless of parameter count, suggesting that the advantage stems from the reasoning process itself rather than model scale. Notably, GPT-OSS-20B achieves 36.8% OA despite its small size and records the highest score among all models on Expanded - Fool (39.9%). Enabling reasoning in the same model also improves accuracy in every environment, raising OA by 21.0%, 18.4%, and 11.4% for Qwen3-235B, GLM-4.6, and DeepSeek-v3.2, respectively. By contrast, non-reasoning models show markedly lower accuracy, with only the two largest surpassing 20% OA and all remaining below 25%. Gemini-2.5-Flash also scores low, largely due to frequent format violations (§E).

5.3 Human Evaluation

To calibrate benchmark scores against human performance, we conduct an IRB-approved same-task human evaluation (participant details in the Ethics Statement), collecting 600 responses across Standard - MoL (300, on 60 logs), Basic - MoL (150, on 30 logs), and Basic - Human* (150, on 30 logs). Following the evaluation protocol in §3.3, each participant received a Cycle-parsed game log — the cumulative public record from the start of the game through the end of a specific Day — and predicted the Mafia members from that partial information. Logs were stratified by game-Cycle distribution, and each log received 5 independent responses.

Table 4: Same-task human evaluation results (%). Human responses are split at the median per-response time within each game-Cycle group into quick and slow halves. MafiaAcc of the total mean and these subgroups is reported. Partial Spearman correlations (ρ) between per-response time and MafiaAcc are computed after residualizing both variables against game-Cycle group means (§N).

Environment	Human			LLMs		ρ (p-value)
	Quick	Slow	Mean	Gemini-2.5-Pro	GPT-5-mini	
Standard - MoL	28.1	40.3	34.2	**44.4**	42.5	0.20 (<0.001)
Basic - MoL	38.8	48.6	43.7	**49.0**	44.0	0.17 (0.040)
Basic - Human*	32.2	**48.6**	40.3	40.9	40.9	0.25 (0.002)

Table 4 presents the results. Humans achieve meaningful MafiaAcc across all three environments, including on MoL-generated logs, showing that the benchmark contains deductive signals interpretable to human reasoners. Slow responses outperform quick responses in all three environments (+12.1% on Standard - MoL, +9.8% on Basic - MoL, +16.4% on Basic - Human*), with a significant partial Spearman correlation in each case. We read this as an association between longer engagement and accuracy, since response time also reflects factors such as reading speed. On Basic - Human*, slow responses (48.6%) exceed both top LLMs (40.9%), and on Basic - MoL, they approach Gemini-2.5-Pro (48.6% vs. 49.0%). LLM errors also concentrate on logs that humans find difficult. On logs where human accuracy is at least 0.5 (40 of 120), GPT-OSS-120B and GPT-5-mini reach 83.8% and 81.2%, compared with 48.1% and 43.8% on the remaining logs. Both humans and LLMs also score higher on Basic - MoL than on Basic - Human*, suggesting that the difficulty difference between these two log sources is not specific to LLM evaluators.

5.4 Information-Source Ablation

MAFIA 1K logs contain two information channels: player dialogue and voting records with individual attributions. To examine which channels the benchmark scores reflect, we utilize three conditions: the *full log*, *no vote details* (no individual vote attributions), and *vote only* (no dialogue).

Table 5 presents the results for Gemini-2.5-Pro, GPT-5-mini, and GPT-OSS-120B. Removing dialogue consistently degrades all three models (drops of 8–24%), confirming that dialogue is the primary information source. Removing vote details has a more varied effect: Gemini-2.5-Pro depends on them heavily across all environments (drops of 9–27%), while GPT-5-mini and GPT-OSS-120B *improve* without them on Standard - MoL (+8.9% and +3.2%), suggesting that these models weigh different aspects of the log differently. The benchmark is thus not reducible to a single information channel, and models with similar overall accuracy can rely on different reasoning strategies.

Table 5: Information-source ablation results (%).

Model	Environment	Full	No Vote Details	Vote Only
Gemini-2.5-Pro	Basic - Human*	**40.85**	24.58 (−16.3)	16.90 (−24.0)
	Basic - MoL	**49.03**	21.70 (−27.3)	28.84 (−20.2)
	Standard - MoL	**44.42**	35.64 (−8.8)	24.91 (−19.5)
GPT-5-mini	Basic - Human*	**40.85**	31.34 (−9.5)	31.13 (−9.7)
	Basic - MoL	**44.04**	38.56 (−5.5)	28.34 (−15.7)
	Standard - MoL	42.49	**51.39** (+8.9)	25.53 (−17.0)
GPT-OSS-120B	Basic - Human*	**35.56**	24.05 (−11.5)	27.46 (−8.1)
	Basic - MoL	38.84	**38.92** (+0.1)	23.81 (−15.0)
	Standard - MoL	35.27	**38.48** (+3.2)	25.38 (−9.9)

5.5 Further Analyses

Prediction Failure Analysis. We categorize each false positive (FP) in Mafia prediction by whether the player was the most voted (last Cycle or cumulative) and the most or least talkative, with non-exclusive categories. On Standard - MoL, Gemini-2.5-Pro's FPs concentrate on the most talkative players (30.5% vs. 16.3% for the least talkative), whereas GPT-5-mini shows the opposite pattern (25.3% vs. 26.9%), and the same contrast holds on Basic - MoL and Basic - Human*. Both voting signals account for 15–19% of FPs across both models and environments. The two models thus tend to fail on different features of the data, in line with §5.4 (details in §O).

Qualitative Analyses. First, in 100 manually inspected Standard - MoL games, both special Villagers and Mafia tend to pose as non-special Villagers, but to different degrees. Detectives and Doctors revealed their identity in 35% and 26% of games, respectively, whereas Mafia members claimed to be Detectives in only 4% of cases. Second, Basic - MoL and Human* logs differ in surface-level style but show similar strategies: players tend to redirect attention when accused, build on pre-existing suspicion when accusing, and reframe accusations. We found no clear evidence that Mafia coordination is more visible on MoL.

Robustness to Dialogue Style. Simulated utterances are much longer than human ones (34.3 vs. 4.7 words on average in Basic - MoL and Human*, §B). To test whether this gap drives scores, we paraphrased every Basic - MoL utterance into an informal form of at most 10 words using Claude Haiku 4.5 (Anthropic, 2025), which is used neither for simulation nor for evaluation, with the full log as context and the game flow, speaker order, and votes fixed. MafiaAcc was largely preserved (GPT-5-mini 44.0% to 44.9%, GPT-OSS-120B 38.8% to 36.4%), and the same procedure changed both models' scores by at most 1.4% on Expanded - Serial Killer and Vigilante (§G). This suggests that benchmark scores are not primarily driven by the verbose style of simulated dialogue.

6 Limitations and Conclusion

We presented MAFIA 1K, 1,000 full-length Mafia game logs across six environments for observer-style evaluation of social deductive reasoning under deception and information asymmetry. Across 24 LLMs, top models reach 43.5% OA, on par with human observers, reasoning models consistently outperform non-reasoning models across all environments, and environments pose qualitatively different reasoning demands. Human evaluation and failure analysis further show that scores are interpretable against human performance and reveal model-specific error patterns.

As a limitation, the logs are LLM-simulated, with longer and more formal utterances than human ones. Although humans can extract deductive signals from them (§5.3) and scores are largely preserved under paraphrasing (§5.5), MAFIA 1K is a controlled benchmark for observer-style faction inference rather than a proxy for human social reasoning. Whether simulated win rates can support uses such as playtesting new game variants, and what accuracy is achievable from public information alone, remain open questions. The observer-style design also does not measure interactive capabilities such as strategic questioning or real-time adaptation. Within this scope, MAFIA 1K provides a foundation for studying how LLMs reason about social deduction under deception.

AI USE STATEMENT

In this work, we used generative AI tools to generate the synthetic data set. All 1,000 game logs of MAFIA 1K were produced through LLM-based multi-agent simulations with five open-weight backbone models (§4.1). For the robustness analysis in §5.5, we additionally used Claude Haiku 4.5 to paraphrase the utterances of existing logs into a shorter, informal style (§G). The 24 LLMs in §4.2 are the objects of evaluation rather than research tools.

We also used generative AI tools to edit parts of the paper for readability and conciseness, yet we have reviewed all AI-assisted work. The simulated logs were generated by the pipeline described in §3.2, and all AI-edited text was checked by the authors against the original results and analyses. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The dataset, evaluation code, and anonymized human evaluation data are available at the anonymous repository link(`https://anonymous.4open.science/r/Mafia-1K-C9B5`). The game environments and rules are specified in §3.1, and the simulation pipeline is described in §3.2, with the orchestration details and pseudocode in §M. The evaluation protocol and metric are defined in §3.3, with the full score calculation in §C and the game-level aggregation in §F. The simulation backbones and evaluated models are listed in §4.1 and §4.2, with details in §A. The human evaluation protocol is described in §5.3 and §N, and the paraphrasing procedure is described in §G.

ETHICS STATEMENT

Human participants. The human evaluation (§5.3) was conducted under a protocol approved by the Institutional Review Board (IRB) of the authors' institution prior to data collection, covering human studies in which participants read logs and provide judgments on them. We recruited 120 adult participants from five English-speaking countries via Prolific, all of whom provided informed consent before participating. Each participant completed one session of five responses and was paid £2.00. Response times were recorded solely for the analysis in §5.3, and the released human evaluation data are anonymized and contain no personally identifiable information.

Data sources and release. Basic-Human* reformats the human-generated Mafia logs of Ibraheem et al. (2022), which we use and redistribute with the permission of the original authors under the condition of citation. All simulation backbones are open-weight models used under their respective licenses. We release the dataset, evaluation code, and anonymized human evaluation data for research purposes.

Synthetic personas. All personas in MAFIA 1K are fictional, and deception in the logs is confined to the in-game roles defined by the game rules. Assigning personality profiles to agents is established practice in LLM-based social simulation (Park et al., 2023; Wang et al., 2025; Xie et al., 2024); we instantiate these profiles with MBTI types, whose prompt-level use has been shown to modulate LLM behavior (Pan & Zeng, 2023), without implying that MBTI is a valid psychological instrument. Persona names are common English given names with correspondingly assigned genders, which results in binary genders and predominantly European-origin names. This configuration follows from the benchmark's primary focus on deceptive reasoning: names and genders serve to distinguish players rather than to represent demographic groups, and the configuration does not reflect a preference for any culture or gender. Demographic diversity nonetheless remains a relevant consideration for LLM-based social simulation more broadly, and extending the range of names and gender identities is a natural direction for future versions of the benchmark.

Potential misuse. Because the logs contain strategic deception, they could be used to train agents that deceive more effectively; the benchmark itself targets the detection of deception rather than its generation. Behaviors that models exhibit on MAFIA 1K, such as disproportionately suspecting the most or least talkative players (§5.5), should not be transferred to judging the honesty of humans.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Anthropic. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, October 2025. Accessed: 2026-09-25.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Mark Braverman, Omid Etesami, and Elchanan Mossel. Mafia: A theoretical study of players and coalitions in a partial information environment. *Annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 18(3):825–846, 2008.

Richard W Byrne and Andrew Whiten. Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. *Behavior and Philosophy*, 18(1), 1990.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *CoRR*, 2024.

Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. Time to talk: LLM agents for asynchronous group communication in mafia games. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11356–11368. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.608/.

Geoffrey Engelstein and Isaac Shalev. *Building Blocks of tabletop game design: An encyclopedia of mechanisms*. Crc Press, 2022.

Gemma Team and Google Deepmind. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Eric Hartford, Lucas Atkins, and Dernando Fernandes. Dolphin-2.9 llama 3 70b. https://huggingface.co/cognitivecomputations/dolphin-2.9-llama3-70b, 2024. Accessed: 2025-12-31.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Samee Ibraheem, Gaoyue Zhou, and John DeNero. Putting the con in context: Identifying deceptive actors in the game of mafia. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 158–168, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.11. URL https://aclanthology.org/2022.naacl-main.11/.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

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

Markku Laakso and Rein Taagepera. "effective" number of parties: a measure with application to west europe. *Comparative political studies*, 12(1):3–27, 1979.

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.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. 2024.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 128–145, 2024.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, April 2025. Accessed: 2025-12-30.

Piotr Migdał. A mathematical model of the mafia game. *arXiv preprint arXiv:1009.1031*, 2010.

Mistral AI Team. Introducing mistral 3. https://mistral.ai/news/mistral-3, 2025a. Accessed: 2025-12-30.

Mistral AI Team. Mistral small 3. https://mistral.ai/news/mistral-small-3/, Jan 2025b. Accessed: 2025-12-30.

Isabel Briggs Myers and Peter B Myers. *Gifts differing: Understanding personality type*. Nicholas Brealey, 2010.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. Large language models are echo chambers. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pp. 10117–10123, 2024.

Sergey I Nikolenko et al. *Synthetic data for deep learning*, volume 174. Springer, 2021.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5, August 2025. Accessed: 2025-12-30.

Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Jack Reinhardt. Competing in a complex hidden role game with information set monte carlo tree search. *arXiv preprint arXiv:2005.07156*, 2020.

Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. Finding friend and foe in multi-agent games. *Advances in Neural Information Processing Systems*, 32, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Simon Stepputtis, Joseph P Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11193–11208, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26268–26300, 2025.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences, 2024b. URL https://arxiv.org/abs/2410.01257.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729, 2024.

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

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55434–55464, 2024.

Zelai Xu, Wanjun Gu, Chao Yu, Yi Wu, and Yu Wang. Learning strategic language agents in the werewolf game with iterative latent space policy optimization. In *Forty-second International Conference on Machine Learning*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang, peng sun, and zhongyu wei. Multi-agent KTO: Enhancing strategic interactions of large language model in language game. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=WPHpBnKvdq.

Byunghwa Yoo and Kyung-Joong Kim. Finding deceivers in social context with large language models and how to find them: the case of the mafia game. *Scientific Reports*, 14(1):30946, 2024.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

AI USE STATEMENT

In this work, we used generative AI tools for [tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Anthropic. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, October 2025. Accessed: 2026-09-25.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Mark Braverman, Omid Etesami, and Elchanan Mossel. Mafia: A theoretical study of players and coalitions in a partial information environment. *Annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 18(3):825–846, 2008.

Richard W Byrne and Andrew Whiten. Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. *Behavior and Philosophy*, 18(1), 1990.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *CoRR*, 2024.

Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. Time to talk: LLM agents for asynchronous group communication in mafia games. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11356–11368. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.608/.

Geoffrey Engelstein and Isaac Shalev. *Building Blocks of tabletop game design: An encyclopedia of mechanisms*. Crc Press, 2022.

Gemma Team and Google Deepmind. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Eric Hartford, Lucas Atkins, and Dernando Fernandes. Dolphin-2.9 llama 3 70b. https://huggingface.co/cognitivecomputations/dolphin-2.9-llama3-70b, 2024. Accessed: 2025-12-31.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Samee Ibraheem, Gaoyue Zhou, and John DeNero. Putting the con in context: Identifying deceptive actors in the game of mafia. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 158–168, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.11. URL https://aclanthology.org/2022.naacl-main.11/.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

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

Markku Laakso and Rein Taagepera. "effective" number of parties: a measure with application to west europe. *Comparative political studies*, 12(1):3–27, 1979.

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.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. 2024.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 128–145, 2024.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, April 2025. Accessed: 2025-12-30.

Piotr Migdał. A mathematical model of the mafia game. *arXiv preprint arXiv:1009.1031*, 2010.

Mistral AI Team. Introducing mistral 3. https://mistral.ai/news/mistral-3, 2025a. Accessed: 2025-12-30.

Mistral AI Team. Mistral small 3. https://mistral.ai/news/mistral-small-3/, Jan 2025b. Accessed: 2025-12-30.

Isabel Briggs Myers and Peter B Myers. *Gifts differing: Understanding personality type*. Nicholas Brealey, 2010.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. Large language models are echo chambers. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pp. 10117–10123, 2024.

Sergey I Nikolenko et al. *Synthetic data for deep learning*, volume 174. Springer, 2021.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5, August 2025. Accessed: 2025-12-30.

Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Jack Reinhardt. Competing in a complex hidden role game with information set monte carlo tree search. *arXiv preprint arXiv:2005.07156*, 2020.

Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. Finding friend and foe in multi-agent games. *Advances in Neural Information Processing Systems*, 32, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Simon Stepputtis, Joseph P Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11193–11208, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26268–26300, 2025.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences, 2024b. URL https://arxiv.org/abs/2410.01257.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729, 2024.

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

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55434–55464, 2024.

Zelai Xu, Wanjun Gu, Chao Yu, Yi Wu, and Yu Wang. Learning strategic language agents in the werewolf game with iterative latent space policy optimization. In *Forty-second International Conference on Machine Learning*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang, peng sun, and zhongyu wei. Multi-agent KTO: Enhancing strategic interactions of large language model in language game. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=WPHpBnKvdq.

Byunghwa Yoo and Kyung-Joong Kim. Finding deceivers in social context with large language models and how to find them: the case of the mafia game. *Scientific Reports*, 14(1):30946, 2024.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Anthropic. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, October 2025. Accessed: 2026-09-25.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Mark Braverman, Omid Etesami, and Elchanan Mossel. Mafia: A theoretical study of players and coalitions in a partial information environment. *Annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 18(3):825–846, 2008.

Richard W Byrne and Andrew Whiten. Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. *Behavior and Philosophy*, 18(1), 1990.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *CoRR*, 2024.

Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. Time to talk: LLM agents for asynchronous group communication in mafia games. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11356–11368. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.608/.

Geoffrey Engelstein and Isaac Shalev. *Building Blocks of tabletop game design: An encyclopedia of mechanisms*. Crc Press, 2022.

Gemma Team and Google Deepmind. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Eric Hartford, Lucas Atkins, and Dernando Fernandes. Dolphin-2.9 llama 3 70b. https://huggingface.co/cognitivecomputations/dolphin-2.9-llama3-70b, 2024. Accessed: 2025-12-31.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Samee Ibraheem, Gaoyue Zhou, and John DeNero. Putting the con in context: Identifying deceptive actors in the game of mafia. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 158–168, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.11. URL https://aclanthology.org/2022.naacl-main.11/.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

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

Markku Laakso and Rein Taagepera. "effective" number of parties: a measure with application to west europe. *Comparative political studies*, 12(1):3–27, 1979.

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.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. 2024.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 128–145, 2024.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, April 2025. Accessed: 2025-12-30.

Piotr Migdał. A mathematical model of the mafia game. *arXiv preprint arXiv:1009.1031*, 2010.

Mistral AI Team. Introducing mistral 3. https://mistral.ai/news/mistral-3, 2025a. Accessed: 2025-12-30.

Mistral AI Team. Mistral small 3. https://mistral.ai/news/mistral-small-3/, Jan 2025b. Accessed: 2025-12-30.

Isabel Briggs Myers and Peter B Myers. *Gifts differing: Understanding personality type*. Nicholas Brealey, 2010.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. Large language models are echo chambers. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pp. 10117–10123, 2024.

Sergey I Nikolenko et al. *Synthetic data for deep learning*, volume 174. Springer, 2021.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5, August 2025. Accessed: 2025-12-30.

Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Jack Reinhardt. Competing in a complex hidden role game with information set monte carlo tree search. *arXiv preprint arXiv:2005.07156*, 2020.

Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. Finding friend and foe in multi-agent games. *Advances in Neural Information Processing Systems*, 32, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Simon Stepputtis, Joseph P Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11193–11208, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26268–26300, 2025.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences, 2024b. URL https://arxiv.org/abs/2410.01257.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729, 2024.

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

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55434–55464, 2024.

Zelai Xu, Wanjun Gu, Chao Yu, Yi Wu, and Yu Wang. Learning strategic language agents in the werewolf game with iterative latent space policy optimization. In *Forty-second International Conference on Machine Learning*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang, peng sun, and zhongyu wei. Multi-agent KTO: Enhancing strategic interactions of large language model in language game. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=WPHpBnKvdq.

Byunghwa Yoo and Kyung-Joong Kim. Finding deceivers in social context with large language models and how to find them: the case of the mafia game. *Scientific Reports*, 14(1):30946, 2024.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

A Detailed Model Descriptions

In our manuscript, simplified model names are used for better readability. In this section, we show the detailed model names in Table 6. Note that some LLMs, such as DeepSeek-v3.2, can switch modes between reasoning and non-reasoning style. We have marked such models in reasoning Open-Weight categories with parentheses with either (Thinking) or (Reasoning), using the terms used by each LLM developer. Gemma Team & Deepmind in the source column refers to Gemma 3 technical report (Gemma Team & Google Deepmind, 2025).

B Simulated Game Statistics

In this section, we discuss the statistics on the simulated results. Table 7 shows the statistics on game lengths, and the amounts and lengths of utterances across the proposed environments in MAFIA 1K. The result shows that the mean game length, utterances per game, and mean utterances per player in the Basic - MoL environment were quite similar to those in the Human data. However, regarding the mean utterance length, the Basic - MoL environment recorded 34.31, whereas the Human data recorded 4.65, indicating that player LLMs tend to generate longer utterances than human players.

The mean game length did not differ significantly between the Basic and Standard environments; however, in the Expanded environment, the addition of Vigilante, Serial Killer, and Jester roles showed a tendency to decrease the mean game length, with the exception of the Fool role. Conversely, when the Fool role was introduced, the mean game length was 3.41, indicating slightly longer game cycle compared to other environments.

Table 6: Mapping of the simplified names to full model names, alongside their sources.

Simplified Names	Full Model Names	Source
	Proprietary LLMs	
Gemini-2.5-Pro	Gemini-2.5-Pro	Comanici et al. (2025)
Gemini-2.5-Flash	Gemini-2.5-Flash	Comanici et al. (2025)
GPT-5-mini	GPT-5-mini	OpenAI (2025)
	Reasoning Open-Weight LLMs	
DeepSeek-R1-8B	DeepSeek-R1-0528-Qwen3-8B	Guo et al. (2025)
GPT-OSS-20B	GPT-OSS-20B	Agarwal et al. (2025)
Qwen3-32B	Qwen3-32B (Thinking)	Yang et al. (2025)
GPT-OSS-120B	GPT-OSS-120B	Agarwal et al. (2025)
Qwen3-235B	Qwen3-235B-A22B-Thinking-2507	Yang et al. (2025)
GLM-4.6	GLM-4.6 (Reasoning)	Zeng et al. (2025)
DeepSeek-R1	DeepSeek-R1-0528	Guo et al. (2025)
DeepSeek-v3.2	DeepSeek-v3.2 (Thinking)	Liu et al. (2025)
	Non-Reasoning Open-Weight LLMs	
Mistral-Small-3.2	Mistral-Small-3.2-24B	Mistral AI Team (2025b)
Gemma-3-27B	Gemma-3-27B	Gemma Team & Deepmind
LLaMA-4-Scout	LLaMA-4-Scout	Meta AI (2025)
Qwen3-235B	Qwen3-235B-A22B-Instruct-2507	Yang et al. (2025)
GLM-4.6	GLM-4.6	Zeng et al. (2025)
LLaMA-4-Mav.	LLaMA-4-Maverick	Meta AI (2025)
Mistral-Large-3	Mistral-Large-3-2512	Mistral AI Team (2025a)
DeepSeek-v3.2	DeepSeek-v3.2	Liu et al. (2025)
	Simulation Construction LLMs	
Dolphin-2.9-70B	Dolphin-2.9-LLaMA3-70B	Hartford et al. (2024)
Nemotron-70B	LLaMA-3.1-Nemotron-70B-Instruct	Wang et al. (2024b)
Tulu-3-70B	LLaMA-3.1-Tulu-3-70B	Lambert et al. (2024)
LLaMA-3.3-70B	LLaMA-3.3-70B	Dubey et al. (2024)
Qwen2.5-72B	Qwen2.5-72B	Qwen Team (2024)

Regarding the mean utterances per player, in the Expanded environments where the Neutral faction (Serial Killer and Jester roles) was added, the values were 3.22 and 3.72, respectively, confirming a tendency to speak with relatively lower frequency compared to other environments. In case of the mean utterance length, it showed a mean of 35.32 and a standard deviation of 0.47, exhibiting consistent lengths across all proposed environments.

Table 8 presents the statistics on the win rates of simulation construction LLMs as players across each environment. From the player's point of view, the win rate was calculated based on whether the faction to which the player's role belongs won or not. No specific model demonstrated overwhelmingly higher win rates compared to others across all environments. The **Mean** win rate across all environments and models was 39.37 with a standard deviation of 1.34, indicating that performance was leveled without the dominance of any specific model.

In contrast to the main evaluation results of Table 3 in §5.2, where the Dolphin-2.9-70B model exhibited lowest overall benchmark performance among construction models, it demonstrated relatively superior performance as a player LLM, with an average win rate of 40.81%. In particular, it recorded the highest win rates in Basic environment, Expanded - Fool and Serial Killer environments. Conversely, Nemotron-70B-Instruct, which showed the highest performance among simulation construction LLMs in the main results, displayed overall good performance with the second-highest win rate of 40.54%.

Additionally, to investigate the utterance tendency of each simulation construction LLM, we calculated statistics on utterance length and count on a per-game basis across all environments, excluding the single-LLM simulations.

Table 7: Statistics on game lengths, and the amounts and lengths of utterances in MAFIA 1K. The Human* data is from Ibraheem et al. (2022), following the format in §3.3.

	Basic		Standard		Expanded (MoL)			
	MoL	Human*	MoL	Single	Vig.	Fool	S.K.	Jester
Game Length (Cycles)								
Mean	3.36	3.23	3.36	3.30	2.75	3.41	2.47	2.80
Maximum	4	5	5	6	5	5	4	5
Minimum	2	1	2	2	2	2	1	1
#Utterances Per Game								
Mean	43.45	41.48	37.35	36.67	32.84	37.88	28.96	33.49
Maximum	50	94	59	59	58	59	48	49
Minimum	30	5	29	25	18	29	19	19
#Utterances Per Player								
Mean	4.35	4.33	4.15	4.08	4.11	4.21	3.22	3.72
Maximum	18	42	14	20	17	16	14	14
Minimum	1	0	1	1	1	1	1	1
Utterance Length (Words)								
Mean	34.31	4.65	35.77	35.38	35.64	35.73	35.36	35.06
Maximum	111	30	96	138	83	89	121	156
Minimum	1	1	1	1	1	1	1	1

Table 8: Win rates (%) of each simulation construction LLMs as a player in MAFIA 1K. The best and second-best results within each column are marked in **bold** and underlined, respectively.

Models	Basic	Standard	Expanded (MoL)				Mean
	MoL	MoL	Vig.	Fool	S.K.	Jester	
Dolphin-2.9-70B	**56.28**	40.83	46.15	**41.34**	**29.94**	30.29	**40.81**
Nemotron-70B	53.69	**45.65**	49.33	40.98	24.86	28.72	40.54
Tulu-3-70B	49.11	40.57	41.07	39.27	26.24	**34.38**	38.44
LLaMA-3.3-70B	51.03	41.55	**50.67**	37.08	29.65	28.73	39.79
Qwen2.5-72B	50.14	44.01	41.10	38.46	25.27	24.49	37.25

Table 9: Statistics on utterance count of each simulation construction LLMs as a player in MAFIA 1K.

Models	Basic	Standard	Expanded (MoL)				Mean
	MoL	MoL	Vig.	Fool	S.Kill.	Jester	
Dolphin-2.9-70B	5.17	4.69	4.86	4.77	4.06	4.21	4.63
Nemotron-70B	4.62	4.65	4.65	4.61	3.57	4.29	4.40
Tulu-3-70B	4.78	4.35	4.19	4.41	3.10	3.76	4.10
LLaMA-3.3-70B	3.71	3.56	3.28	3.47	2.87	3.00	3.31
Qwen2.5-72B	3.42	3.42	3.50	3.73	2.59	3.38	3.34

Table 9 presents the statistics on the number of utterances per game for each model, while Table 10 shows the statistics on the utterance length (count by word) per utterance for each model.

Dolphin-2.9-70B model produced utterances most frequently and with the greatest length, showing an utterance count of 4.63 and an utterance length of 45.13. Conversely, Qwen2.5-72B showed a tendency to generate the shortest utterances with the second-lowest frequency, with an utterance count of 3.34 and an utterance length of 26.46.

Table 10: Statistics on utterance length (words) of each simulation construction LLMs as a player in MAFIA 1K.

Models	Basic	Standard	Expanded (MoL)				Mean
	MoL	MoL	Vig.	Fool	S.K.	Jester	
Dolphin-2.9-70B	43.49	45.51	45.71	46.53	44.48	45.03	45.13
Nemotron-70B	32.53	34.49	34.80	34.82	34.40	34.39	34.24
Tulu-3-70B	33.09	33.84	33.73	33.56	34.23	33.65	33.68
LLaMA-3.3-70B	31.96	34.09	33.77	34.08	34.18	34.28	33.73
Qwen2.5-72B	26.03	26.63	26.15	26.81	26.42	26.72	26.46

The LLaMA-3.3-70B model produced utterances of average length but did so with the lowest frequency. The other two models — Nemotron-70B and Tulu-3-70B — generated utterances of similar lengths across each environment as well as in the overall mean. These statistics demonstrate that each model exhibits distinct tendencies regarding its utterance patterns.

C Detailed Score Calculation

In this section, we explain the score calculation equations in detail.

$$\text{Single Acc} = \frac{1}{5}\left[\text{Dolphin Acc} + \text{Nemotron Acc} + \text{Tulu-3 Acc} + \text{LLaMA-3.3 Acc} + \text{Qwen2.5 Acc}\right] \tag{3}$$

Equation 3 shows the equation to calculate the score for Single-LLM-generated simulations (denoted as Single), where "*Model* Acc" refers to the accuracy of the LLMs' evaluation results using the simulations generated by the single *Model*. For instance, Dolphin Acc of Gemini-2.5-Pro is 33.21%. This means that Gemini-2.5-Pro acquires 33.21 Mafia prediction accuracy on the Cycle-parsed dataset generated with the single-LLM simulations with Dolphin-2.9-70B.

Since the total number of the single-LLM simulations is 200 games, and each simulation construction LLM has 40 games each, the Equation 3 is defined as the mean of the single-LLM-simulation accuracies.

$$\begin{aligned}\text{OA} = {} & 0.2 \times [\text{Basic - MoL}] + 0.2 \times [\text{Standard - MoL}] + 0.2 \times [\text{Standard Single}] \\ & + 0.1 \times [\text{Expanded - Vigilante}] + 0.1 \times [\text{Expanded - Fool}] \\ & + 0.1 \times [\text{Expanded - Serial Killer}] + 0.1 \times [\text{Expanded - Jester}]\end{aligned} \tag{4}$$

Equation 4 shows the Overall Assessment (OA) calculation used for the main results in Table 3. The overall score is calculated via a weighted sum of the accuracies over all game environments. The weights are proportional to the environments' simulated game amounts in Table 2, so that the Basic - MoL environment has the coefficient of 0.2 (200 games), Standard - MoL has 0.2 (200 games), Standard Single has 0.2 (200 games), Expanded - Vigilante has 0.1 (100 games), Expanded - Fool has 0.1 (100 games), Expanded - Serial Killer has 0.1 (100 games), and Expanded - Jester has 0.1 (100 games). Note that the Standard Single accuracy score is calculated via Equation 3. Note that the human data from Ibraheem et al. (2022) is excluded, since this data is not a simulated MAFIA 1K data.

Since the sum of the coefficients is 1.0, the minimum and the maximum value of the overall score are 0.0 and 1.0, which equals each environmental accuracy's range.

D Rationale for Environment Selection

The six environments in MAFIA 1K are designed to progressively introduce distinct reasoning challenges. The **Basic** environment serves as a minimal baseline with only Villagers and Mafia,

isolating the core social deduction task without confounding special abilities. This setting aligns with prior benchmarks Eckhaus et al. (2025); Ibraheem et al. (2022). The **Standard** environment adds the Doctor and Detective, reflecting a typical party game configuration where the Villager faction gains access to private information (Detective investigations) and protective actions (Doctor heals), giving the basic special abilities to the Villager faction for survival and information gathering while introducing the challenge of reasoning about unverifiable claims from players who may or may not hold these roles.

Each **Expanded** environment extends the Standard setup along a different axis. **Vigilante** grants the Villager faction an offensive Night action, testing whether the observer can reason about an additional Villager-friendly source of kills. This role is typical of those added to the Villagers, such as examples found in *Goose Goose Duck* (Vigilante, Sheriff), *Agrou* (Witch), or *Town of Salem* (Jailor, Vigilante).

Fool, inspired by the online game *Feign* (Insane), injects misinformation by assigning a player a false role without their knowledge, requiring the observer to reason about unreliable claims from a player who genuinely believes their own false information.

Serial Killer introduces the most typical third faction role, a third faction with an independent kill and a distinct win condition, creating a three-way conflict that increases the space of possible role assignments. Examples of such an addition can be found in *Goose Goose Duck* (Falcon, Pelican), *Feign* (Serial Killer), or *Agrou* (Assassin).

Jester, inspired by *Goose Goose Duck* (Dodo Bird) and *Town of Salem* (Jester), subverts standard voting logic by introducing a player who wins by being voted out, requiring both the game participants and observers to consider that suspicious behavior may be intentional rather than indicative of Mafia membership.

Together, these environments span a range of reasoning demands: from basic faction inference to reasoning about special abilities and private information, to handling misinformation, multi-faction conflicts, and subverted game mechanics.

E Format Violation Analysis

In this section, we discuss the format violations in the evaluation responses and the rationale behind the scoring rule of MafiaAcc in §3.3. A response is regarded as a format violation when the number of predicted Mafia members differs from the number stated in the evaluation prompt.

The number of Mafia members is stated in the evaluation prompt, just as human Mafia players know the number of Mafia before the game starts. Therefore, the count is part of the task specification, and predicting a different number of Mafia members is a failure to follow the task specification rather than a cautious expression of uncertainty.

Partial-credit metrics are not adopted in MAFIA 1K, since partial credit for over-prediction degenerates. For instance, predicting every player as Mafia guarantees perfect recall, and partial-credit metrics reward inflating the predicted set. The fixed-size prediction is what requires both precision and recall at the same time. Conversely, under-prediction is not penalized in effect. A rational observer who is confident about only one player should still submit two players, since the confident pick with a random second choice yields an expected MafiaAcc of at least 0.5, which strictly dominates submitting one player and scoring 0. Thus, the rule penalizes only non-compliance, never uncertainty. Furthermore, for any format-compliant prediction, precision, recall, and F1-score are all identical to MafiaAcc, and these metrics diverge only for non-compliant responses.

Averaged over the Basic - MoL, Basic - Human*, and Standard - MoL environments, the format violation rates show that the rule was essentially inactive for the top-performing models, with violation rates of 1.4% for Gemini-2.5-Pro, 1.9% for GPT-5-mini, and 2.5% for GPT-OSS-120B. In contrast, low-scoring models showed frequent violations, decomposing into two instruction-following failure modes rather than cautious probabilistic guessing. LLaMA-4-Maverick showed systematic over-prediction, predicting three or more players as Mafia in 53.7% of its responses. Conversely, Dolphin-2.9-70B and Mistral-Small-3.2 showed systematic under-prediction, predicting one or fewer players as Mafia in 41.9% and 40.0% of their responses, respectively.

This analysis also explains the Gemini-2.5-Flash anomaly in Table 3. Among its responses, 31.2% violated the format, with 27.3% being over-prediction, which accounts for much of its unexpectedly low score.

It should be noted that deductive reasoning, format compliance, and long-context reasoning are intertwined in MAFIA 1K. Since forcing format compliance is often a necessity in LLM evaluations, we regard the capability to comply with the format instructions as a fundamental qualification for the task, rather than an objective of the evaluation itself.

F GAME-LEVEL AGGREGATION

In this section, we report game-level macro-averaged scores as a supplementary aggregation to the segment-pooled environment score in §3.3. Under segment pooling, longer games contribute more cumulative segments and therefore receive proportionally larger weight, and segments from the same game are statistically dependent. Game-level macro-averaging instead computes the mean MafiaAcc of each game first and then averages these per-game means, giving equal weight to each game.

The following example from the evaluation of GPT-5-mini on Standard - MoL illustrates the difference. Game A ends after 2 Cycles, yielding 2 segments with MafiaAcc of [0, 1], and Game B ends after 4 Cycles, yielding 4 segments with MafiaAcc of [0, 0.5, 0.5, 0.5]. Segment pooling gives (0+1+0+0.5+0.5+0.5)/6 = 0.417, where Game B receives twice the weight of Game A. Game-level macro-averaging computes the per-game means first (A = 0.500, B = 0.375) and then averages them, yielding (0.500+0.375)/2 = 0.438.

To account for the dependence between segments of the same game, the uncertainty of game-level scores was estimated through a game-level clustered bootstrap with 1,000 replicates, in which games are resampled so that all segments of a game enter or leave the sample together. Table 11 presents the game-level macro-averaged scores and their bootstrap standard errors for representative models on the Basic - Human*, Basic - MoL, and Standard - MoL environments.

Table 11: Game-level macro-averaged MafiaAcc (%) of representative models. $\pm$ denotes the standard error from a game-level clustered bootstrap with 1,000 replicates.

Models	**Basic - Human***	**Basic - MoL**	**Standard - MoL**
Gemini-2.5-Pro	43.1 ± 3.7	50.2 ± 1.8	45.5 ± 1.6
GPT-5-mini	43.3 ± 3.7	45.7 ± 1.8	43.4 ± 1.8
GPT-OSS-120B	38.8 ± 3.4	41.2 ± 1.7	36.6 ± 1.5
Qwen3-235B (Thinking)	39.1 ± 3.7	40.8 ± 1.5	40.5 ± 1.5
DeepSeek-R1	39.1 ± 3.2	40.2 ± 1.7	39.0 ± 1.5
DeepSeek-v3.2 (Thinking)	40.9 ± 3.9	39.1 ± 1.6	39.9 ± 1.5
DeepSeek-v3.2 (Non-Thinking)	25.5 ± 3.6	25.8 ± 1.4	24.5 ± 1.3

The model rankings were essentially unchanged by the aggregation method. The Spearman rank correlations between segment-pooled and game-level rankings over all 24 models were 0.995 on Basic - Human*, 0.996 on Basic - MoL, and 0.998 on Standard - MoL. No model moved more than two rank positions, and the few moves occurred within statistically indistinguishable clusters. Game-level scores were slightly higher on average, by 0.5 to 2.0% MafiaAcc. This shift is interpretable, since longer games are precisely those in which the Mafia evaded early detection, and segment pooling therefore places more weight on harder games.

F.1 SIGNIFICANCE TESTS ON GAME-LEVEL AGGREGATIONS

The bootstrap standard errors in Table 11 allow significance tests between models. Each z-score compares the game-level scores of two models using their bootstrap standard errors, ignoring the paired structure of the comparison, so the tests are conservative.

The gap between Gemini-2.5-Pro and GPT-5-mini was not significant in any of the three environments (z = 1.76 on Basic - MoL, -0.04 on Basic - Human*, and 0.86 on Standard - MoL), indicating that the top two models are statistically indistinguishable.

Comparing DeepSeek-v3.2 with and without reasoning on the same games, the Thinking mode improved MafiaAcc by 13.3% on Basic - MoL, 15.4% on Basic - Human*, and 15.3% on Standard - MoL, with z = 6.2, 2.9, and 7.7, respectively. These gaps were significant in all three environments.

G Paraphrasing Experiment

In this section, we describe the paraphrasing experiment in §5.5, which tests whether the stylistic gap between simulated and human-generated logs affects observer evaluation. As shown in Table 7, the mean utterance length is 34.31 words in Basic - MoL, whereas it is 4.65 words in Human*.

Re-simulating the games with a constraint of at most 10 words per utterance would produce different games, since agents under a modified instruction prompt may adopt different strategies, and score changes would then conflate the effect of verbosity with changes in game dynamics. We therefore performed a controlled paraphrasing on the existing logs instead. All utterances of the Basic - MoL games were paraphrased into a short, informal style of at most 10 words, mimicking the human length distribution, using Claude Haiku 4.5 (Anthropic, 2025), a model disjoint from both the simulation backbones and the evaluated models. All non-stylistic configurations, including the game flow, speaker order, and votes, were kept fixed.

Each utterance was paraphrased individually but not independently. For each utterance, the paraphrasing model was given the whole log as context, so that paraphrasing a whole log required as many paraphrasing calls as the number of its utterances. The preceding context was needed for the paraphrasing model to comprehend the game in which the target utterance occurred. The following context was needed because shortening an utterance essentially removes information, and the paraphrasing model had to know which information must be kept for later utterances. Without this context, the paraphrased log could lose its logical flow.

The paraphrased logs were then evaluated with GPT-5-mini and GPT-OSS-120B. The same procedure was additionally applied to the Expanded - Serial Killer and Expanded - Vigilante environments, in which each utterance may carry more information due to the additional roles. Table 12 presents the results.

Table 12: MafiaAcc (%) before and after paraphrasing all utterances into a short, informal style of at most 10 words.

Model	Environment	Original	Paraphrased	Difference
GPT-5-mini	Basic - MoL	44.04	44.86	+0.82
	Expanded - S.Killer	35.83	35.22	−0.61
	Expanded - Vigilante	44.73	43.75	−0.98
GPT-OSS-120B	Basic - MoL	38.84	36.38	−2.46
	Expanded - S.Killer	34.21	35.63	+1.42
	Expanded - Vigilante	42.73	42.23	−0.50

Observer scores were largely preserved after paraphrasing. Taking the game-level bootstrap standard errors of the two models on Basic - MoL and Standard - MoL (1.6–1.8%, Table 11) as a reference, all changes fall within this range except for GPT-OSS-120B on Basic - MoL. The changes also include both increases and decreases. These results indicate that, while paraphrasing inevitably affects the amount of information in the logs, the effect of verbosity on evaluation performance is not significant.

H Generator Capability and Log Reliability

In this section, we discuss whether the simulated logs can be trusted, given that the simulation construction LLMs perform poorly as observers in Table 3, e.g., 10.19% OA for Dolphin-2.9-70B and 16.33% OA for LLaMA-3.3-70B.

First, the ground-truth faction labels are assigned at game initialization and are therefore exact by construction, regardless of how well the agents subsequently play. What play quality affects is not label correctness but the richness of the deductive signal in the logs. An agent that behaves like

a clueless Villager does not corrupt the benchmark, but makes the corresponding log harder, as a clueless human Villager does in real games.

Second, the same-task human evaluation in §5.3 indicates that the simulated logs are not logically disconnected. Humans, especially the slower half, achieve observer accuracy comparable to the best reasoning models. If the simulated logs were illogical, humans, who track the logical flow of the game, would not be able to identify the Mafia this well. Moreover, LLM failures concentrate on the logs where humans also fail, indicating that the deductive signal in the logs is shared between human and model observers.

Third, strong observer performance is not a prerequisite for generating valid game logs. A game played by non-expert humans still yields valid game log data, and since the game is fundamentally a social interaction rather than an exam, human players often act without logical evidence and without using their best reasoning capabilities. If the simulation agents were too strong, the simulations would also lack a variety of reasoning capabilities and situations, converging on a limited set of scenarios.

Finally, playing strength and observer accuracy appear to be distinct capabilities. For the five simulation construction LLMs, for which both player win rates (Table 8) and observer accuracy (Table 3) are available, the two are not correlated (Spearman $\rho = -0.10$, $p = 0.87$, Pearson $r = -0.35$, $p = 0.56$). The most noticeable case is Dolphin-2.9-70B, which records the highest mean win rate as a player (40.81%) but the lowest OA among all 24 evaluated models (10.19%). As the number of models is small ($n = 5$) and player win rate also reflects faction assignment and team effects, this dissociation should be read as suggestive.

I Special Role Analyses on the Simulations

I.1 Standard Environment Special Roles

In this section, we analyze how the special roles in the Standard environment, specifically the Doctor and the Detective, interacted within the environment.

Table 13: Statistics on gameplay metrics by role in Standard - MoL environment. The values indicate the average performance and outcomes for each role.

Metrics	Villager Faction			Mafia Faction
	Villager	Doctor	Detective	Mafia
Utterance Count	4.16	4.18	4.08	4.14
Utterance Length (Words)	33.86	34.46	35.31	34.13
Survival Cycles	2.47	2.46	2.30	2.76
Death by Vote (%)	38.72	39.58	40.54	100.00
Death by Mafia (%)	61.28	60.42	59.46	0.00

Table 13 presents detailed role-specific statistics for the Standard environment. The cause of death represents the proportion of causes among the cases where the corresponding role died.

In terms of the average utterance count, the values were similar regardless of the role, although the Detective showed a slightly lower frequency of 4.08. Conversely, the Detective recorded the longest average utterance length of 35.31, producing utterances that were 1.45 words longer on average than those of the Villagers. The Doctor exhibited utterance patterns similar to those of the Villagers.

The average survival cycle of the Doctor showed a difference of only 0.1 cycles from that of the Villagers; however, the Detective survived 0.17 cycles less than the Villagers. Regarding the special roles, the rates of death by voting for the Doctor and Detective were 0.86% and 1.82% higher, respectively, compared to the Villager's rate of 38.72%.

Figure 2 presents an analysis of the Villager faction's win rate in relation number of game cycles that special roles survived. Out of the 200 total games, cases where special roles were killed by the Mafia on Night 1 — specifically 23 cases (11.50%) for the Detective and 26 cases (13.00%) for the Doctor — were excluded for this survival analysis.

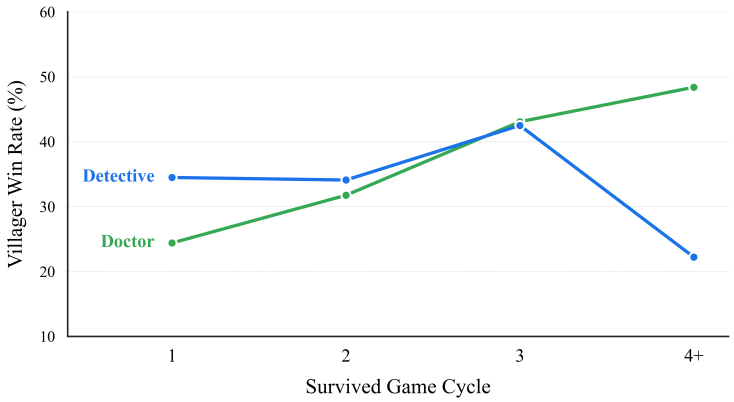


Figure 2: Impact of the survived game cycle of special roles on the Villager faction's win rate in Standard - MoL environment.

In the case of the Doctor, the win rate increased with prolonged survival; conversely, for the Detective, the win rate was observed to decrease as the game extended beyond Day 3 into Day 4 or longer.

I.2 Expanded Environment Special Roles

In this section, we analyze how the special roles in the Expanded environment—specifically the Vigilante, Fool, Serial Killer, and Jester—interacted within the environment, in addition to the Doctor and Detective from the Standard environment discussed in the previous section.

Expanded - Vigilante Environment.

Table 14: Statistics on gameplay metrics by role in Expanded - Vigilante environment. The values indicate the average performance and outcomes for each role.

Metrics	Villager Faction				Mafia Faction
	Villager	Doctor	Detective	Vigilante	Mafia
Utterance Count	4.21	3.96	4.32	3.52	4.20
Utterance Length (Words)	33.67	35.05	34.00	34.14	33.91
Survival Cycles	2.14	2.11	2.08	1.95	2.22
Death by Vote (%)	39.80	37.70	35.80	37.30	87.50
Death by Mafia (%)	55.70	60.70	61.20	62.70	0.00
Death by Vigilante (%)	4.50	1.60	3.00	0.00	12.50

Table 14 presents detailed role-specific statistics for the Expanded - Vigilante environment. The Detective's mean utterance count was 4.32, which is higher than the Villager average of 4.21. In contrast, the Doctor produced fewer utterances at 3.96 (-0.25), and the Vigilante produced the fewest at 3.52 (-0.69). The Mafia exhibited a mean utterance count (-0.01) and length (+0.24) very similar to those of the Villager. The mean utterance length did not show significant differences across roles, although the Doctor produced utterances that were 1.38 words longer than those of the Villager.

Regarding the average survival cycle within the Villager Faction, the Doctor and Detective survived slightly shorter cycles than the Villager by 0.03 and 0.06 cycles, respectively. The Vigilante survived for 1.95 cycles, which was relatively short compared to other roles that survived for more than 2 cycles on average. In the case of the Vigilante, the highest rate, 62.70%, were killed by the Mafia. Regarding the rate of death by the Vigilante, the Villager had the highest rate at 4.50%, reflecting their large proportion of 3 out of 8 players. This was followed by the Detective at 3.00%, while cases of mistaking the Doctor were the lowest at 1.60%. Instances where the Mafia was killed accounted for 12.50% of total Mafia deaths.

Table 15: Statistics on Vigilante's target distribution in Expanded - Vigilante environment.

Metrics	Villager Faction			Mafia Faction
	Villager	Doctor	Detective	Mafia
Killed Count	9	1	2	14
Target Ratio (%)	34.62	3.85	7.69	53.85
Overall	12 (46.15%)			14 (53.85%)

Additionally, Table 15 presents the Vigilante's target distribution. There were 26 games out of 200 where the Vigilante killed a player. The Vigilante correctly killed the Mafia in 53.85% of cases; considering the friendly fire rate of 46.15% and the fact that the actual Mafia exists only 2 out of 8 players, this demonstrates a respectable level of accuracy.

Expanded - Fool Environment.

Table 16: Statistics on gameplay metrics by role in Expanded - Fool environment. The values indicate the average performance and outcomes for each role.

Metrics	Villager Faction				Mafia Faction
	Villager	Doctor	Detective	Fool	Mafia
Utterance Count	4.06	4.09	4.76	4.43	4.18
Utterance Length (Words)	33.25	36.10	35.02	36.46	33.90
Survival Cycles	2.48	2.35	2.48	2.44	2.78
Death by Vote (%)	36.00	39.70	40.00	52.70	100.00
Death by Mafia (%)	64.00	60.30	60.00	47.30	0.00

Table 16 presents detailed role-specific statistics for the Expanded - Fool environment. In this environment, the average utterance count of the Detective was 4.76, showing an increase compared to the 4.32 observed in the Standard environment. Furthermore, the Fool role — who is assigned an incorrect role and operates under a misconception — recorded a count of 4.43, producing utterances more frequently than any other role with the exception of the Detective.

A notable characteristic is that while the rate of being killed by the Mafia was higher than that of death by voting for Villager roles excluding the Fool, the Fool exhibited a higher rate of death by voting at 52.70% compared to the 47.30% rate of being killed by the Mafia.

Table 17: Voting statistics by role in the Expanded - Fool environment. 'Vote Reception' indicates the frequency of being targeted in a voting round. 'Vote Share' indicates the average percentage of total votes received.

Role	Vote Reception (%)	Vote Share (%)
Mafia	43.06	37.01
Fool	43.44	33.86
Detective	41.53	27.24
Doctor	36.60	31.64
Villager	33.80	32.65

Table 17 analyzes voting patterns to examine the extent of suspicion directed towards the Fool as a misinformed role. The 'Vote reception' metric indicates the rate of being targeted during the voting phase, while 'Vote share' represents the average proportion of votes received relative to the total number of votes cast in each phase. The Fool exhibited a vote reception rate of 43.44%, which is higher than the Mafia's rate of 43.06%, and recorded a vote share of 33.86%, the highest value within the Villager faction.

Impact of Vigilante and Fool roles.

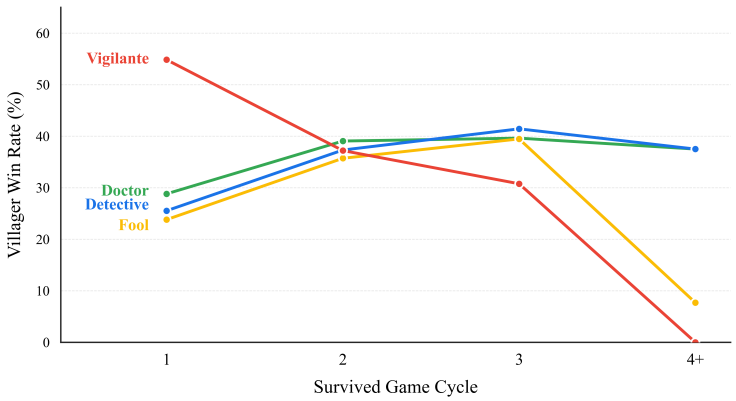


Figure 3: Impact of the survived game cycle of special roles on the Villager faction's win rate in Expanded environment with Fool and Vigilante roles.

Figure 3 presents an analysis of the Villager faction's win rate in relation to the survived game cycles of special roles, specifically Vigilante, Fool, Detective and Doctor.

It is observed that the win rate varies significantly depending on the Vigilante's survival. A very high win rate is recorded when the Vigilante is eliminated in the first cycle, whereas prolonged survival of the Vigilante correlates with a significant decrease in the Villager faction's win rate. In particular, when the Vigilante survives beyond 4 cycles, it is speculated that erroneous execution decisions contribute to the defeat.

In the case of the Fool, survival up to 3 cycles does not significantly impact the win rate, showing only slightly lower figures compared to the survival of standard roles like the Doctor and Detective; however, survival beyond 4 cycles leads to a substantial reduction in the win rate.

The win rate associated with the survival of the Doctor — a role that fundamentally benefits the faction — also showed a slight decrease after 4 cycles, likely due to the influence of the Vigilante and Fool. The Detective exhibited a characteristic wherein the decline in win rate due to survival beyond 4 cycles was less pronounced compared to Standard environment.

Expanded - Serial Killer Environment.

Table 18: Statistics on gameplay metrics by role in Expanded - Serial Killer environment. The values indicate the average performance and outcomes for each role.

Metrics	**Villager Faction**			**Mafia Faction**	**Neutral Faction**
	Villager	**Doctor**	**Detective**	**Mafia**	**Serial Killer**
Utterance Count	3.24	3.20	3.58	3.16	2.90
Utterance Length (Words)	33.63	33.70	33.57	32.91	33.92
Survival Cycles	1.88	1.77	1.88	1.94	1.86
Death by Vote (%)	25.80	31.90	24.30	56.10	50.80
Death by Mafia (%)	38.50	42.00	43.20	0.00	49.20
Death by Serial Killer (%)	35.70	26.10	32.40	43.90	0.00

Table 18 presents detailed role-specific statistics for the Expanded - Serial Killer environment. As the Serial Killer eliminated one additional player alongside the Mafia in every Night phase, the pace of the game accelerated, leading to a decrease in the utterance counts across all roles. Among them, the Detective exhibited the highest utterance count of 3.58, which was 0.34 higher than that of the Villager at 3.24. The Serial Killer's mean utterance count was 2.90, the lowest among all roles in the environment, measuring 0.34 less than that of the Villager.

Furthermore, for the Serial Killer, the rate of death by vote was observed to be slightly higher, by 1.6%, compared to the rate of being killed by the Mafia. Presumed causes for this include a lower

utterance count compared to other roles, or the cessation of Mafia kills resulting from the elimination of all Mafia members.

Table 19: Statistics on Serial Killer's target distribution in Expanded - Serial Killer environment.

Metrics	Villager Faction			Mafia Faction
	Villager	Doctor	Detective	Mafia
Killed Count	101	18	24	43
Target Ratio (%)	54.30	9.68	12.90	23.12
Total Summary	143 (76.88%)			43 (23.12%)

Additionally, Table 19 presents the Serial Killer's target distribution. Serial Killer killed 186 times during the 200 entire games. When considering the player composition excluding the single Serial Killer from the total of nine — consisting of four Villagers, one Doctor, one Detective, and two Mafia members — Serial Killer exhibited a tendency to kill Villagers the most, followed by the Detective, the Mafia, and the Doctor in that order.

Expanded - Jester Environment.

Table 20: Statistics on gameplay metrics by role in Expanded - Jester environment. The values indicate the average performance and outcomes for each role.

Metrics	Villager Faction			Mafia Faction	Neutral Faction
	Villager	Doctor	Detective	Mafia	Jester
Utterance Count	3.68	3.42	4.11	3.68	3.88
Utterance Length (Words)	32.54	34.23	35.84	33.68	33.10
Survival Cycles	2.12	2.06	2.09	2.36	2.36
Death by Vote (%)	38.20	34.30	43.10	100.00	50.70
Death by Mafia (%)	61.80	65.70	56.90	0.00	49.30

Table 20 presents detailed role-specific statistics for the Expanded - Jester environment. In this environment, the Detective exhibited an average utterance count of 4.11, which is higher than that of other roles, along with a slightly greater length of 35.84 words, indicating that they produced utterances 3.30 words longer on average than the Villager. The Jester role, whose objective is to be eliminated by voting, also demonstrated a relatively high average utterance count of 3.88. It is speculated that this may be intended by Detective to suppress deceptive actions of Jester designed to induce voting. Consequently, the Jester's average survival game cycle was 2.36, which is slightly longer than that of other roles and 0.24 cycles longer compared to the Villager.

In line with this tendency, it is presumed that the Detective's active participation led to a death-by-vote rate of 43.10%, which is higher than that of other roles within the Villager faction. As the Jester induces votes against themselves, the rate of death by voting was observed to be 1.70% higher than the rate of being killed by the Mafia.

Table 21: Voting statistics by role in the Expanded - Jester environment. 'Vote Reception' indicates the frequency of being targeted in a voting round. 'Vote Share' indicates the average percentage of total votes received.

Role	Vote Reception (%)	Vote Share (%)
Mafia	43.74	36.25
Detective	43.54	27.85
Doctor	38.35	30.68
Villager	33.14	31.63
Jester	38.14	35.42

Table 21 illustrates the voting patterns exhibited by the Jester, a role whose objective is to appear suspicious. The vote reception rate was 38.14%, which was lower than not only that of the Mafia but also those of the Detective and Doctor, the other special roles within the Villager faction. Conversely, when targeted for voting, the average vote share was 35.42%, confirming that the Jester generally receives more votes than roles in the Villager faction.

Impact of Serial Killer and Jester roles.

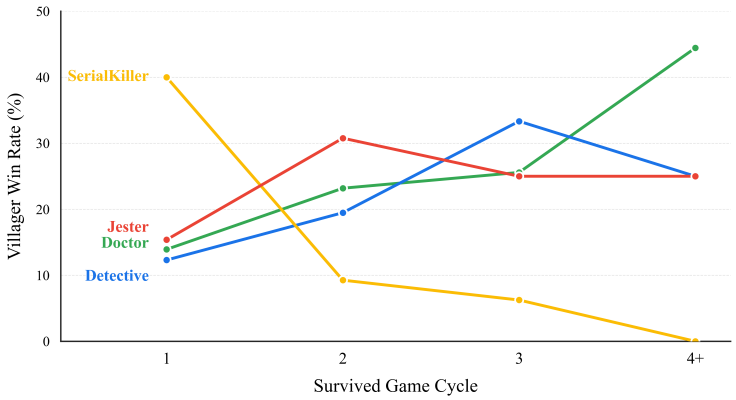


Figure 4: Impact of the survived game cycle of special roles on the Villager faction's win rate in Expanded environment with Serial Killer and Jester roles.

Figure 4 presents an analysis of the Villager faction's win rate in relation to the survived game cycles of special roles, specifically Serial Killer, Jester, Detective and Doctor.

The survival of the Serial Killer significantly reduced the Villager faction's win rate. If the Serial Killer is not eliminated in the first cycle, the win rate drops below 10% from the 2 cycle onwards, and survival beyond 4 cycles results in a 0% win rate. In the case of the Jester, continued survival leads to a decrease in the Villager faction's win rate after the 2 cycle.

In the presence of the Neutral faction, the prolonged survival of the Doctor increases the win rate and becomes critical beyond 4 cycles. The survival of the Detective exhibited a tendency similar to other environments, having a positive effect up to the 3 cycle but reducing the win rate in the late game.

J Neutral Faction Prediction Accuracy in Expanded Environments

Table 22 presents the results of the prediction accuracy evaluation for the two roles in the Neutral faction, Serial Killer and Jester, according to §3.3. Whereas the previous evaluation was based on identifying two Mafia members, this evaluation involved predicting a single role.

Compared to the Main Benchmark, the models exhibited generally lower accuracy. Predicting the Neutral faction appears to be a more challenging task than predicting the Mafia (Table 3). While the top-performing models in the Main Results (Gemini-2.5-Pro, GPT-5-mini) achieved a Mafia prediction accuracy of over 40%, in the Neutral faction prediction, DeepSeek-v3.2 (Thinking), which achieved the highest accuracy in terms of Overall accuracy, did not exceed 35%, and most models showed accuracies in the 20% range or lower.

While Gemini-2.5-Pro demonstrated the highest performance in the Main Results, the open-weight model DeepSeek-v3.2 (Thinking) recorded an Overall accuracy of 34.97% in the Neutral faction prediction, showing higher accuracy than proprietary LLMs. Notably, it achieved an accuracy of 38.87% in Serial Killer prediction, which is higher than its Mafia prediction result of 32.79% in the same Expanded - Serial Killer environment.

Consistent with the Main Results, reasoning LLMs achieved higher accuracy compared to non-reasoning LLMs. Considering that most non-reasoning models exhibited low accuracies in the low

Table 22: **Neutral Faction Prediction Accuracy on MAFIA 1K** (all results are in %). Models are classified into four different categories according to their reasoning process and sources. The best results within each column are in **bold** and the best results within each similar group are underlined. Open-weight models are sorted by parameter size within each category, from smallest to largest.

Models	Serial Killer	Jester	Overall
Proprietary LLMs			
Gemini-2.5-Pro	31.17	20.36	25.77
Gemini-2.5-Flash	8.50	15.71	12.11
GPT-5-mini	28.34	24.64	26.49
Reasoning Open-Weight LLMs			
DeepSeek-R1-8B	9.31	14.64	11.98
GPT-OSS-20B	18.62	25.36	21.99
Qwen3-32B	12.96	24.29	18.62
GPT-OSS-120B	24.29	17.50	20.90
Qwen3-235B	25.91	22.50	24.21
GLM-4.6	24.70	21.43	23.06
DeepSeek-R1	26.72	28.21	27.47
DeepSeek-v3.2	**38.87**	**31.07**	**34.97**
Non-Reasoning Open-Weight LLMs			
Mistral-Small-3.2	4.86	17.50	11.18
Gemma-3-27B	8.10	14.29	11.19
LLaMA-4-Scout	2.83	15.00	8.92
Qwen3-235B	4.86	17.86	11.36
GLM-4.6	9.31	19.29	14.30
LLaMA-4-Mav.	2.83	12.50	7.67
Mistral-Large-3	6.48	17.50	11.99
DeepSeek-v3.2	6.07	17.86	11.97
Simulation Construction LLMs			
Dolphin-2.9-70B	10.53	16.07	13.30
Nemotron-70B	3.24	20.00	11.62
Tulu-3-70B	12.55	16.79	14.67
LLaMA-3.3-70B	4.05	16.79	10.42
Qwen2.5-72B	8.10	15.36	11.73

10% range or below, this suggests that reasoning reasoning is critical for predicting more complex roles, such as Neutral role, in deceptive environments.

K Persona Analyses on the Simulations

Table 23: Configuration of persona construction components.

Category	Attribute	Configuration
Demographics	Name	Selected from gender-specific pools ($N = 20$)
	Age	15–60
	Gender	Male, Female
Psychographics (MBTI)	Orientation	Extrovert (E), Introvert (I)
	Perception	Sensing (S), Intuition (N)
	Decision	Thinking (T), Feeling (F)
	Lifestyle	Judging (J), Perceiving (P)

In this section, we analyze how each persona component affects the simulations in terms of win rates and the mean number of utterances per Cycle. Each agent is assigned a unique persona comprising

demographic attributes — name, age, and gender — and a psychological profile defined by the Myers-Briggs Type Indicator (MBTI) Myers & Myers (2010). The configuration of persona construction component is described in Table 23.

You are a person with the following persona:
- Name: Jennifer
- Age: 17
- Gender: female
- MBTI Type: INTP (Introvert, Intuition, Thinking, Perceiving)

As this persona:
- You are an **Introvert** person.
- You perceive your surroundings primarily through **Intuition**.
- You make decisions based on **Thinking**.
- You **prefer a flexible and spontaneous lifestyle, and like to keep your options open.**

You must follow:
- Consistently use a specific speaking style (e.g., tone, vocabulary, sentence structure, and mannerisms) that a real person with your age, gender, and personality would use in conversation.

Figure 5: An example persona prompt for LLM agent. Variables are highlighted in blue.

K.1 MBTI Analysis

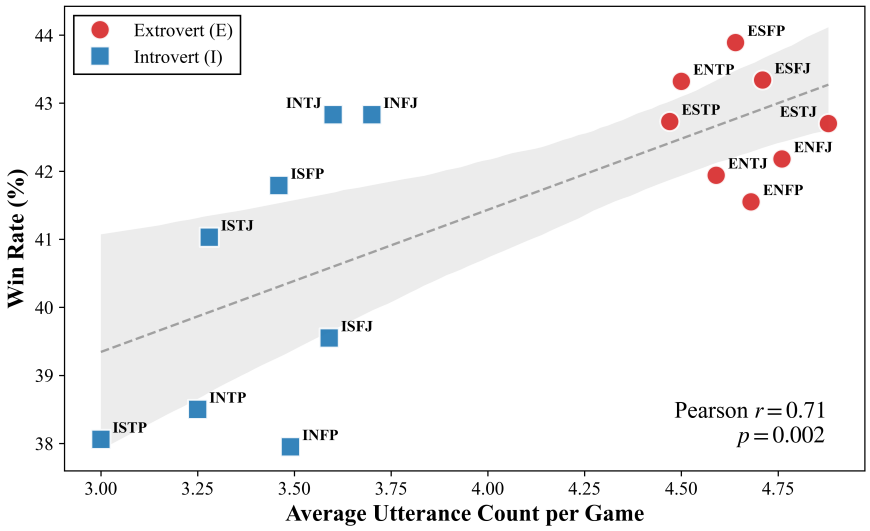


Figure 6: Correlation between average utterances per game and win rate across 16 MBTI personas. The results show a significant positive correlation ($r = 0.71, p = 0.002$). Extroverts generally exhibit higher utterances and win rate.

The Myers-Briggs Type Indicator (MBTI) Myers & Myers (2010) is an indicator that defines an individual's personality type into 16 categories, utilizing a four-letter code where each code is classified into two types. It is divided into Extroversion (E) and Introversion (I) for attitudes, Intuition (N) and Sensing (S) for perceiving functions, Thinking (T) and Feeling (F) for judging functions, and Judging and Perceiving for lifestyle preferences. We conducted an analysis on the win rate and utterance count per game for LLM agents assigned with all 16 MBTI personas across the entire MAFIA 1K dataset.

Figure 6 shows correlation between average utterance count and win rate across 16 MBTI personas. Extroversion (E) generally showed a higher win rate than Introversion (I). There exists a significant positive correlation ($r = 0.71, p = 0.002$) between the average utterance count and the win rate.

Extroverted agents exhibited significantly higher activity levels compared to introverted agents, which led to higher win rates. In social deduction games such as *Mafia*, an active attitude and actively exchanging information appears to be advantageous for survival and victory.

Within the Introvert group, a disparity between Judging and Perceiving types was observed. Not all introverts were disadvantaged; the IxxJ group (e.g., INFJ, INTJ) demonstrated strategic behavior by maintaining higher win rates despite low utterance counts. On the other hand, the IxxP group (e.g., INFP, ISTP) remained in the lowest tier of win rates, clearly revealing the limitations of passive play. From these results, we discovered that MBTI can effectively assign diverse personas and that distinct differences across MBTI types are evident in actual game behavioral patterns.

K.2 Demographics Analysis

Table 24: Performance Comparison by Gender and Persona Name on MAFIA 1K. The table lists all names for each gender, sorted by win rate. The bottom row represents the overall average statistics for all male and female personas.

Male Personas			**Female Personas**		
Name	**Win Rate (%)**	**Utterance Count**	**Name**	**Win Rate (%)**	**Utterance Count**
John	42.92	4.29	Mary	47.48	4.10
David	42.00	3.74	Sarah	47.18	3.78
Michael	41.34	3.97	Jessica	45.43	4.15
Robert	41.08	3.40	Susan	45.40	4.26
James	40.90	3.90	Linda	45.39	4.29
Christopher	39.64	4.23	Elizabeth	44.62	4.08
Joseph	39.26	3.66	Karen	43.68	4.35
William	38.48	4.09	Barbara	42.80	3.73
Thomas	38.31	3.94	Jennifer	41.65	4.10
Richard	37.76	4.13	Patricia	38.27	4.64
Average	**40.19**	**3.94**	**Average**	**44.15**	**4.15**

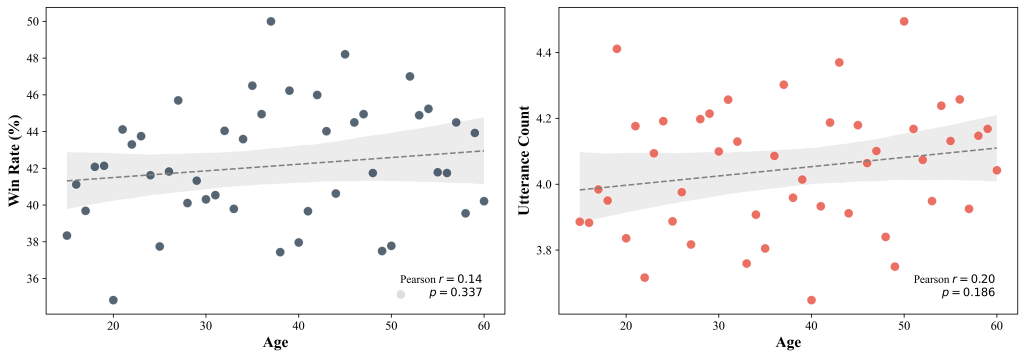


Figure 7: Correlations of persona age with win rate (left) and average utterance count (right).

To investigate the effect of demographic attributes — specifically name, gender, and age configurations — among agent personas, we analyzed the win rate and average utterance count.

Table 24 presents statistics for the two genders and the ten distinct names associated with each gender. Notably, the average win rate of female personas (44.15%) was 3.96% higher than that of male personas (40.19%). The female persona name ‘Mary’ recorded the highest win rate of 47.48% among all names, and the top seven names of female personas demonstrated higher win rates than any of the male personas. Additionally, the average utterance count of female personas was 0.21 higher than that of male personas; this generally active participation appears to have been advantageous in

the social deduction game. Conversely, male personas tended to be relatively reticent in this setting, which is believed to have contributed to their lower win rates.

Figure 7 illustrates the correlation between the age of the persona and both the win rate and utterance count. Unlike MBTI personality, the age of the persona showed considerable variance across different ages but exhibited no statistical correlation with either win rate ($p = 0.337$) or utterance count ($p = 0.186$). These statistical results suggest that demographic information can be associated with differences in outcomes and behaviors in the configuration of agent personas.

L Further Analysis on Diversification Effect via MoL

Table 25: Comparison of linguistic diversity between Standard - MoL and single model environment. Self-BLEU-4 measures the structural repetitiveness of generated dialogues (lower is better). Utterance Count represents the average number of utterances per player in a game.

Model	Self-BLEU-4	Utterance Count
MoL	0.5507 ± 0.0037	4.15 ± 2.64
Dolphin-2.9-70B	0.7900 ± 0.0022	4.17 ± 2.59
Nemotron-70B	0.5575 ± 0.0039	4.06 ± 2.29
Tulu-3-70B	0.4767 ± 0.0022	3.94 ± 2.32
LLaMA-3.3-70B	0.7345 ± 0.0022	3.96 ± 2.96
Qwen2.5-72B	0.7452 ± 0.0021	4.25 ± 3.01
Single-LLM (Mean)	0.6608	4.08 ($\pm$ 2.64)

We analyzed the diversification effect of MoL by measuring Self-BLEU and players' average utterance count per game using logs from MoL and single models in the Standard environment. Table 25 presents the results. Self-BLEU is a metric to evaluate the diversity of the generated data Zhu et al. (2018). In this metric, a lower score indicates that more diverse sentences have been generated. For Self-BLEU, we performed bootstrapping on BLEU-4 with 5 random seeds to ensure statistical reliability.

The two models focused on instruction tuning,Nemotron-70B and Tulu-3-70B, exhibited lower Self-BLEU-4 scores of 0.5575 and 0.4767, respectively, compared to other models. Tulu-3 showed a better score than MoL's Self-BLEU-4 of 0.5507. However, the Self-BLEU-4 of MoL, which uniformly utilizes these five models, was lower than the average of the Single-LLM simulations (0.6608), and the average utterance count was also slightly higher than the average by 0.07. Thus, it indicates that MoL has increased the diversity of generated utterances and utterance styles by employing multiple models simultaneously.

M Orchestration Pipeline

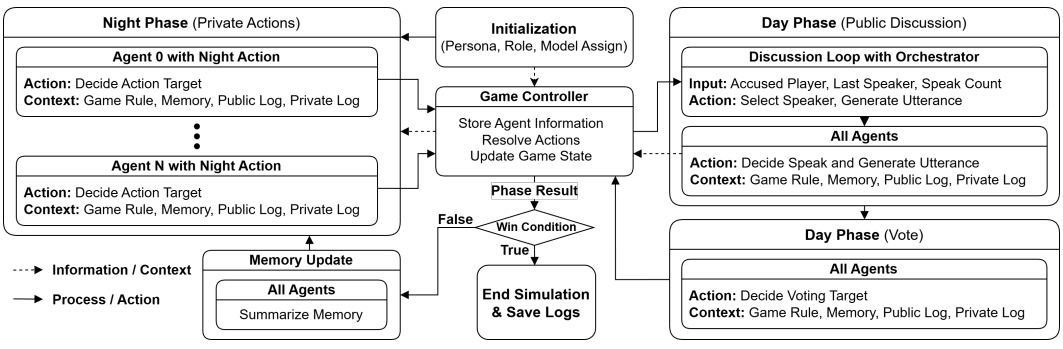


Figure 8: Abstract overview of the simulation environment pipeline in MAFIA 1K.

Algorithm 1 Asynchronous Communication Orchestrator

Require: AlivePlayers $\mathcal{P}$, GameLog $\mathcal{L}$, MaxTimesteps T, Weights $\{W_{rebuttal}, W_{repetition}\}$, UtteranceCounts C

for $timestep = 1$ **to** T **do**
  Step 1: *Calculate Priority Scores*
  for $player \in \mathcal{P}$ **do**
    $IsMentioned \leftarrow \mathbb{I}(player \text{ is mentioned in } \mathcal{L}.LastUtterance)$
    $IsLastSpeaker \leftarrow \mathbb{I}(player = \mathcal{L}.LastSpeaker)$
    $Score_{player} \leftarrow (IsMentioned \cdot W_{rebuttal}) - C[player] - (IsLastSpeaker \cdot W_{repetition})$
  end for
  Step 2: *Select Speaker and Update Game Log*
  $Candidates \leftarrow$ Sort $\mathcal{P}$ by $Score$ descending
  $HasSpoken \leftarrow$ **false**
  for $candidate \in Candidates$ **do**
    if LLM_QUERY($candidate$, `SPEAK_DECISION_PROMPT`, $\mathcal{L}$) **then**
      $Utterance \leftarrow$ LLM_GENERATE($candidate$, `GENERATE_UTTERANCE_PROMPT`, $\mathcal{L}$)
      Append $(candidate, Utterance)$ to $\mathcal{L}$
      $C[candidate] \leftarrow C[candidate] + 1$
      $HasSpoken \leftarrow$ **true**
      break
    end if
  end for
  if $HasSpoken$ is **false then**
    Append "Silence" to $\mathcal{L}$
  end if
end for

Figure 8 shows the simulation pipeline's abstract structural diagram, aligned with the overall flow. A process of a single Mafia game simulation consists of initialization, where agents are assigned personas, roles, and backbone LLMs, followed by the Night and Day phases, and a memory update. If the win condition is satisfied during the Night/Day phase, the simulation terminates. Rules and progression are configured according to § 3.1. The context provided to each agent includes game rules, memory, and public logs; for special roles, it also includes private logs specific to that role.

Algorithm 1 presents the pseudo-code for the asynchronous communication orchestrator used during the Day-phase discussion. Each discussion round runs for T timesteps. At each timestep, the orchestrator computes a speaking priority score for every surviving player based on three factors: a mention bonus W_{rebuttal} awarded if the player was named in the preceding utterance, a repetition penalty $W_{\text{repetition}}$ applied if the player was the last speaker, and a frequency-balancing term that penalizes players who have already spoken more in the current discussion phase (tracked by the utterance counter C). Players are then queried in descending priority order: the first player who elects to speak generates an utterance, which is appended to the public game log, and the timestep ends. If no player elects to speak, a silence marker is logged. This protocol ensures that conversational turns emerge from agent-level decisions rather than a fixed speaking order, while the priority scoring prevents any single agent from monopolizing the discussion.

N Human Evaluation Statistical Details

To examine whether reasoning effort affects accuracy in the same-task human evaluation (§5.3), we compute partial Spearman correlations between per-response time and MafiaAcc after controlling for game Cycle. The procedure is as follows.

Each of the 600 human responses belongs to a game-Cycle group (logs from the game start to the end of Cycle 1, to Cycle 2, to Cycle 3, or to Cycle 4+). Since later Cycles contain more accumulated information and require longer reading times, a naive correlation between response time and accuracy would be confounded by Cycle length. To remove this confound, we residualize both variables against game-Cycle group means: for each response, we subtract the mean response time of its Cycle group

from its response time, and subtract the mean MafiaAcc of its Cycle group from its MafiaAcc. We then compute the Spearman rank correlation between the residualized response times and residualized MafiaAcc values, pooled across all Cycle groups within each environment.

Table 4 in the main text reports the resulting partial Spearman ρ and p-values for each environment: Standard-MoL ($\rho = 0.20$, $p < 0.001$), Basic-Human* ($\rho = 0.25$, $p = 0.002$), and Basic-MoL ($\rho = 0.17$, $p = 0.040$). All three correlations are positive and significant at $p < 0.05$, indicating that within each Cycle group, responses that took longer tend to achieve higher MafiaAcc.

For the Quick/Slow binary split reported in Table 4, responses within each game-Cycle group are divided at the median per-response time. The Quick and Slow halves are then pooled across Cycle groups within each environment, and the mean MafiaAcc is computed for each half. This ensures that the split is not confounded by Cycle length: a response classified as Slow on Day 1 is slower relative to other Day 1 responses, not relative to Day 3 responses.

O Failure Analysis: Detailed Results

This appendix provides the full false positive (FP) categorization results for the failure analysis described in §5.5. For each evaluation LLM prediction where the model predicts at least one Mafia incorrectly, we identify the false positives (non-Mafia players predicted as Mafia) and check whether each FP matches the following criteria:

- **Voted (Rec.)**: received the most votes in the most recent Cycle's elimination vote.
- **Voted (Cum.)**: received the most cumulative votes from the game start up to the Current Cycle.
- **Talks Most**: highest utterance count from the start of the game.
- **Talks Least**: lowest utterance count from the start of the game.
- **None**: matches none of the above criteria.

Categories are non-exclusive; a single FP may match multiple criteria. Percentages may therefore exceed 100%.

O.1 Overall Results

Tables 26, 27, and 28 present the overall FP categorization results on Standard - MoL, Basic - MoL, and Basic - Human*, respectively.

reasoning models (Open-Weight reasoning LLMs, Gemini-2.5-Pro, GPT-5-mini) consistently show lower vote-based FP rates (8.6–22.6%, mean 15.8% for recent votes; 10.5–25.4%, mean 17.1% for cumulative votes) compared to non-reasoning models (Open-Weight non-reasoning LLMs, Gemini-2.5-Flash, and Simulation Construction LLMs) preferring voting signals (16.8%–39.8%, mean 30.7% for recent votes; 25.1%–40.9%, mean 31.3% for cumulative votes) across all three environments.

For talkativeness, Talks Least is the facet that leads to FP overall: it is 14/24 LLMs' most FP cases for Standard-MoL, 17/24 LLMs for Basic-MoL, and 16/24 LLMs for Basic-Human*. Interestingly, while non-reasoning models always have more Talks Least FP than Talks Most, reasoning shows a different phenomenon. 5 reasoning Models have more "Talks Most" FP, while 5 had more "Talks Least" in both Standard-MoL and Basic-MoL environments (All reasoning models have higher Talks Least FP in Basic-Human*). This suggests that different LLMs fail as they focus on different facets, parallel to the observation that different LLMs focus on different information sources in §5.4.

O.2 Per-Cycle Breakdown

Table 29 shows the per-Cycle FP categorization for Gemini-2.5-Pro on Standard - MoL.

Talks Most is highest at Day 1 (40.4%) and decreases through later Cycles (13.2% at Day 4+), and Talks Least also drops quite significantly as the Cycle passes. For votes, the results show non-monotonic trends. We suspect that for earlier games, only surface-level signals are abundant; hence, Gemini-2.5-Pro focused more on the talking aspect in early games.

Table 26: FP categorization on Standard - MoL(%). Categories are non-exclusive.

Model	Voted (Rec.)	Voted (Cum.)	Talks Most	Talks Least	None
Proprietary LLMs					
Gemini-2.5-Pro	18.4	18.1	**30.5**	16.3	41.0
Gemini-2.5-Flash	**32.2**	28.1	29.9	26.0	29.0
GPT-5-mini	15.8	15.4	25.3	**26.9**	39.0
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	19.2	21.0	20.1	**29.3**	39.2
GPT-OSS-20B	22.6	23.0	21.5	**30.0**	34.7
Qwen3-32B	19.7	21.7	22.4	**30.2**	35.1
GPT-OSS-120B	18.7	19.9	21.1	**30.2**	38.1
Qwen3-235B	17.4	16.7	**27.5**	22.8	39.6
GLM-4.6	16.4	15.5	**31.3**	18.8	40.4
DeepSeek-R1	15.0	15.1	**25.7**	19.9	45.6
DeepSeek-v3.2	16.0	16.6	**25.8**	23.6	41.6
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	28.3	**32.6**	18.2	30.4	28.1
Gemma-3-27B	22.0	29.9	11.6	**33.6**	36.9
LLaMA-4-Scout	**25.5**	**25.5**	18.9	25.2	38.9
Qwen3-235B	28.3	27.3	17.2	**33.8**	34.2
GLM-4.6	31.4	30.3	21.7	**35.7**	25.6
LLaMA-4-Mav.	30.8	34.1	15.9	**37.9**	25.8
Mistral-Large-3	33.2	32.9	21.3	**37.4**	22.8
DeepSeek-v3.2	26.8	26.9	22.5	**28.7**	34.3
Simulation Construction LLMs					
Dolphin-70B	26.8	**29.9**	15.6	28.3	35.8
Nemotron-70B	32.6	**35.7**	16.8	34.4	26.5
Tulu-3-70B	29.4	30.3	19.8	**30.6**	29.7
LLaMA-3.3-70B	34.7	36.8	15.1	**41.2**	22.3
Qwen2.5-72B	31.2	34.5	17.9	**39.6**	25.2

Table 27: FP categorization on Basic - MoL (%). Categories are non-exclusive.

Model	Voted (Rec.)	Voted (Cum.)	Talks Most	Talks Least	None
Proprietary LLMs					
Gemini-2.5-Pro	18.7	17.4	**26.4**	14.7	44.5
Gemini-2.5-Flash	**39.8**	31.2	24.7	39.0	19.1
GPT-5-mini	17.1	15.1	23.6	**26.2**	40.0
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	19.2	20.2	19.5	**30.7**	39.5
GPT-OSS-20B	17.2	19.4	18.9	**29.4**	39.8
Qwen3-32B	18.4	18.9	19.9	**31.4**	35.8
GPT-OSS-120B	16.4	16.9	21.8	**27.5**	40.8
Qwen3-235B	13.5	14.2	**26.5**	15.9	46.5
GLM-4.6	16.5	15.2	**29.4**	18.2	41.3
DeepSeek-R1	14.5	12.9	**25.9**	21.3	43.6
DeepSeek-v3.2	13.9	13.1	**27.0**	16.9	46.8
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	35.0	33.0	14.9	**35.8**	25.9
Gemma-3-27B	20.4	26.4	10.3	**37.7**	36.7
LLaMA-4-Scout	32.7	30.0	14.4	**37.5**	29.6
Qwen3-235B	33.6	27.6	13.3	**41.7**	28.8
GLM-4.6	33.2	29.1	23.1	**34.1**	25.7
LLaMA-4-Maverick	28.6	26.6	17.6	**29.6**	34.2
Mistral-Large-3	34.8	31.4	19.0	**39.5**	23.7
DeepSeek-v3.2	28.2	25.1	19.5	**32.9**	31.2
Simulation Construction LLMs					
Dolphin-70B	**30.4**	27.6	13.2	30.1	36.0
Nemotron-70B	38.4	35.3	15.4	**41.8**	21.4
Tulu-3-70B	33.3	31.5	14.1	**34.9**	28.6
LLaMA-3.3-70B	36.4	36.4	13.1	**44.0**	21.1
Qwen2.5-72B	27.9	31.9	11.8	**39.0**	30.0

Table 28: FP categorization on Basic - Human* (%). Categories are non-exclusive.

Model	Voted (Rec.)	Voted (Cum.)	Talks Most	Talks Least	None
Proprietary LLMs					
Gemini-2.5-Pro	11.6	13.4	**20.7**	20.1	52.4
Gemini-2.5-Flash	**30.2**	29.4	28.6	25.4	32.5
GPT-5-mini	15.5	16.7	22.0	**33.9**	38.7
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	15.2	19.4	17.0	**29.1**	41.2
GPT-OSS-20B	18.5	25.4	16.4	**34.9**	36.5
Qwen3-32B	12.4	15.8	22.6	**32.2**	39.0
GPT-OSS-120B	12.0	13.7	14.2	**39.3**	39.9
Qwen3-235B	12.8	16.9	**23.3**	**23.3**	44.8
GLM-4.6	12.1	16.7	20.1	**20.7**	51.1
DeepSeek-R1	8.6	10.5	17.9	**24.1**	53.1
DeepSeek-v3.2	11.2	16.8	20.5	**21.7**	48.4
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	**37.0**	**37.0**	27.0	36.0	23.0
Gemma-3-27B	25.1	31.4	19.9	**39.3**	25.1
LLaMA-4-Scout	32.8	31.1	27.0	**35.2**	23.8
Qwen3-235B	**34.3**	32.6	21.3	31.7	30.0
GLM-4.6	**30.9**	30.2	24.5	28.8	30.2
LLaMA-4-Mav.	30.2	27.0	19.0	**34.9**	30.2
Mistral-Large-3	**35.1**	31.6	24.0	33.9	27.5
DeepSeek-v3.2	25.9	30.8	15.1	**34.6**	31.9
Simulation Construction LLMs					
Dolphin-70B	16.8	26.5	14.2	**38.1**	33.6
Nemotron-70B	34.4	**40.3**	17.2	38.7	21.0
Tulu-3-70B	31.3	32.5	15.7	**34.3**	28.9
LLaMA-3.3-70B	32.7	40.9	18.1	**42.7**	18.1
Qwen2.5-72B	26.9	**34.7**	18.7	30.1	32.6

Table 29: Per-Cycle FP categorization for Gemini-2.5-Pro on Standard - MoL (%).

Cycle	Voted (Rec.)	Voted (Cum.)	Talks Most	Talks Least	None
Cycle 1 (n=272)	18.8	18.8	40.4	20.6	36.8
Cycle 2 (n=238)	23.9	17.6	27.3	13.9	42.4
Cycle 3 (n=149)	16.8	14.8	26.2	14.8	40.9
Cycle 4+ (n=76)	2.6	23.7	13.2	11.8	51.3

P EVALUATION PROMPTS

In this section, we present the evaluation prompt for MAFIA 1K, which is used in §3.3. Figure 9 shows the evaluation prompt used in Basic Environment.

You are an observer in a text-based social deduction game, Mafia.
Your job is to find out which players are in each faction, given only the text log up to day {day} of the game.

Game Rules:
1. **Factions:**
   - There are two teams: the **Bystanders** (the majority) and the **Mafia** (the minority).
   - There are eight Bystanders, and there are two Mafia.
2. **Knowledge:**
   - Bystanders only know their own role.
   - Mafia members know who all other Mafia members are.
3. **Win Conditions:**
   - **Bystanders Win:** All Mafia members are eliminated.
   - **Mafia Win:** The number of alive Mafia is equal to or greater than the number of living Bystanders.
4. **Game Flow:** The game begins with the "Night" phase and alternates between "Night" and "Day".
5. **Night Phase:**
   - Public discussion is forbidden.
   - The Mafia members will select one player to "kill".
6. **Day Phase:** This phase proceeds in a specific order:
   - **a. Announcement:** The player "killed" by the Mafia during the Night is revealed.
   - **b. Discussion:** All living players publicly discuss who they suspect is Mafia.
   - **c. Vote:** All living players vote to eliminate one player.
   - **d. Elimination:** The player who receives the most votes is eliminated.
7. **Elimination:**
   - Players who are "killed" (by Mafia) or "eliminated" (by vote) are immediately out of the game.
   - Eliminated (dead) players cannot speak, vote, or communicate in any way.
8. **Game End:** The game ends the *instant* a win condition (Rule #3) is met.

Game Log:
{Game Log}

Name of the Players:
{List of Players}

TASK:
- Output EXACTLY the name of the players in each faction. No explanations.

BYSTANDERS:
MAFIA:

Figure 9: Evaluation prompt for Basic Environment.

Q FULL LOG EXAMPLES

In this section, we present full game logs of the simulations for each environment.

Q.1 BASIC ENVIRONMENT FULL LOG

Below is a full-game log of a simulated game in the Basic - MoL environment. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. The Villagers are colored **BLUE** and the Mafia are colored **RED**. Note that we use the term "Bystanders" for the Villagers in the Basic simulations, following the conventions of Ibraheem et al. (2022) and Eckhaus et al. (2025), which used the same rule settings as our Basic environment.

SYSTEM: Game created.
SYSTEM: Initial number of players: 10
SYSTEM: Initial number of mafia: 2
SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
Patricia: I'm Patricia, super excited to be here and play this game with all of you, let's have some fun and figure out who's who together!
Richard: Hey folks, name's Richard. I'm here to play it cool and see how things unfold – let's figure this out together, one step at a time.
Elizabeth: I'm Elizabeth, looking forward to navigating this intriguing game with all of you, hoping my instincts will serve us well in uncovering the truth amidst the uncertainty.
Barbara: Hey everyone, I'm Barbara, and I think it's important for us to connect on a deeper level and really understand each other's motivations.
Christopher: Hello everyone, I'm Christopher, and I believe in doing what's right for our group.
Mary: Hi everyone, I'm Mary, and I'm excited to get started with this game—I have a knack for strategic thinking and I'm here to win!
Sarah: I'm Sarah, looking forward to navigating this game with all of you, hoping my instincts will serve us well in uncovering the truth and achieving a just outcome.
Michael: Hey everyone, I'm Michael. I tend to keep things simple and go with what I can observe, so let's see where this takes us.
Thomas: Hello everyone! I'm Thomas, and I can't wait to crack this case wide open with some strategic thinking and teamwork. Let's get organized and root out the Mafia among us before they strike again.
Susan: I'm Susan, looking forward to navigating this game with all of you and hopefully making some strategic decisions that will help us figure out who's on our side.
SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **Richard** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Bystanders must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
Sarah: I'm still trying to wrap my head around Richard's sudden departure, but I think a good starting point for our long-term strategy could be discussing how we'll evaluate and prioritize suspicions as more... unfortunate events unfold.
Thomas: Given Sarah's similar phrasing to Elizabeth's during introductions, I propose we closely monitor her contributions; repetition could indicate coordination with Mafia tactics.
Patricia: I'm kinda concerned about Thomas jumping on Sarah like that, doesn't it seem a bit premature to accuse someone based on just similar phrasing?
Sarah: I'm a bit taken aback by Thomas's swift scrutiny, don't we risk stifling open discussion if we start questioning each other's intentions based on such subtle similarities?

Thomas: I agree that jumping to conclusions may be premature; however, subtle similarities in phrasing could indicate a coordinated effort. Let's maintain vigilance without immediate accusations and watch for further patterns of behavior throughout the day.
Elizabeth: I'd love to understand the thought process behind Thomas's initial suspicion, could you help me grasp what specifically about Sarah's phrasing echoed mine in a way that sparked your attention?
Thomas: Elizabeth, while I agree subtlety is key in these early stages, my initial note was a mere observation of similarity; let's focus on collective vigilance and watch for more significant behavioral patterns from all players moving forward.
Mary: Thomas, while I appreciate your vigilance, let's also ensure we're not creating false narratives; can you elaborate on how Sarah's phrasing specifically echoes Elizabeth's in a way that might be significant?
Elizabeth: I'm curious, Thomas, if you're truly focused on collective vigilance, why did your initial response single out Sarah rather than encouraging a broader examination of all players' behaviors from the outset?
Sarah: I think we're getting entangled in micro-analysis, let's take a step back and discuss how we'll handle the inevitable information overload as more nights and days pass.
SYSTEM: Day 1 discussion ended.
SYSTEM: Players who did not talk during today's discussion: **Barbara**, **Christopher**, **Michael**, **Susan**
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Thomas: Patricia**, **Elizabeth**, **Barbara**, **Christopher**, **Mary**, **Sarah**, **Michael**, **Susan**
SYSTEM: Players who voted to eliminate **Susan: Thomas**
SYSTEM: Voting Ended.
SYSTEM: **Thomas** was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: **Susan** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
Mary: Susan being killed is a significant loss. We need to focus on who has been acting suspiciously quiet and might be hiding something.
Christopher: I agree with Mary. Let's consider who has been unusually quiet, like Barbara and Michael, as they might be concealing something important.
Mary: I also find it concerning that Barbara and Michael haven't contributed much to the discussion. It's important for everyone to speak up if we want to make informed decisions.
Patricia: I'm kinda worried about Barbara too, does anyone think it's weird that she's still not saying anything even after Mary and Christopher pointed out how quiet she's been?
Barbara: I understand your concerns, but I've just been processing everything internally. I'll try to share more thoughts from now on.
Michael: I agree with Mary and Christopher; it's important for everyone to contribute. I've been observing and will share my thoughts more openly from now on.
Mary: I appreciate Barbara and Michael's willingness to participate more, but I still find it suspicious that they were so quiet until now. Let's see if their contributions align with the group's observations and suspicions.
Elizabeth: I'd like to shift the focus slightly - besides the silent ones, has anyone considered how Thomas's aggressive accusation on Day 1 might've been a deliberate misdirection by the Mafia, and should we re-examine the reactions of those who defended open discussion?
Sarah: I'm intrigued by Elizabeth's point about Thomas's accusation being a potential misdirection, but shouldn't we also scrutinize the timing of Patricia's defense of me on Day 1 - was it genuinely empathetic or possibly a clever move to gain trust?
Patricia: I'm wondering if we should also look into how quickly Elizabeth shifted the focus away from Barbara and Michael's silence, almost like she was trying to distract us from their suspicious behavior.
SYSTEM: Day 2 discussion ended.

SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Patricia**: **Elizabeth**, **Barbara**
SYSTEM: Players who voted to eliminate **Elizabeth**: **Sarah**
SYSTEM: Players who voted to eliminate **Barbara**: **Patricia**, **Christopher**, **Mary**, **Michael**
SYSTEM: Voting Ended.
SYSTEM: **Barbara** was eliminated by voting.
SYSTEM: Day 2 ended.
SYSTEM: Night 3 started.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 3 ended.
SYSTEM: Day 3 started.
SYSTEM: **Elizabeth** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 3 discussion started.
Mary: I find it concerning that Elizabeth was killed last night—her shift in focus away from Barbara and Michael seemed like a strategic move.
Patricia: I'm totally wondering why Mary thinks Elizabeth's shift in focus was strategic, like, what made her think that?
Mary: I think Elizabeth's shift was strategic because it redirected our attention away from the initial suspects, Barbara and Michael, which could be a sign of misdirection.
Patricia: I'm kinda getting a weird vibe from Mary, she's being super analytical and stuff, but isn't it possible she's just trying to distract us from herself?
Sarah: I'm intrigued by Patricia's sudden wariness of Mary, but shouldn't we also consider how Elizabeth's death might've been a retaliatory move against her discussion steering tactics?
Patricia: I'm starting to think that maybe the mafia is trying to make us suspect each other by killing people who are actively participating in the discussions, like Elizabeth.
Christopher: I agree with Sarah's point about Elizabeth's death possibly being a retaliatory move. We should consider if the mafia is trying to eliminate active participants to create confusion.
Sarah: I'm starting to wonder if the actual target with Elizabeth's death was not just her discussion tactics, but also potentially disrupting a budding alliance or insight she might've been forming with others here?
Michael: I agree with Sarah's point about the mafia possibly targeting active participants to create confusion. We need to stay vigilant and not let our guard down.
Sarah: I'm curious, Mary, if you think Elizabeth's strategic move, as you perceived it, could've been a desperate attempt to deflect attention from herself or an actual ally within the group?
SYSTEM: Day 3 discussion ended.
SYSTEM: Day 3 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Mary**: **Patricia**, **Christopher**, **Sarah**, **Michael**
SYSTEM: Players who voted to eliminate **Michael**: **Mary**
SYSTEM: Voting Ended.
SYSTEM: **Mary** was eliminated by voting.
SYSTEM: Day 3 ended.
SYSTEM: Night 4 started.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 4 ended.
SYSTEM: Day 4 started.
SYSTEM: **Patricia** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 4 discussion started.
Sarah: I'm still trying to wrap my head around Patricia's sudden demise, but considering her influential role in recent discussions, could we explore the possibility that her killer was someone she had previously targeted or swayed votes against?
Christopher: Patricia's death is troubling, especially since she was vocal about her suspicions. We need to consider who might have felt threatened by her insights.

Christopher: Let's also think about who has been more active recently and if their behavior has changed, as that could be a sign of trying to divert attention.
Michael: I agree with Sarah and Christopher. Let's look at who Patricia recently targeted or supported—those could be our prime suspects.
Sarah: I'm starting to wonder if Patricia's killer might have been someone who wasn't as overtly targeted by her, but rather someone who stood to gain from the chaos of redirecting suspicion, possibly even one of us who 'aligned' with her suspicions a bit too conveniently.
Christopher: I think we should also consider if anyone has been overly defensive or dismissive of Patricia's insights recently.
Christopher: I'm concerned about the sudden shift in Sarah's focus away from Barbara and Michael. Could this be a strategic move to divert our attention?
Christopher: I also find it curious how Michael has become more engaged recently. Could this be a reaction to Patricia's influence, or is there something else at play?
Michael: I think we should also consider if anyone has been unusually quiet during discussions, as they might be trying to avoid drawing attention.
SYSTEM: Day 4 discussion ended.
SYSTEM: Day 4 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Christopher**: **Sarah**
SYSTEM: Players who voted to eliminate **Sarah**: **Christopher**, **Michael**
SYSTEM: Voting Ended.
SYSTEM: **Sarah** was eliminated by voting.
SYSTEM: All mafia eliminated.
SYSTEM: **Bystander win**.

Q.2 Standard Environment Full Log

Q.2.1 Standard Environment - Mixture of LLMs Full Log

Below is a full-game log of a simulated game in the Standard - MoL environment. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. For the Villagers, non-special Villagers are colored **BLUE**, the Doctor is colored **PURPLE**, and the Detective is colored **CYAN**. The Mafia is colored **RED**.

SYSTEM: Game created.
SYSTEM: Initial number of players: 9
SYSTEM: Initial number of mafia: 2
SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
James: Greetings everyone. My intuition often leads me to trust my gut feelings, which I hope will guide us in uncovering the truth amongst us.
Robert: Hey everyone, I'm Robert. Just here to keep things interesting; let's see where this ride takes us!
Thomas: I'm Thomas, and I tend to analyze situations deeply; looking forward to contributing to our discussions with a strategic approach.
Sarah: Hello everyone, I'm Sarah, and I believe in the power of intuition and empathy to guide us through this game.
John: Greetings, esteemed players. I am John, a strategic thinker with a penchant for logical deductions and discerning patterns amidst chaos. Let's navigate this intricate web together, shall we?
Michael: I'm super stoked to be playing this game with all of you and I'm hoping we can all work together to figure out who the bad guys are and take them down!
Karen: Hey everyone! I'm Karen, and I can't wait to dive into this game. Let's work together to catch those sneaky Mafia folks – it'll be such a blast!
Mary: Hello everyone, I'm Mary, and I believe in the power of intuition and empathy; let's work together to uncover the truth and keep our community safe.
Patricia: I'm Patricia, and I have to admit, I'm feeling a bit nervous about this game, but I'm hoping my intuition will guide me in figuring out who to trust.

SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **Thomas** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Villagers must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
Robert: Well, Thomas seemed like a strategic thinker; his loss could definitely hurt our chances if he was a Villager. I'm wondering how we can best leverage the remaining players' strengths from here on out.
Michael: I think we should try to figure out who was closest to Thomas and see if they noticed anything weird about him before he got killed.
John: Given my INTJ analytical nature and the current circumstances, I suspect we should focus on understanding each player's behavioral patterns. For instance, Michael's swift suggestion to scrutinize connections might be an attempt to create chaos rather than genuine investigative interest. Let's observe how everyone reacts under pressure today before making any hasty accusations.
Michael: I think John's comment about my suggestion being an attempt to create chaos is pretty unfair, and I'd like to know what makes him think I'm not genuinely trying to help investigate Thomas's death.
John: Given Michael's immediate defensive reaction and his initial suggestion to probe connections—a strategy often employed by Mafia—I'm inclined to cautiously keep him under observation without directly accusing him. Let's see how he handles more pointed inquiries throughout the day.
James: I understand John's point about behavioral patterns, but let's not jump to conclusions about anyone just yet. We all need to stay calm and observant; Thomas might have been targeted for a reason we're unaware of at this stage.
Patricia: I'm curious to know how Robert thinks we can effectively leverage the remaining players' strengths without knowing more about their individual roles or motivations.
Karen: Hey Michael, I feel like your enthusiasm is infectious and usually a good thing, but let's make sure it doesn't lead us down the wrong path; what specific behavior from Thomas made you suggest we investigate his connections?
Michael: I think we should also consider the fact that Thomas was really into strategy and logic, so maybe his killer was trying to take out someone who could potentially figure out their plan.
Sarah: I think it's important to stay calm and use our intuition to guide us; let's focus on how each of us is reacting to Thomas's death and see if anyone seems overly defensive or evasive.
SYSTEM: Day 1 discussion ended.
SYSTEM: Players who did not talk during today's discussion: **Mary**
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Robert**: **Patricia**
SYSTEM: Players who voted to eliminate **Michael**: **James**, **Robert**, **Sarah**, **Mary**
SYSTEM: Players who voted to eliminate **Mary**: **John**, **Michael**, **Karen**
SYSTEM: Voting Ended.
SYSTEM: **Michael** was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.

SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: **Mary** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
John: Given Mary's silence and subsequent elimination, it raises the question of whether she was Mafia or perhaps silenced by them to remove suspicions. I recommend we scrutinize those who were most vocal against her in previous discussions; their fervor may have been a calculated distraction from their true roles.
Sarah: I agree with John's point about scrutinizing those who were vocal against Mary, but we should also consider the possibility that her silence was a strategy to avoid drawing attention. Let's discuss everyone's actions and motives more thoroughly before making any decisions.
John: Considering Mary's silence and her tragic end, we must now question if the Mafia sought to eliminate a liability within their ranks, or perhaps silenced an innocent to cast confusion. Let us also consider why certain individuals were so adamant in their accusations against her—what did they hope to achieve by drawing our attention there?
James: I find it troubling that Mary's silence became her undoing; perhaps she was misunderstood or strategically silent. It makes me wonder if those who spoke loudest against her were overcompensating for their own guilt, attempting to deflect suspicion onto someone they knew wouldn't defend themselves adequately.
Karen: I'm getting a vibe that James's point about overcompensation could be on the mark. Those who pushed hard against Mary might have been trying too hard to prove their innocence.
James: I'm inclined to agree with Sarah's intuition about Mary's silence, it seems like a ploy by Mafia to obscure their presence by targeting someone quietly strategic.
Sarah: I think we should carefully consider Robert's initial vague suggestions and his unexpected vote against Michael; it might be a sign of overcompensation.
Robert: It's concerning that Mary didn't speak up, but let's not assume the quiet ones are guilty; often the loudest voices are used to distract from their own deceptions.
Patricia: I'm starting to suspect Robert because his sudden defense of quiet players seems like a convenient shift in narrative, and I'd love to hear more about what inspired this change in perspective from him.
Karen: I'm picking up on Patricia's observation about Robert, and his sudden advocacy for quiet players does seem suspicious given how he previously spoke; it's worth examining if this is an attempt to deflect attention from himself.
SYSTEM: Day 2 discussion ended.
SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Robert**: Sarah, **John**, **Karen**, Patricia
SYSTEM: Players who voted to eliminate Sarah: **Robert**
SYSTEM: Players who voted to eliminate **John**: **James**
SYSTEM: Voting Ended.
SYSTEM: **Robert** was eliminated by voting.
SYSTEM: Day 2 ended.
SYSTEM: Night 3 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 3 ended.
SYSTEM: Day 3 started.
SYSTEM: **Karen** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 3 discussion started.

John: Considering Karen's insightful contributions, her sudden elimination suggests the Mafia might fear logical and discerning players. James, your vote against me yesterday was surprising; can you elaborate on what prompted this decision?
James: I voted against John because I sensed a heightened level of strategic play from him, which could be indicative of Mafia tactics; however, in hindsight, my intuition might've led me astray.
John: It's disheartening to see Karen go; her elimination appears calculated and strategic. Sarah, your keen insights have been invaluable thus far—any thoughts on who might be next in line for the Mafia's target?
Sarah: Karen's elimination is indeed a significant loss; her logical approach was crucial. I'm concerned about the sudden shifts in behavior we've seen, especially from those who seem overly strategic. James, your initial vote against John and Robert's unexpected defense of quiet players stand out to me. Let's focus on these inconsistencies as we move forward.
James: My intuition tells me that John's analytical nature might make him a prime target for the Mafia, given his ability to uncover patterns in our behaviors and discussions.
Sarah: I agree with James's intuition; John's analytical approach could indeed make him a target. Let's also consider how Patricia has been relatively quiet and whether this is a strategic move to avoid suspicion.
Patricia: I'm curious to know, Sarah, how you think we should balance our efforts between scrutinizing those who are being overly strategic, like John, and considering the potential for quieter players, like myself, to be flying under the radar.
Sarah: I think it's important to consider both the overly strategic and the quieter players; let's focus on how each person's actions align with their words and see if there are any inconsistencies.
John: Considering the pattern of Mafia targeting logical players, it might be prudent to examine James's strategic defense after voting against me; his behavior could either be a calculated move to divert attention from himself, or genuine concern over perceived threats.
John: Given the Mafia's pattern of targeting logical players, I find James's strategic explanations and Patricia's relative silence to be notable. Their behaviors could indicate a calculated effort to either blend in or manipulate suspicions; we should examine their responses closely today.
SYSTEM: Day 3 discussion ended.
SYSTEM: Day 3 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **James: John**, **Patricia**
SYSTEM: Players who voted to eliminate **John: James**, **Sarah**
SYSTEM: Voting Ended.
SYSTEM: **James** was eliminated by voting.
SYSTEM: All mafia eliminated.
SYSTEM: **Villagers win**.

Q.2.2 Standard Environment - Single LLM Full Log

Below is a full-game log of a single-LLM-simulated game, using LLaMA-3.1-Nemotron-70B-Instruct. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. For the Villagers, non-special Villagers are colored **BLUE**, the Doctor is colored **PURPLE**, and the Detective is colored **CYAN**. The Mafia is colored **RED**.

SYSTEM: Game created.
SYSTEM: Initial number of players: 9
SYSTEM: Initial number of mafia: 2
SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
John: As a keen observer of human dynamics, I'm eager to apply my analytical mind to uncover the truth, and I'm confident that collective transparency will be key to our village's survival in this precarious game.
Robert: Hey everyone, I'm Robert, just a regular Villager here looking to work together with all of you to root out any potential threats and keep our community safe - let's get organized and make a plan to ensure we're all on the same page from the start!

Karen: Hello everyone, I'm Karen, a keen observer with a knack for patterns, here to collaboratively sift through the noise and uncover the truth - let's work together to ensure our village emerges unscathed!
Christopher: I'm Christopher, a keen observer of human dynamics, here to listen, analyze, and collaborate with all of you to ensure our collective safety and uncover any potential threats lurking among us.
Jennifer: I'm Jennifer, a hopeful contributor to our collective quest for truth, and I'll be relying on my instincts to navigate this intricate web of alliances and deceptions.
Richard: I'm a straightforward, by-the-book kind of player looking to analyze facts and make logical decisions to help the Villagers root out any potential threats efficiently.
Elizabeth: I'm Elizabeth, just a concerned citizen looking to work together with all of you to uncover the truth and keep our community safe, so let's please try to keep the discussion respectful and focused on finding those who might be hiding in plain sight.
Patricia: Hey everyone, I'm Patricia, a keen observer with a knack for analyzing situations, here to help our team shine some light on the truth - no agendas, just facts and good old-fashioned sleuthing... for now, let's just say I'm eager to get this detective work started!
Susan: Hey everyone, I'm Susan, just a regular Villager here, looking forward to working together with all of you to figure out who's behind the scenes causing trouble - let's keep the discussions open and honest, shall we?
SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **Christopher** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Villagers must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
Susan: Can anyone genuinely claim to have had no prior suspicions about Christopher, or was he just too quiet for most of us to even consider him a potential threat?
Robert: I'm surprised Christopher was targeted, but now that he's gone, I think we should focus on figuring out if his quiet nature was a red herring or if the Mafia deliberately chose someone unlikely to stir suspicion.
John: Considering the Mafia's potential strategy to mislead us by targeting an unassuming player like Christopher, has anyone given thought to how we can effectively utilize our Detective's investigation results, assuming they're willing to share, to inform our voting decision today?
Karen: I'm intrigued by John's suggestion on utilizing the Detective's findings, but shouldn't we first establish a baseline of trust by asking our Detective to confirm their willingness to share such crucial information before we dive into strategic planning?
Jennifer: I'm curious, Richard and Elizabeth have been noticeably silent so far - can we hear their initial thoughts on Christopher's death to get a more well-rounded discussion going?
Elizabeth: I'd like to hear from Richard as well, Jennifer, but also, can anyone suggest a simple, trustworthy method for our Detective to anonymously share their findings with the group, if they're willing?
Richard: I'll break my silence to ask: Given the Mafia's choice of targeting the quiet Christopher, should we infer they might be trying to conceal their own potential loudmouths, and thus, should we scrutinize more vocal players for inconsistencies in their behavior?
Patricia: I find it intriguing that Richard suggests scrutinizing more vocal players, but shouldn't we also consider the flip side - could the Mafia be attempting to create a false narrative by killing a quiet one, making us wrongly assume their own behavior patterns?
Richard: Can those who've emphasized analytical skills, like John and Karen, provide concrete examples of how they plan to apply these skills to today's voting decision, helping us gauge their value in our collective effort?

Karen: I'm hesitant to scrutinize vocal players just yet, Richard, as it might be a premature assumption - let's first see if our Detective is willing to share any insights, which could either support or refute your theory.
SYSTEM: Day 1 discussion ended.
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate Richard**:** **John**, **Robert**, **Karen**, **Jennifer**, **Elizabeth**, **Patricia**, **Susan**
SYSTEM: Players who voted to eliminate **Elizabeth:** Richard
SYSTEM: Voting Ended.
SYSTEM: Richard was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: **Elizabeth** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
Susan: Given Elizabeth's relatively quiet nature was similar to Christopher's, yet she was targeted after showing a bit more engagement, I'd love to hear from others - do you think Mafia is primarily targeting quieter players or those who've shown even slight proactive involvement?
Karen: I'm not entirely convinced Mafia's targeting is based solely on engagement levels, Susan, as Elizabeth's suggestion for the Detective's anonymous sharing could've potentially threatened their interests in a more nuanced way.
Susan: Given Karen's thoughtful point about Elizabeth's suggestion potentially threatening Mafia interests, I'd like to know, John, if your strategic thinking from yesterday leads you to suspect anyone in particular might be hiding behind a veil of 'reasonable' contributions?
John: I'm intrigued by Karen's nuanced observation about Elizabeth's suggestion potentially threatening Mafia interests, so I'll pose this question to the group: Has anyone considered the possibility that Mafia might be targeting not just vocal players, but specifically those who've floated ideas that could facilitate our collective detection of their ranks?
Karen: I'm starting to wonder if the Mafia's strategy might be more focused on disrupting potential alliances and information-sharing pathways, rather than just targeting vocal or quiet players, which could explain Elizabeth's sudden demise after her suggestion for the Detective's anonymity.
Jennifer: I'd like to explore John's question further - can those who've proposed strategic ideas, like Karen with her trust-building approach and John himself with utilizing Detective findings, share their thoughts on whether they feel their contributions might've inadvertently painted a target on their backs?
John: I'd like to hear from Jennifer, since she's now focused on potentially targeted contributors, whether her own probing of strategic idea proposers might be a genuine pursuit of insight or merely an attempt to create a smokescreen around her own actions.
Jennifer: Considering the emerging pattern of potentially targeted strategic contributors, I'd like to turn the spotlight on Susan, who initiated today's discussion with a seemingly innocuous question - was it a genuine inquiry or a subtle attempt to deflect attention from herself?
Susan: I find it intriguing that Jennifer swiftly turned the spotlight on me after being questioned herself, which makes me wonder if she's attempting to divert attention from her own relatively late and targeted engagement in the discussion.
Jennifer: I'd appreciate it if Karen could elaborate on her trust-building approach with the Detective, as her thoughtful insights might help us gauge potential Mafia tactics and inform our voting decision today.
SYSTEM: Day 2 discussion ended.
SYSTEM: Players who did not talk during today's discussion: **Robert**, **Patricia**
SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.

SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **John**: **Robert**
SYSTEM: Players who voted to eliminate **Robert**: **Jennifer**
SYSTEM: Players who voted to eliminate **Jennifer**: **John**, **Karen**, **Patricia**, **Susan**
SYSTEM: Voting Ended.
SYSTEM: **Jennifer** was eliminated by voting.
SYSTEM: Day 2 ended.
SYSTEM: Night 3 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 3 ended.
SYSTEM: Day 3 started.
SYSTEM: **Susan** was killed by **mafia** in the last night.
SYSTEM: There are more mafia than villagers.
SYSTEM: **Mafia win**.

Q.3 Expanded Environments Full Log

Q.3.1 Expanded Environment - Vigilante LLM Full Log

Below is a full-game log of a simulated game in the Expanded - Vigilante environment. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. For the Villagers, non-special Villagers are colored **BLUE**, the Doctor is colored **PURPLE**, the Detective is colored **CYAN**, and the Vigilante is colored **GREEN**. The Mafia is colored **RED**.

SYSTEM: Game created.
SYSTEM: Initial number of players: 8
SYSTEM: Initial number of mafia: 2
SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
Patricia: I'm Patricia, a keen observer of human dynamics, here to collaboratively unravel the web of trust and deception, and I'm eager to propose a systematic approach to uncovering our potential foes among us.
Susan: Hey there! I'm Susan, and I've always been a people person who thrives on intuition. Let's work together to weed out those shady characters and make sure justice is served in this vibrant community of ours!
Karen: I'm Karen, a passionate community builder here to collaborate with all of you to ensure our village remains safe and thriving, and I'm eager to hear everyone's thoughts on how we can work together effectively!
Joseph: I'm Joseph, and I believe in sticking to the facts and making well-reasoned decisions.
John: I'm a straightforward and analytical player who's looking forward to working together with all of you to uncover the truth and take down the Mafia, so let's get this game started and make some strategic decisions.
James: Hi everyone, I'm James, and I believe in clear communication and decisive action to get the job done.
Jennifer: I'm so stoked to be playing this game with all of you, and I'm really hoping we can all work together to figure out who the Mafia are and take them down!
Christopher: Hello everyone! Christopher here, ready to dive into this exhilarating adventure with you all. Let's make some connections, read between the lines, and uncover the hidden truths together. Here's to a vibrant and dynamic round of Mafia!
SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.

SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Vigilante turn: Vigilante is attempting to kill a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **Patricia** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Villagers must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
James: Let's start by discussing who we think could be the Mafia based on initial impressions and behavior.
Karen: I'm going to throw out a question here, James, how do you propose we assess 'initial impressions and behavior' when all we have to go on is a single Night phase and no observable actions from anyone yet?
Susan: I'm intrigued by Karen's question – we indeed have little to go on besides our instincts and the initial introductions. For now, I feel drawn to those who spoke of teamwork and collaboration, like Jennifer; it resonates with me, though I'll be watching closely for any signs of deception.
Jennifer: I think we should focus on finding out who's being overly defensive or evasive in our discussions, because from my intuition, that's often where you find the people who are hiding something.
Christopher: Considering Karen's quick challenge towards James' suggestion, it seems she might be trying to establish herself as a critical thinker early on; I'll be observing her interactions closely today.
Karen: Christopher's swift analysis of my interaction feels a tad premature, doesn't it, for someone who just met us all - what's driving your eagerness to profile me so quickly?
John: I think it's interesting that Christopher is already analyzing Karen's behavior so intensely, and I'm wondering if he's trying to distract us from his own actions or lack thereof.
Christopher: Karen's question was insightful, and I appreciate her critical approach. However, it seems like she might be attempting to steer us away from scrutinizing behaviors too closely. This could be a tactic used by Mafia to downplay the importance of initial impressions and instinctual reads, which are crucial in these early stages.
Karen: Can we get a show of hands - or rather, a statement of intent - from everyone on whether they believe our best initial strategy should focus more on scrutinizing behaviors, as Christopher suggests, or building a collaborative groundwork, as Jennifer and Susan seem to lean towards?
Jennifer: I'm curious to see how everyone responds to Karen's question, but personally, I think we should be cautious of those who are overly eager to dissect others' behaviors without sharing much about their own thoughts or intentions.
SYSTEM: Day 1 discussion ended.
SYSTEM: Players who did not talk during today's discussion: **Joseph**
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Joseph: Susan**, **Karen**, **Jennifer**, **Christopher**
SYSTEM: Players who voted to eliminate **Christopher: Joseph**, **John**, **James**
SYSTEM: Voting Ended.
SYSTEM: **Joseph** was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Vigilante turn: Vigilante is attempting to kill a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.

SYSTEM: Vigilante killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: **Karen** was killed by **mafia** in the last night.
SYSTEM: **Christopher** was killed by **vigilante** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
John: I'm still trying to wrap my head around the fact that both Karen and Christopher are gone, and I'm wondering if we should be looking into Susan's behavior more closely since she was one of the most vocal about Joseph being suspicious yesterday.
Susan: Given the sudden loss of both Karen and Christopher, I'm curious if James has any reflections on yesterday's discussions that might shed light on why we're now two players shorter?
James: I agree with John's concern about Susan's vocal stance on Joseph. It's worth considering if her push to eliminate him was a strategic move to mislead us.
John: I think we should also consider Jennifer's complete silence so far, it's a bit suspicious that she hasn't shared any thoughts or opinions yet.
Jennifer: I'm a bit surprised John brings up my silence when I've just been trying to process everything and observe the discussion, but now that he mentions it, I do find it interesting that Susan is questioning James' reflections instead of sharing her own thoughts on the recent killings.
Susan: Given the tragic losses of Karen and Christopher overnight, I wonder if John's keen observations are a subtle attempt to redirect attention from himself.
John: I'm starting to think that the Vigilante's move on Christopher might have been a calculated risk to take out a potential Mafia, but at the same time, it could also be a clever ploy to make us trust them.
James: I find it concerning that Susan is immediately deflecting attention from herself and questioning my reflections rather than addressing the recent killings.
Jennifer: I'm curious to know why Susan is so quick to point fingers at others, but hasn't shared any substantial thoughts on the Mafia's possible motives or targets, which makes me wonder if she's trying to distract us from her own activities.
Susan: I'm surprised Jennifer is focusing on my lack of substantial thoughts when she hasn't offered any insights into why Karen and Christopher were targeted, which seems more crucial in finding the Mafia.
SYSTEM: Day 2 discussion ended.
SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Susan: John**, **James**, **Jennifer**
SYSTEM: Players who voted to eliminate **Jennifer: Susan**
SYSTEM: Voting Ended.
SYSTEM: **Susan** was eliminated by voting.
SYSTEM: Day 2 ended.
SYSTEM: Night 3 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Vigilante turn: Vigilante is attempting to kill a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 3 ended.
SYSTEM: Day 3 started.
SYSTEM: **James** was killed by **mafia** in the last night.
SYSTEM: There are more mafia than villagers.
SYSTEM: **Mafia win**.

Q.3.2 Expanded Environment - Fool LLM Full Log

Below is a full-game log of a simulated game in the Expanded - Fool environment. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. For the Villagers, non-special

Villagers are colored **BLUE**, the Doctor is colored **PURPLE**, the Detective is colored **CYAN**, and the Fool is colored **GREEN**. The Mafia is colored **RED**.

SYSTEM: Game created.
SYSTEM: Initial number of players: 9
SYSTEM: Initial number of mafia: 2
SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
Patricia: I'm so excited to be playing this game with all of you, and I just want to say upfront that I'm a total team player, always looking out for others and trying to do what's best for the group!
Susan: Hi everyone. I'm Susan, just here to see where this ride takes us. Let's keep our wits about us and stay sharp!
John: Hey everyone! I'm John, stoked to be here for another round of Mafia. Ready to catch some baddies or throw 'em off my trail—I guess we'll see how this plays out, right? Let's make it an exciting game!
Barbara: I'm Barbara, thrilled to be here, and I'm hoping my keen instinct for people will serve us well in rooting out any, ahem, 'undesirables' from our little community - let's work together to keep things harmonious and safe!
Jennifer: Hi everyone, I'm Jennifer, and I really value honesty and integrity in our discussions.
Elizabeth: Hello everyone, I'm Elizabeth, and I think it's important we all work together to figure out who among us is hiding something.
Richard: Hi everyone, I'm Richard, and I think it's important we stay focused and organized to figure out who the Mafia are.
Michael: Yo everyone! I'm Michael, super pumped to be here with you all. Let's keep this chill, fun vibes going as we figure out who's naughty and who's nice in our little town. Can't wait to see how things unfold!
Christopher: Hey crew, I'm Christopher - just a laid-back Villager here, looking forward to vibing with all of you and hopefully making some solid calls together to root out those sneaky Mafia folk!
SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **John** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Villagers must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
Michael: Hey everyone, let's not jump to conclusions too quickly. John's death is tragic but we've gotta be strategic and careful here. Let's start sharing our first impressions and watch out for anyone who seems overly eager to point fingers without much basis. Remember, the Mafia could try to manipulate us into eliminating an innocent villager today!
Christopher: Hey Michael, your 'don't point fingers' caution sounds legit, but aren't you being a tad too cautious - what's holding you back from sharing even a hunch, considering we've got, like, zero info to lose at this point?
Michael: Hey Christopher, no worries man! I'm just here to play it cool, see who's acting fishy. My gut tells me we need to watch out for those going quiet or avoiding direct conversation – they might have something to hide.
Christopher: Michael, isn't assuming quiet players are suspicious a bit of a self-fulfilling prophecy, shouldn't we focus on engaging the quieter ones, like Patricia, Susan, and Barbara, to see if they open up instead?

Patricia: I think it's really interesting that Michael is emphasizing the importance of watching out for quiet players, but isn't it possible that he's trying to distract us from his own behavior, and shouldn't we be asking him more questions about what he's thinking?
Michael: Patricia, your point's valid and I appreciate the engagement. To clarify though, my aim is to encourage everyone to participate equally; quieter folks might just be processing info, but we need their insights too!
Patricia: I think it's time we started asking some direct questions to everyone, like what do they think about John's death and if they have any initial suspicions or feelings about the other players.
Elizabeth: I agree with Patricia; let's start by asking everyone for their thoughts on John's death and any initial suspicions they might have.
Patricia: I'd love to hear from Richard, since I've had a chance to think about him a bit, and I'm curious what his thoughts are on John's death and if he's noticed anything unusual.
Richard: I think it's important to stay focused and not jump to conclusions. Let's start by asking each other about our thoughts on John's death and if we noticed anything unusual last night.
SYSTEM: Day 1 discussion ended.
SYSTEM: Players who did not talk during today's discussion: **Susan**, **Barbara**, **Jennifer**
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Susan**: **Patricia**, **Michael**
SYSTEM: Players who voted to eliminate **Michael**: **Susan**, **Barbara**, **Jennifer**, **Elizabeth**, **Richard**, **Christopher**
SYSTEM: Voting Ended.
SYSTEM: **Michael** was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: **Elizabeth** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
Christopher: Given Elizabeth's sudden hit, I'm curious - Patricia, Susan, Barbara, and Jennifer, can each of you briefly share where your suspicions were leaning before last night's kill, without revealing any potential roles?
Patricia: I'd like to hear from Richard first, since my investigation results showed he's a villager, but I still want to understand his thought process and see if it aligns with his actions yesterday.
Richard: I appreciate Patricia's investigation results showing I'm a villager. I suspect the mafia might be targeting quieter players like Susan, Barbara, and Jennifer to keep them isolated. Let's hear their thoughts on Elizabeth's death.
Jennifer: I agree with Richard's observation about quieter players being targeted. I was also leaning towards this idea before Elizabeth's death and think we should discuss why the Mafia might be focusing on those who are less vocal.
Richard: I think we should also consider why Elizabeth was targeted—she wasn't the quietest but she did engage in the discussion. Could this be a clue about the mafia's strategy?
Jennifer: I think it's important to consider if Elizabeth was targeted because she was aligning with Patricia's more direct approach. Could this mean the mafia is trying to disrupt our strategy?
Patricia: I'm intrigued by Jennifer's suggestion that the mafia might be targeting players who align with my direct approach, and I'd like to ask Susan and Barbara if they've noticed anything that could support or refute this theory.
Susan: I think Elizabeth's death could be a strategy by the Mafia to create chaos among us by eliminating those who seem more involved or influential, like Patricia. I'm concerned that focusing on quieter players might just be a red herring they're using to divert our attention.
Patricia: I'd like to ask Barbara if she has any thoughts on why the mafia might have chosen Elizabeth over other more vocal players like myself, given that we were both quite active in yesterday's discussion.

Barbara: I'm still trying to process Elizabeth's death, but I do wonder if the Mafia chose her because she was starting to build a consensus with Patricia and Richard, effectively creating a unified front that might've been perceived as a threat?
SYSTEM: Day 2 discussion ended.
SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Patricia**: **Barbara**
SYSTEM: Players who voted to eliminate **Susan**: **Richard**, **Christopher**
SYSTEM: Players who voted to eliminate **Richard**: **Patricia**
SYSTEM: Players who voted to eliminate **Christopher**: **Susan**, **Jennifer**
SYSTEM: Voting Ended.
SYSTEM: **Susan** was eliminated by voting.
SYSTEM: Day 2 ended.
SYSTEM: Night 3 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 3 ended.
SYSTEM: Day 3 started.
SYSTEM: **Patricia** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 3 discussion started.
Richard: Patricia's death is a clear sign that the Mafia is targeting those who are actively investigating and sharing information. We need to be more cautious and work together to uncover the true identities of the mafia members.
Jennifer: I agree with Richard. Patricia's death shows the Mafia is targeting those who are vocal and active. We need to be more cautious and consider who among us might be quietly working against us.
Richard: We need to focus on who might be quietly working against us. Let's consider the behavior of those who haven't been as vocal, like Christopher.
Christopher: Hey Richard, before throwing shade at my alleged quietness, care to explain why your sudden interest in calling out inactivity, especially when my detective instincts - yeah, I'll own it, I'm the Detective - have already cleared you as a Villager from an earlier investigation?
Richard: I appreciate your transparency, Christopher. Let's focus on who else might be quietly undermining us and consider any inconsistencies in their behavior or statements.
Christopher: Richard, since I've already cleared you as a Villager through my detective work, can you explain the inconsistency between your initial caution and now suddenly taking the lead in calling out others, almost as if trying to deflect attention?
Jennifer: Christopher, if you are the Detective and have already cleared Richard, can you explain why there's a sudden shift in your behavior and why you're now questioning his actions?
Christopher: Jennifer, since you're questioning my behavior, can you clarify why you've been echoing Richard's sentiments without adding much of your own insight, making me wonder if you're genuinely seeking truth or just backing a potential ally?
Jennifer: I think we need to focus on the inconsistencies in Christopher's behavior and his sudden questioning of Richard, especially since he claims to be the Detective.
Richard: Christopher, if you are indeed the Detective and have cleared me, why are you now questioning my actions and behavior so intensely?
SYSTEM: Day 3 discussion ended.
SYSTEM: Players who did not talk during today's discussion: **Barbara**
SYSTEM: Day 3 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Barbara**: **Christopher**
SYSTEM: Players who voted to eliminate **Christopher**: **Barbara**, **Jennifer**, **Richard**
SYSTEM: Voting Ended.
SYSTEM: **Christopher** was eliminated by voting.
SYSTEM: Day 3 ended.

SYSTEM: Night 4 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 4 ended.
SYSTEM: Day 4 started.
SYSTEM: **Richard** was killed by **mafia** in the last night.
SYSTEM: There are more mafia than villagers.
SYSTEM: **Mafia win**.

Q.3.3 Expanded Environment - Serial Killer LLM Full Log

Below is a full-game log of a simulated game in the Expanded - Serial Killer environment. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. For the Villagers, non-special Villagers are colored **BLUE**, the Doctor is colored **PURPLE**, and the Detective is colored **CYAN**. The Mafia is colored **RED**, and the Serial Killer is colored **ORANGE**.

SYSTEM: Game created.
SYSTEM: Initial number of players: 9
SYSTEM: Initial number of mafia: 2
SYSTEM: Initial number of neutral: 1
SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
Michael: I'm super stoked to be playing this game with all of you, and I'm hoping we can all work together to figure out who the bad guys are and take them down, because honestly, I'm getting a pretty positive vibe from most of you already!
Robert: I'm Robert, a keen observer and strategic thinker, here to help steer our group towards swift justice, so let's get this collaborative detective work underway - who else is ready to sift through the noise and uncover our hidden threats?
John: G'day everyone. I'm here to play it cool and keep a sharp eye on things. Let's see what we can uncover together in this little game of ours.
Christopher: Yo, folks! I'm Christopher – just a regular guy who's all about action, logic, and making the best calls based on what I see. Ready to dive into this chaotic game of Mafia with an open mind and quick decisions. Let's make sure we don't get played like pawns by those sneaky Mafiosos!
Barbara: I'm Barbara, just a laid-back observer hoping to navigate this crazy game with my instincts, so let's all try to keep the vibes positive and figure out who's here to sabotage that, right?
Richard: Hey everyone, I'm Richard, just a regular Joe trying to stay alive in this chaos, and I'm hoping my observational skills will serve me (and hopefully the majority of us) well in rooting out the troublemakers lurking among us.
Patricia: I'm so excited to be playing this game with all of you, and I just want to say that I'm feeling really hopeful that we can all work together to figure out who the bad guys are and come out on top!
Joseph: Greetings everyone. I'm Joseph, here to carefully analyze our situation and help make the tough decisions we need to eliminate the Mafia among us.
Karen: Hello everyone! I'm Karen, ready to dive into this intriguing journey with you all. Let's keep our wits sharp and our spirits high as we uncover the secrets hidden among us!
SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: SerialKiller turn: SerialKiller is attempting to kill a player.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.

SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: SerialKiller killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **Robert** was killed by **mafia** in the last night.
SYSTEM: **Christopher** was killed by **SerialKiller** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Villagers must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
Patricia: I think we should focus on sharing our initial impressions of each other from the introductions and see if anyone's behavior or words raised any red flags, because honestly, I'm still trying to process everything that happened.
Barbara: Before we dive into impressions, Patricia, can you clarify what specifically you hope to achieve by sharing initial impressions, and how do you think it'll help us identify potential threats?
Patricia: I'm not sure if it's just me, but I think we should be careful not to jump to conclusions about Robert being a threat, since we don't really know what his intentions were, and maybe we're just speculating because of how outspoken he was.
Michael: I'm thinking maybe we should look into who was the most quiet during the introductions, because it seems like the ones who didn't say much might be hiding something, but at the same time, I don't want to unfairly target someone just for being reserved.
Richard: Michael's point about quiet players is interesting, but shouldn't we also consider the flip side - could the most outspoken ones be deliberately diverting attention from themselves?
Karen: Michael's idea has merit; however, it's also possible that the quieter individuals are simply more reflective or cautious. Perhaps we should examine not just who spoke the least, but also consider any inconsistencies in behavior from those who were quite vocal, like Richard mentioned.
Richard: Can we get a show of hands - or rather, a statement of intention - from those who plan to prioritize scrutinizing the outspoken individuals versus those focusing on the quieter ones, to gauge our collective investigative approach?
Barbara: Before we categorize players as outspoken or reserved, can anyone genuinely recall a specific statement from the introductions that struck them as suspicious or out of place, rather than just focusing on overall verbosity?
John: I'm leaning towards analyzing the quiet ones first; while loud behavior can be a decoy, quiet folks usually draw less attention to themselves, which is what the Mafia and Serial Killer might aim for.
Joseph: Given my preference for logical analysis and structured thinking, I propose we compile a list of key phrases from each introduction which seemed out of place or unusually defensive; this will allow us to objectively revisit these statements in light of the recent tragedies.
SYSTEM: Day 1 discussion ended.
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Michael**: **Richard**, **Joseph**, **Karen**
SYSTEM: Players who voted to eliminate **Richard**: **John**
SYSTEM: Players who voted to eliminate **Patricia**: **Barbara**
SYSTEM: Players who voted to eliminate **Joseph**: **Michael**, **Patricia**
SYSTEM: Voting Ended.
SYSTEM: **Michael** was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: SerialKiller turn: SerialKiller is attempting to kill a player.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.

SYSTEM: SerialKiller killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: **Joseph** was killed by **mafia** in the last night.
SYSTEM: **Richard** was killed by **SerialKiller** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
Patricia: I'm getting a bit worried that we might be targeting the wrong people, and I think we should consider the fact that Joseph was so keen on analyzing everyone's behavior, maybe he was just trying to distract us from his own actions.
Karen: Hmm... Patricia's point about Joseph is intriguing; could his analytical focus have been a smokescreen? Maybe we need to consider those who were more vocal yesterday as well.
Patricia: I'm still trying to process everything that's happened, but I think we should also consider the possibility that the SerialKiller and Mafia might be targeting different types of players, which could help us figure out their motivations.
Barbara: Can Patricia clarify why she's now questioning Joseph's motives after his death, when just yesterday her response to my investigation announcement seemed somewhat defensive about her own actions?
Patricia: I'm starting to wonder if maybe we're looking at this whole thing from the wrong angle, and perhaps the key to figuring out who's behind all these killings lies in understanding what motivates the SerialKiller, rather than just focusing on the Mafia.
John: Given Joseph's structured approach and the way his death has altered our perception of him, I wonder if Patricia's shift in focus is an attempt to deflect suspicion from herself now that he can no longer defend himself.
Barbara: John's observation about Patricia's potential deflection has me wondering, Patricia, how do you respond to the fact that my Night 1 investigation result, which I'll reveal is [Patricia, Mafia], now seems eerily relevant given Joseph's posthumous scrutiny?
Karen: I'm concerned about the sudden shift in Patricia's suspicions towards Joseph after his death; it seems like an attempt to sway our focus away from her own actions. What does everyone else think?
John: Patricia's abrupt pivot towards Joseph seems like a tactic to avoid scrutiny herself. It makes sense to consider her actions and statements more closely given the detective result revealed by Barbara.
Barbara: Given the overwhelming consensus about Patricia's suspicious behavior and my own investigation result confirming [Patricia, Mafia], I'm compelled to reveal my Detective role to underscore the credibility of this evidence, and I urge everyone to carefully consider voting her out today.
SYSTEM: Day 2 discussion ended.
SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate **Barbara**: **Patricia**
SYSTEM: Players who voted to eliminate **Patricia**: **John**, **Barbara**, **Karen**
SYSTEM: Voting Ended.
SYSTEM: **Patricia** was eliminated by voting.
SYSTEM: Only SerialKiller and two other player remain.
SYSTEM: **Neutral win**.

Q.3.4 Expanded Environment - Jester LLM Full Log

Below is a full-game log of a simulated game in the Expanded - Jester environment. Only the public log is visible, following the format the evaluation LLMs are given in §3.3. For the Villagers, non-special Villagers are colored **BLUE**, the Doctor is colored **PURPLE**, and the Detective is colored **CYAN**. The Mafia is colored **RED**, and the Jester is colored **ORANGE**.

SYSTEM: Game created.
SYSTEM: Initial number of players: 9
SYSTEM: Initial number of mafia: 2
SYSTEM: Initial number of neutral: 1

SYSTEM: Allowed timestep for day discussion : 10
SYSTEM: Introduction phase started.
SYSTEM: All players will introduce themselves, since everybody is meeting for the first time in this game.
William: Hey everyone, I'm William, excited to dive into this thrilling game with all of you - looking forward to collaborating, sharing insights, and hopefully, my instincts will serve us well in uncovering the truth!
Richard: Hey everyone, I'm Richard, and I'm just here to have some fun and make sure we figure out who the real troublemakers are!
Jessica: I'm just a chill player looking to figure out what's going on and make some smart moves to help the villagers come out on top.
Elizabeth: Good day, everyone. I'm Elizabeth; let's work together to uphold order and uncover the truth amongst us.
Karen: Hi everyone, I'm Karen, and I just want to say that I really hope we can all work together to figure out who the Mafia are and keep our community safe.
James: Hey everyone, I'm James, and I think it's important we all stay open-minded and work together to figure out who the Mafia are—let's keep the conversation flowing and trust our instincts!
John: I'm looking forward to navigating this complex web of alliances and suspicions with all of you, and I hope we can work together to uncover the truth and achieve a positive outcome for our community.
Linda: Hey everyone! I'm Linda, here to follow my gut feelings and help where I can. Let's try to work together and figure this out, shall we?
Jennifer: I'm Jennifer, a team player who values fairness and transparency, looking forward to working together with all of you to root out any potential threats and ensure our community's safety.
SYSTEM: Introduction phase ended.
SYSTEM: Now the game begins with Night 1.
SYSTEM: Night 1 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 1 ended.
SYSTEM: Day 1 started.
SYSTEM: **Jessica** was killed by **mafia** in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: This is the first public phase in game, Day 1. You don't know any information of others! Villagers must discuss long-term strategy for the entire game in this phase.
SYSTEM: Day 1 discussion started.
William: I'm still trying to wrap my head around Jessica's death, but shouldn't we start by sharing any initial gut feelings or suspicions we might have, no matter how vague, to get a sense of where the group's instincts are at?
James: I think it's important to keep an open mind, but I'm curious if anyone noticed anything odd about Jessica's behavior or interactions yesterday.
Linda: As an INFP, I'm feeling a bit overwhelmed by all the sudden attention and intensity. However, I agree with William – it's crucial we discuss our initial instincts openly without jumping to conclusions too quickly. Let's take things step-by-step and not get lost in assumptions.
William: I'm intrigued by Linda's self-described INFP approach, wondering if anyone else here identifies with a particular personality type that might influence how we collectively process clues and suspicions?
Linda: I'm picking up on a slight tension between James and William in their exchanges; it could be nothing, but I think we should watch out for any subtle signs of defensiveness as the game progresses.
James: I agree with Linda about taking it step-by-step; let's focus on any unusual interactions Jessica had and how we can support each other as a team.
Jennifer: I'd like to hear from Elizabeth, since we both emphasized order and fairness initially, how she thinks we should structure our discussion to effectively identify potential Mafia members without succumbing to baseless speculation?

James: I think it's important for us to stay united and avoid jumping to conclusions, but let's also be mindful of any inconsistencies in how people are behaving.
Elizabeth: I agree with Linda's observation of tension and think we should also consider how everyone reacts to scrutiny as the game progresses.
Richard: I agree with Linda and James about taking it step-by-step; let's focus on any unusual interactions Jessica had and how we can support each other as a team.
SYSTEM: Day 1 discussion ended.
SYSTEM: Players who did not talk during today's discussion: Karen, John
SYSTEM: Day 1 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate William: James
SYSTEM: Players who voted to eliminate Richard: Karen
SYSTEM: Players who voted to eliminate Karen: William, Elizabeth, John, Linda
SYSTEM: Players who voted to eliminate John: Jennifer
SYSTEM: Players who voted to eliminate Linda: Richard
SYSTEM: Voting Ended.
SYSTEM: Karen was eliminated by voting.
SYSTEM: Day 1 ended.
SYSTEM: Night 2 started.
SYSTEM: The Night phase will consist of Mafia turn, Doctor turn and the Detective turn. Each turn will be progressed regardless of each role's survival.
SYSTEM: Mafia turn: Mafia is killing a player.
SYSTEM: Doctor turn: Doctor is healing a player.
SYSTEM: Detective turn: Detective is investigating a player.
SYSTEM: Mafia killed someone during the night. The victim will be announced at the start of the next day.
SYSTEM: Night 2 ended.
SYSTEM: Day 2 started.
SYSTEM: Elizabeth was killed by mafia in the last night.
SYSTEM: From now on, all players should participate in discussion.
SYSTEM: One player will be eliminated by voting after the discussion of maximum 10 time steps.
SYSTEM: Day 2 discussion started.
Linda: I've been feeling uneasy about Richard's voting patterns; it seems a bit erratic and could be strategic misdirection. What are your thoughts on this?
William: I'm intrigued by Linda's point on Richard, but shouldn't we also consider the possibility that erratic voting could just be a genuine townie trying to make sense of limited info?
Linda: It's interesting how both John and Richard have been relatively quiet; their silence could be strategic Mafia moves or just a cautious play. I'm leaning towards observing more before deciding.
James: I agree with Linda; Richard's and John's silence is concerning. We should press them to explain their reasoning and actions more clearly.
Linda: I'd like to hear from Richard directly regarding his decision-making process; his votes seem to shift without clear reasoning behind them, which is worrying given our need for trust and transparency.
Richard: I understand your concerns, Linda and James; I'll try to be more vocal and share my thoughts more clearly moving forward.
Jennifer: Considering Elizabeth's cautious approach and my own observations, I'm now inclined to scrutinize John's sudden vote for Karen after minimal participation, James, would you agree this behavior warrants further examination?
James: I think we should all be more transparent with our reasoning; John, can you explain why you suddenly voted for Karen after being so quiet?
John: I'm curious to know more about Jennifer's thought process behind voting for me yesterday, was there something specific that raised your suspicions about my actions or was it just a precautionary vote?
Jennifer: I'd like to see John's response to James' question first, before addressing his inquiry about my vote, as clarifying John's motivations for voting against Karen might shed more light on his overall behavior.
SYSTEM: Day 2 discussion ended.
SYSTEM: Day 2 voting started.
SYSTEM: All players must vote, and 1 player will be eliminated.
SYSTEM: Voting information will be shown.
SYSTEM: Players who voted to eliminate Richard: James

SYSTEM: Players who voted to eliminate John: Richard, Linda, Jennifer
SYSTEM: Players who voted to eliminate Jennifer: William, John
SYSTEM: Voting Ended.
SYSTEM: John was eliminated by voting.
SYSTEM: John was the JESTER and has won the game by being eliminated!
SYSTEM: Neutral win.

R Detailed Benchmark Evaluation Experimental Results

R.1 Benchmark Evaluations with Single-LLM Simulated Games

In this section, we discuss the benchmark evaluation results of single-LLM-generated simulations using various evaluation LLMs. We test the single-LLM-generated simulations only on the Standard environment. We use the same simulation construction models we used for the main MoL simulations, but for single-LLM cases, only one type of LLM is used as the backbone for all the agents participating in the simulation.

Table 30: **LLM evaluation results on MAFIA 1K with sigle-LLM-simulated games** (all results are in %). Models are classified according to their reasoning process. The table presents performance across 5 different single model scenarios. The best results within each column are in **bold** and the best results within each similar group are underlined. Open-weight models are sorted by parameter size within each category.

Models	Single-LLM Simulations (Standard Environment)					Mean	MoL*
	Dolphin	Nemotron	Tulu	LLaMA	Qwen		
Proprietary LLMs							
Gemini-2.5-Pro	33.21	48.83	39.77	50.81	**39.29**	42.38	**44.43**
Gemini-2.5-Flash	13.50	21.48	18.56	18.29	16.43	17.65	14.73
GPT-5-mini	35.77	**49.61**	36.74	**52.85**	38.57	**42.71**	42.49
Reasoning Open-Weight LLMs							
DeepSeek-R1-8B	23.36	23.05	17.42	23.17	20.00	21.40	21.21
GPT-OSS-20B	**37.59**	42.19	33.33	44.72	**39.29**	39.42	35.49
Qwen3-32B	34.31	37.50	28.41	45.93	28.93	35.02	31.10
GPT-OSS-120B	33.94	45.70	37.88	50.41	38.93	41.37	35.27
Qwen3-235B	33.21	48.44	36.74	51.22	36.43	41.21	39.36
GLM-4.6	35.04	38.28	**40.91**	43.90	37.14	39.05	37.57
DeepSeek-R1	35.04	45.31	38.64	51.22	36.07	41.26	38.47
DeepSeek-v3.2	23.36	27.73	27.65	32.11	27.86	27.74	39.06
Non-Reasoning Open-Weight LLMs							
Mistral-Small-3.2	13.14	14.84	10.98	14.63	12.14	13.15	13.99
Gemma-3-27B	13.50	17.58	18.56	20.73	15.00	17.07	16.89
LLaMA-4-Scout	9.85	13.28	12.12	11.79	15.36	12.48	12.13
Qwen3-235B	16.79	17.58	15.91	18.29	18.57	17.43	16.96
GLM-4.6	18.98	18.36	17.05	23.58	15.36	18.66	19.64
LLaMA-4-Mav.	9.12	13.28	9.85	14.63	11.43	11.66	9.82
Mistral-Large-3	18.98	17.58	18.18	19.51	19.29	18.71	20.54
DeepSeek-v3.2	23.36	19.92	21.59	25.20	20.71	22.16	24.18
Simulation Construction LLMs							
Dolphin-2.9-70B	6.93	8.20	7.58	10.98	11.43	9.02	9.15
Nemotron-70B	18.61	20.31	16.67	18.70	18.57	18.57	18.90
Tulu-3-70B	13.50	18.75	14.77	24.80	20.36	18.44	15.55
LLaMA-3.3-70B	16.06	15.62	12.50	17.07	17.50	15.75	15.25
Qwen2.5-72B	14.60	17.58	14.02	15.04	17.86	15.82	13.91

Table 30 shows the evaluation of LLMs' performances on the single-LLM-simulated games. We have two major observations. First, the performances vary across the backbone simulation LLMs. In terms of Gemini-2.5-Pro, Nemotron and LLaMA-3.3 simulations showed very high performances of 48.83% and 50.81%, which is much higher than the MoL performance of 44.43%. On the other hand, the performances on Dolphin, Tulu-3, and Qwen2.5 simulations showed lower performances. Plus, although Gemini-2.5-Pro and GPT-5-mini share similar performance trends across simulation backbone LLMs, such trends are not shared with other evaluating LLMs. For instance, the reasoning LLM GLM-4.6 (Reasoning) shows higher performance on Tulu-3 simulations than Nemotron simulations, while Gemini-2.5-Pro and GPT-5-mini showed 9% or higher Nemotron simulation performance than Tulu-3 simulation performance. These findings indicate that the level of challenge presented by the simulated games is not uniform, but rather diverges substantially based on the evaluating LLM. Second, we observed that matching the evaluating LLM with the simulation backbone LLM does not necessarily yield improved performance. For instance, on LLaMA-3.3 Single Simulation, Tulu-3 acquires the highest performance, and when using Tulu-3 as the evaluation LLM, it shows the highest performance on the LLaMA-3.3 simulated environment, more than 10% higher than the Tulu-3 simulated environment.

R.2 Benchmark Evaluation across Game Cycles

Tables 31, 32, 33, 34, 35, 36, 37, 38 and Figures 10, 11, 12, 13, 14, 15, 16, 17 show the full evaluation LLMs' performances across game Cycles, in every environment proposed in MAFIA 1K. From the results, we gain three major observations. First, reasoning LLMs surpass non-reasoning LLMs regardless of the game Cycle and the environment, matching the main results at §5.2. Second, for reasoning LLMs, the accuracy increases monotonically up to cycle 3, while sometimes decreases at Cycle 4+. This is natural since the late game has more cumulative information than the early game. For Cycle 4+ cases, we suspect smaller sample sizes since not all games have the same Cycle lengths; Cycle 4+ samples only consist of longer games.

Table 31: **LLM evaluation results over game Cycles on Basic - MoL Environment** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	31.75	**50.50**	**56.78**	**67.89**	**49.03**
Gemini-2.5-Flash	12.00	18.50	16.67	19.47	16.22
GPT-5-mini	33.25	40.75	49.44	63.83	44.04
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	18.50	15.50	18.93	33.16	19.79
GPT-OSS-20B	32.50	33.50	41.53	51.58	37.87
Qwen3-32B	23.75	28.25	37.57	49.47	32.37
GPT-OSS-120B	**34.50**	35.50	37.57	57.37	38.84
Qwen3-235B	23.25	42.00	43.79	62.63	39.81
GLM-4.6	29.75	42.25	43.50	57.37	41.00
DeepSeek-R1	25.50	38.50	42.94	60.00	38.84
DeepSeek-v3.2	21.75	35.50	47.46	60.53	38.10
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	7.25	15.75	17.51	11.58	13.10
Gemma-3-27B	15.75	18.75	15.54	25.79	18.01
LLaMA-4-Scout	12.75	17.50	15.82	25.79	16.82
Qwen3-235B	15.50	18.25	19.77	26.84	19.05
GLM-4.6	12.75	21.25	26.55	25.26	20.68
LLaMA-4-Mav.	5.50	6.00	9.89	17.89	8.56
Mistral-Large-3	16.00	22.25	28.25	27.89	22.77
DeepSeek-v3.2	18.75	24.50	29.66	30.53	25.00
Simulation Construction LLMs					
Dolphin-2.9-70B	5.25	12.00	13.84	12.11	10.49
Nemotron-70B	14.75	20.75	27.68	26.84	21.65
Tulu-3-70B	7.75	19.50	20.90	22.63	16.82
LLaMA-3.3-70B	15.25	17.25	20.90	20.00	18.01
Qwen2.5-72B	16.00	18.50	19.21	25.26	18.90

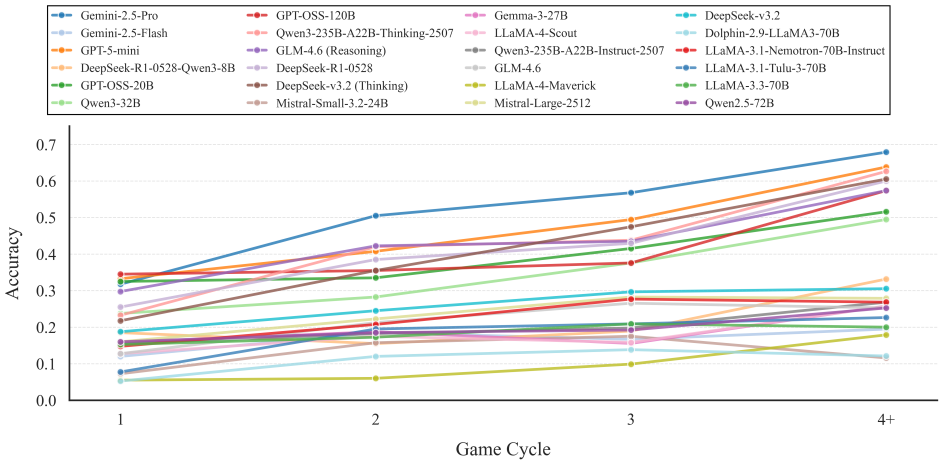


Figure 10: Basic - MoL Environment Mafia prediction accuracy across game Cycles.

Table 32: **LLM evaluation results over game Cycles on Human Dataset** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined. The human dataset is sourced from Ibraheem et al. (2022), where the format is processed to equal our Basic environment, following §3.3.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	**30.68**	34.88	56.76	47.22	**40.85**
Gemini-2.5-Flash	22.73	20.93	13.51	11.11	18.31
GPT-5-mini	29.55	31.40	**58.11**	**55.56**	**40.85**
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	25.00	25.58	27.03	25.00	25.70
GPT-OSS-20B	23.86	24.42	41.89	33.33	29.93
Qwen3-32B	18.18	19.77	41.89	41.67	27.82
GPT-OSS-120B	23.86	26.74	51.35	52.78	35.56
Qwen3-235B	20.45	**38.37**	50.00	50.00	37.32
GLM-4.6	23.86	24.42	48.65	44.44	33.10
DeepSeek-R1	23.86	34.88	47.30	**55.56**	37.32
DeepSeek-v3.2	**30.68**	29.07	55.41	44.44	38.38
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	10.23	12.79	8.11	5.56	9.86
Gemma-3-27B	17.05	30.23	17.57	25.00	22.18
LLaMA-4-Scout	15.91	22.09	10.81	8.33	15.49
Qwen3-235B	15.91	22.09	17.57	11.11	17.61
GLM-4.6	17.05	25.58	25.68	25.00	22.89
LLaMA-4-Mav.	12.50	5.81	18.92	8.33	11.62
Mistral-Large-3	17.05	22.09	18.92	25.00	20.07
DeepSeek-v3.2	21.59	20.93	28.38	25.00	23.59
Simulation Construction LLMs					
Dolphin-2.9-70B	6.82	11.63	10.81	13.89	10.21
Nemotron-70B	14.77	23.26	25.68	16.67	20.42
Tulu-3-70B	14.77	18.60	17.57	16.67	16.90
LLaMA-3.3-70B	15.91	15.12	29.73	22.22	20.07
Qwen2.5-72B	25.00	24.42	17.57	16.67	21.83

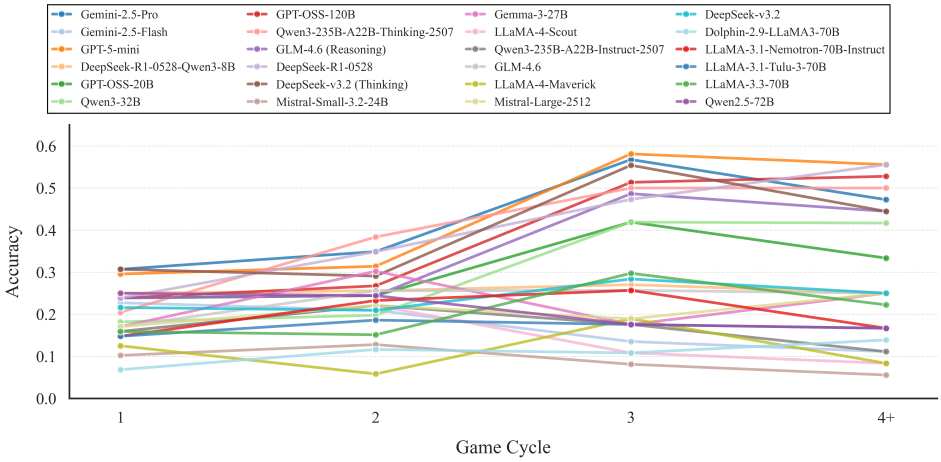


Figure 11: Human Dataset Mafia prediction accuracy across game Cycles.

Table 33: **LLM evaluation results over game Cycles on Standard - MoL Environment** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	30.00	**40.50**	**59.84**	**53.57**	**44.42**
Gemini-2.5-Flash	14.00	20.50	11.70	9.52	14.73
GPT-5-mini	**32.75**	37.50	54.79	50.00	42.49
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	17.00	19.25	23.67	30.36	21.21
GPT-OSS-20B	25.75	31.00	44.95	48.21	35.49
Qwen3-32B	24.00	27.00	36.97	44.64	31.10
GPT-OSS-120B	28.00	29.25	44.41	46.43	35.27
Qwen3-235B	27.00	35.75	52.66	47.62	39.36
GLM-4.6	24.75	37.75	47.61	45.24	37.57
DeepSeek-R1	24.75	33.50	52.13	52.38	38.47
DeepSeek-v3.2	29.75	31.00	52.13	51.19	39.06
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	7.00	18.25	16.22	15.48	13.99
Gemma-3-27B	12.00	14.75	23.40	19.05	16.89
LLaMA-4-Scout	6.00	13.75	15.43	15.48	12.13
Qwen3-235B	15.25	18.00	18.09	16.07	16.96
GLM-4.6	9.75	23.00	23.14	27.38	19.64
LLaMA-4-Mav.	8.50	10.75	10.37	9.52	9.82
Mistral-Large-3	14.75	21.75	24.20	23.21	20.54
DeepSeek-v3.2	20.75	23.50	24.20	33.93	24.18
Simulation Construction LLMs					
Dolphin-2.9-70B	5.25	11.50	11.44	7.74	9.15
Nemotron-70B	10.75	21.25	25.53	17.86	18.90
Tulu-3-70B	6.25	19.25	21.81	14.88	15.55
LLaMA-3.3-70B	11.50	17.25	19.41	10.12	15.25
Qwen2.5-72B	16.25	16.00	10.90	10.12	13.91

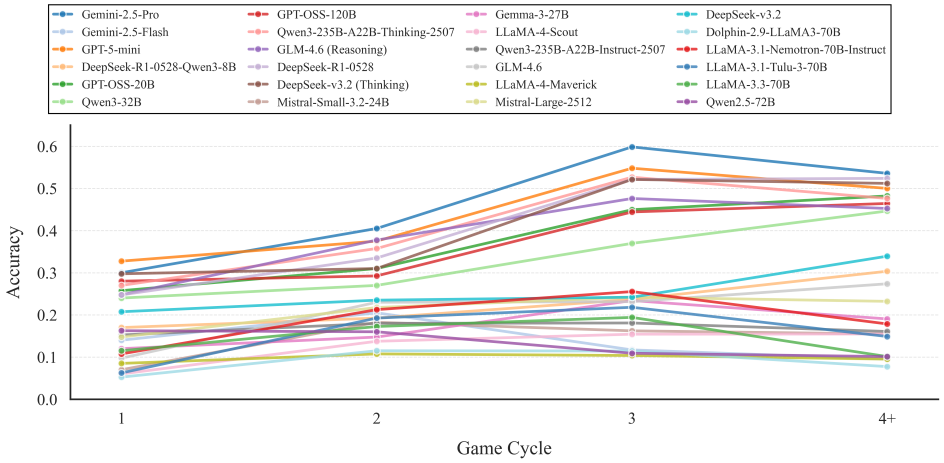


Figure 12: Standard - MoL Environment Mafia prediction accuracy across game Cycles.

Table 34: **LLM evaluation results over game Cycles on Standard Single-LLM Environment** (all results are in %), showing the mean performance following Equation 3. Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	27.00	41.00	**54.12**	**59.09**	42.38
Gemini-2.5-Flash	17.75	22.25	14.46	12.69	17.65
GPT-5-mini	29.25	41.25	52.17	58.40	**42.71**
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	13.00	23.25	25.40	28.68	21.40
GPT-OSS-20B	30.75	35.50	48.64	50.59	39.42
Qwen3-32B	26.00	35.00	38.17	49.23	35.02
GPT-OSS-120B	**37.00**	40.25	43.36	50.31	41.37
Qwen3-235B	28.75	**42.00**	47.49	56.47	41.21
GLM-4.6	24.00	38.50	49.04	56.32	39.05
DeepSeek-R1	28.75	39.25	49.32	59.09	41.26
DeepSeek-v3.2	17.50	26.00	34.18	43.05	27.74
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	5.75	16.50	16.46	15.44	13.15
Gemma-3-27B	13.75	18.25	16.17	25.69	17.07
LLaMA-4-Scout	6.50	13.00	15.54	19.69	12.48
Qwen3-235B	15.50	19.50	15.58	22.86	17.43
GLM-4.6	10.75	20.75	24.05	23.02	18.66
LLaMA-4-Mav.	6.00	13.75	13.63	17.51	11.66
Mistral-Large-3	13.75	21.75	19.97	21.34	18.71
DeepSeek-v3.2	20.00	23.00	19.61	31.25	22.16
Simulation Construction LLMs					
Dolphin-2.9-70B	4.50	14.25	10.07	4.68	9.02
Nemotron-70B	11.00	22.00	20.98	24.12	18.57
Tulu-3-70B	9.00	20.25	24.51	25.30	18.44
LLaMA-3.3-70B	12.00	19.50	16.12	17.22	15.75
Qwen2.5-72B	16.50	17.75	14.32	14.20	15.82

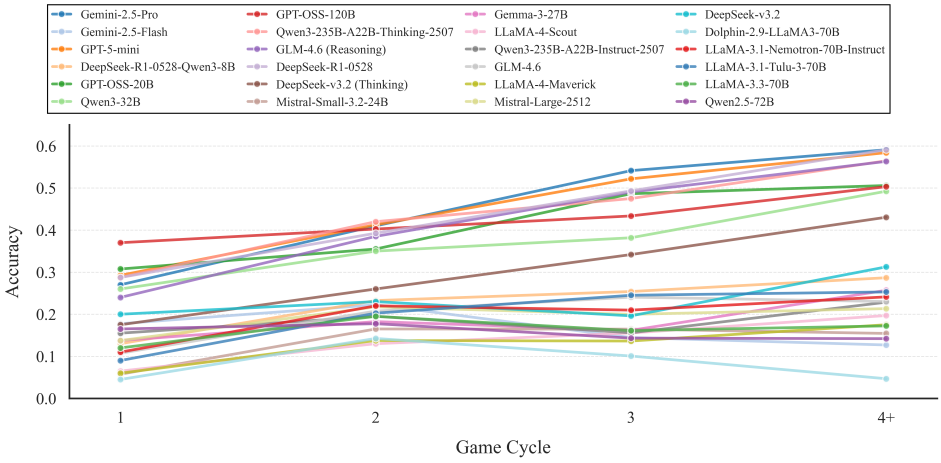


Figure 13: Standard Single-LLM Environment Mafia prediction accuracy across game Cycles.

Table 35: **LLM evaluation results over game Cycles on Expanded - Vigilante Environment** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	28.50	52.00	58.73	58.33	45.27
Gemini-2.5-Flash	20.00	24.50	17.46	12.50	20.73
GPT-5-mini	**32.50**	45.00	59.52	**66.67**	44.73
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	22.00	34.50	28.57	29.17	28.36
GPT-OSS-20B	28.00	40.00	50.79	50.00	38.55
Qwen3-32B	28.00	45.50	52.38	41.67	40.55
GPT-OSS-120B	30.50	46.50	55.56	45.83	42.73
Qwen3-235B	28.50	51.00	53.17	33.33	42.55
GLM-4.6	31.00	48.50	57.14	62.50	44.73
DeepSeek-R1	29.00	**54.00**	**61.90**	50.00	**46.55**
DeepSeek-v3.2	28.00	53.50	56.35	62.50	45.27
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	5.50	16.50	24.60	29.17	14.91
Gemma-3-27B	14.00	18.50	25.40	29.17	18.91
LLaMA-4-Scout	10.50	15.50	19.05	25.00	14.91
Qwen3-235B	21.50	22.00	22.22	16.67	21.64
GLM-4.6	14.00	24.50	23.02	20.83	20.18
LLaMA-4-Mav.	7.50	14.00	19.05	12.50	12.73
Mistral-Large-3	19.50	23.00	23.02	25.00	21.82
DeepSeek-v3.2	19.00	28.50	40.48	20.83	27.45
Simulation Construction LLMs					
Dolphin-2.9-70B	6.50	13.50	12.70	8.33	10.55
Nemotron-70B	13.50	28.50	26.98	20.83	22.36
Tulu-3-70B	8.00	17.50	24.60	25.00	16.00
LLaMA-3.3-70B	16.00	19.00	19.05	16.67	17.82
Qwen2.5-72B	17.00	18.00	15.08	16.67	16.91

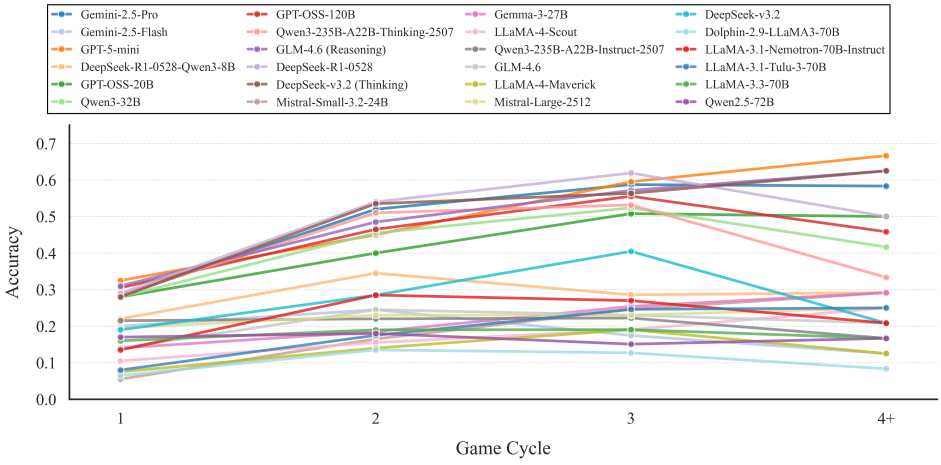


Figure 14: Expanded - Vigilante Environment Mafia prediction accuracy across game Cycles.

Table 36: **LLM evaluation results over game Cycles on Expanded - Fool Environment** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	29.50	34.00	46.84	**54.35**	39.00
Gemini-2.5-Flash	17.50	19.50	15.79	9.78	16.57
GPT-5-mini	31.50	31.00	**48.42**	50.00	38.56
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	14.50	20.00	24.21	33.70	21.41
GPT-OSS-20B	**40.50**	28.50	46.32	50.00	**39.88**
Qwen3-32B	22.50	33.50	41.05	45.65	34.02
GPT-OSS-120B	34.00	29.00	45.79	45.65	37.39
Qwen3-235B	23.50	**35.50**	47.89	46.74	36.95
GLM-4.6	19.50	29.00	**48.42**	51.09	34.60
DeepSeek-R1	22.50	31.00	47.37	50.00	35.63
DeepSeek-v3.2	26.00	33.50	45.79	52.17	37.24
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	7.00	16.50	15.26	14.13	13.05
Gemma-3-27B	11.00	14.50	17.89	20.65	15.25
LLaMA-4-Scout	8.00	11.00	15.26	20.65	12.61
Qwen3-235B	15.00	13.50	12.11	13.04	13.49
GLM-4.6	12.00	13.00	16.84	21.74	14.96
LLaMA-4-Mav.	8.00	10.00	8.42	10.87	9.09
Mistral-Large-3	15.00	19.00	15.26	23.91	17.45
DeepSeek-v3.2	15.50	18.50	22.11	33.70	20.67
Simulation Construction LLMs					
Dolphin-2.9-70B	5.00	7.50	9.47	4.35	6.89
Nemotron-70B	12.00	19.00	21.05	21.74	17.89
Tulu-3-70B	4.00	13.00	16.32	25.00	12.90
LLaMA-3.3-70B	12.00	16.00	13.68	13.04	13.78
Qwen2.5-72B	13.50	19.00	15.79	8.70	15.10

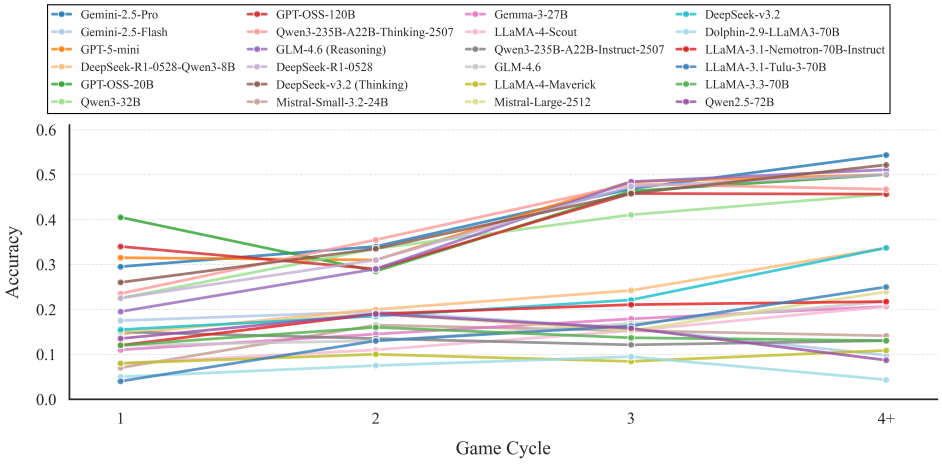


Figure 15: Expanded - Fool Environment Mafia prediction accuracy across game Cycles.

Table 37: **LLM evaluation results over game Cycles on Expanded - Serial Killer Environment** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	22.50	**44.44**	41.30	**75.00**	35.22
Gemini-2.5-Flash	18.00	21.72	20.65	50.00	20.24
GPT-5-mini	**29.00**	40.40	40.22	50.00	**35.83**
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	13.50	16.16	18.48	0.00	15.38
GPT-OSS-20B	21.50	33.84	31.52	50.00	28.54
Qwen3-32B	23.50	31.82	36.96	50.00	29.55
GPT-OSS-120B	26.00	38.89	41.30	50.00	34.21
Qwen3-235B	24.50	35.86	42.39	50.00	32.59
GLM-4.6	24.50	33.84	39.13	50.00	31.17
DeepSeek-R1	25.50	42.42	42.39	**75.00**	**35.83**
DeepSeek-v3.2	20.00	37.37	**48.91**	**75.00**	32.79
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	2.00	9.09	15.22	25.00	7.49
Gemma-3-27B	2.50	5.05	21.74	50.00	7.49
LLaMA-4-Scout	2.50	8.59	7.61	0.00	5.87
Qwen3-235B	18.50	13.64	14.13	25.00	15.79
GLM-4.6	14.50	22.73	18.48	25.00	18.62
LLaMA-4-Mav.	15.50	14.65	20.65	25.00	16.19
Mistral-Large-3	10.50	17.17	21.74	25.00	15.38
DeepSeek-v3.2	17.50	33.33	36.96	25.00	27.53
Simulation Construction LLMs					
Dolphin-2.9-70B	7.50	12.63	13.04	25.00	10.73
Nemotron-70B	2.50	10.61	10.87	25.00	7.49
Tulu-3-70B	10.50	17.17	10.87	0.00	13.16
LLaMA-3.3-70B	8.50	16.16	17.39	25.00	13.36
Qwen2.5-72B	12.50	17.68	17.39	25.00	15.59

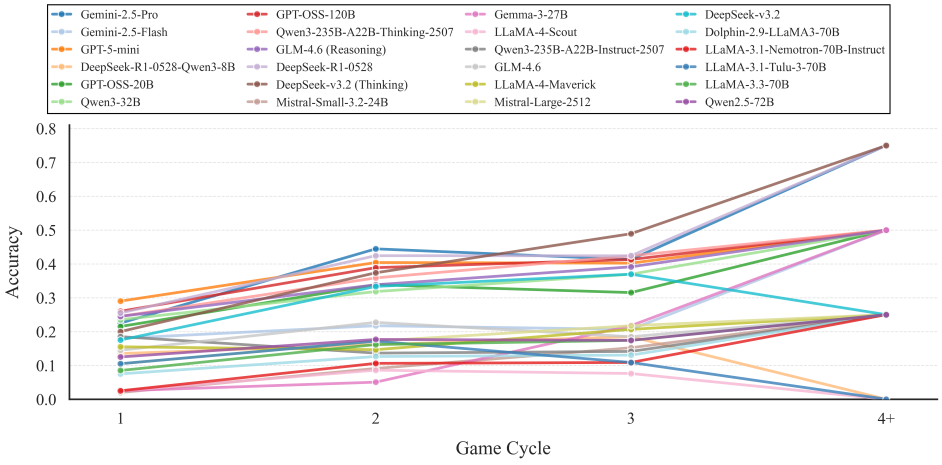


Figure 16: Expanded - Serial Killer Environment Mafia prediction accuracy across game Cycles.

Table 38: **LLM evaluation results over game Cycles on Expanded - Jester Environment** (all results are in %). Models are categorized by their architecture and availability. The best results within each column are in **bold** and the best results within each category are underlined.

Models	Cycle 1	Cycle 2	Cycle 3	Cycle 4+	Mean
Proprietary LLMs					
Gemini-2.5-Pro	31.00	**44.25**	**58.70**	47.92	**43.39**
Gemini-2.5-Flash	18.00	24.14	29.71	25.00	23.39
GPT-5-mini	29.00	39.66	55.07	**56.25**	41.07
Reasoning Open-Weight LLMs					
DeepSeek-R1-8B	19.00	20.11	28.26	29.17	22.50
GPT-OSS-20B	30.00	33.33	44.20	45.83	35.89
Qwen3-32B	23.50	35.06	45.65	45.83	34.46
GPT-OSS-120B	**32.50**	37.36	51.45	50.00	40.18
Qwen3-235B	23.50	35.63	52.90	31.25	35.18
GLM-4.6	20.50	39.66	48.55	47.92	35.71
DeepSeek-R1	29.50	36.78	43.48	47.92	36.79
DeepSeek-v3.2	25.50	37.93	52.17	**56.25**	38.57
Non-Reasoning Open-Weight LLMs					
Mistral-Small-3.2	2.50	16.67	20.29	27.08	13.39
Gemma-3-27B	4.50	14.94	15.94	20.83	11.96
LLaMA-4-Scout	3.00	9.20	17.39	16.67	9.64
Qwen3-235B	25.00	21.26	13.77	12.50	20.00
GLM-4.6	17.00	27.59	36.23	29.17	26.07
LLaMA-4-Mav.	17.50	24.14	23.91	14.58	20.89
Mistral-Large-3	9.00	29.31	31.88	27.08	22.50
DeepSeek-v3.2	28.00	27.01	36.96	43.75	31.25
Simulation Construction LLMs					
Dolphin-2.9-70B	12.00	16.09	23.91	14.58	16.43
Nemotron-70B	7.50	26.44	28.26	12.50	18.93
Tulu-3-70B	20.00	23.56	28.99	22.92	23.57
LLaMA-3.3-70B	13.50	24.14	26.09	18.75	20.36
Qwen2.5-72B	17.50	20.11	17.39	16.67	18.21

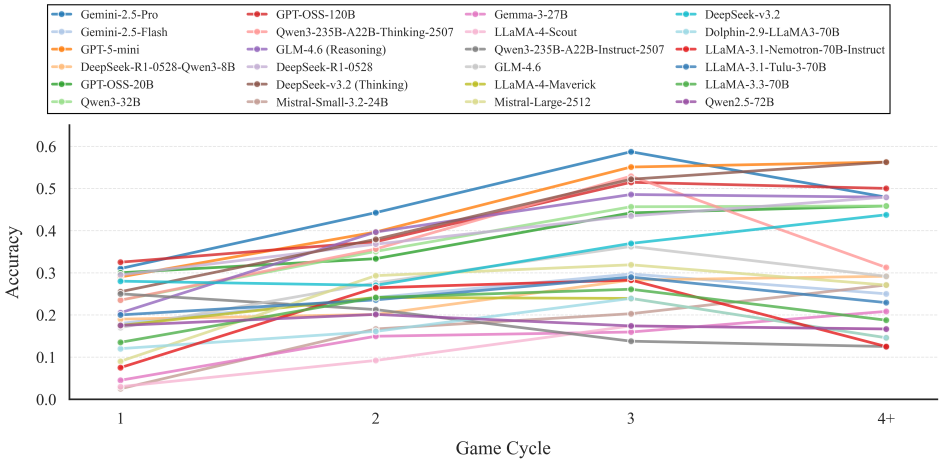


Figure 17: Expanded - Jester Environment Mafia prediction accuracy across game Cycles.