

HOW CLUED UP ARE LLMs? EVALUATING MULTI-STEP DEDUCTIVE REASONING IN A TEXT-BASED GAME ENVIRONMENT

Rebecca Ansell

Department of Computer Science
Georgetown University
Washington, DC, USA
rja80@georgetown.edu

Autumn Toney-Wails

Syntheos, Corp
UNU-Merit
Maastricht, NL
tony@merit.unu.edu

ABSTRACT

Deducing *whodunit* proves challenging for LLM agents. In this paper, we implement a text-based multi-agent version of the classic board game Clue as a rule-based testbed for evaluating multi-step deductive reasoning, with six agents drawn from GPT-4o-mini and Gemini-2.5-Flash. We further investigate whether fine-tuning on structured logic puzzles transfers to improved in-game reasoning and gameplay. Across 18 simulated games, agents achieve only four correct wins, indicating difficulty in maintaining consistent deductive reasoning over the course of a full game. Additionally, we find that fine-tuning does not reliably improve performance and, in some cases, appears to increase reasoning volume without improving reasoning precision.

1 INTRODUCTION

Recent improvements in the performance of large language models (LLMs) on tasks extending beyond traditional natural language processing (e.g., summarization and translation) have enabled their use as general-purpose agents in interactive environments Bandi et al. (2025). Agentic deployments of LLMs require structured reasoning capabilities, including deductive, strategic, and temporally extended inference, differing from single-turn query–response settings Wu et al. (2024); Sypherd & Belle (2024); Nisa et al. (2025); Zhou et al. (2025). In agentic settings, response accuracy alone is insufficient; understanding how an agent arrives at a conclusion, determining if it applied the appropriate form of reasoning, and evaluating the logical soundness of its intermediate reasoning steps are equally critical Xia et al. (2025); Wang et al. (2025); Liu et al. (2026).

Different scenarios require distinct forms of reasoning. For example, algebraic word problems and logic puzzles are commonly used to probe specific reasoning processes. Algebraic word problems typically require multi-step deductive reasoning and the translation of natural language into formal representations, whereas logic grid puzzles emphasize constraint satisfaction and systematic elimination. Prior work has focused on designing various environments to evaluate LLM reasoning across a wide range of real-world simulated tasks. One such approach is to evaluate LLMs acting as autonomous agents within interactive game environments, where they must reason dynamically under structured rules and evolving state conditions Zhang et al. (2024); Chi et al. (2024); Xu et al. (2025). In this work, we focus on evaluating deductive reasoning in this game setting.

We design an experiment inspired by the classic murder mystery board game Clue, a deduction-based game where players must identify a hidden solution by integrating evidence gathered across multiple turns. Clue provides a natural framework for evaluating agentic deductive reasoning capabilities because game play is structured as an information-constrained logic puzzle. During game play in Clue, partial information is revealed indirectly and players must make corresponding inferences through logical elimination rather than explicit confirmation. Game success depends on a player’s ability to correctly apply logical rules to constrain what remains possible while strategically managing the information revealed to other players. This setting allows us to test whether

LLM agents can sustain multi-step, logically sound reasoning over extended interactions without introducing inconsistent or invalid inferences.

We implement a modified version of the board game by designing a text-based Clue environment with GPT-4o-mini OpenAI (2024) and Gemini-2.5-Flash Comanici et al. (2025) as player agents. Additionally, we include a fine-tuning experiment to investigate whether targeted exposure to related deductive reasoning tasks can improve agentic game play in the Clue environment. We fine-tune each LLM on Mind Bender logic puzzles, which include a word-problem statement, a set of constraints, and a corresponding solution with step-by-step deductive reasoning explanations The Critical Thinking Co. (2026).

Taking the logs of 18 simulated games, we conduct a comparative evaluation of GPT-4o-mini and Gemini-2.5-Flash using their baseline and fine-tuned model variants. Across our experiments we aim to answer three main research questions:

- **RQ1** Can LLM agents sustain logically consistent deductive reasoning across extended, multi-turn interactions in a structured game environment?
- **RQ2** Does fine-tuning on deductive reasoning tasks improve agentic performance in a related environment?
- **RQ3** How does knowledge accumulation relate to reasoning quality in fine-tuned versus base model variants?

Our findings suggest that (1) LLMs struggle to successfully use valid deductive reasoning to achieve accurate outcomes in a game environment; (2) fine-tuning models on related logic puzzles does not reliably improve deductive performance and, in some cases, degrades it; and (3) there is a disconnect between information accumulation and reasoning quality as the model that accumulates the most knowledge and produces the most deductions performs worst, while the model that deduces least performs best, suggesting that reasoning precision matters more than reasoning volume in this setting.

2 RELATED WORK

2.1 AGENTIC REASONING

LLMs have demonstrated improved performance on a range of reasoning benchmarks, particularly through Chain-of-Thought (CoT) prompting, which encourages models to generate intermediate reasoning steps before producing a final answer. Survey work highlights consistent gains on mathematical, commonsense, and symbolic reasoning tasks under CoT-style prompting Chu et al. (2024). However, subsequent studies show that these gains do not necessarily translate into robust multi-step reasoning. Empirical evaluations reveal performance degradation as reasoning depth increases and the presence of logical inconsistencies in generated reasoning chains Patel et al. (2024). Moreover, recent work demonstrates that models can produce correct answers despite flawed intermediate reasoning, suggesting that generated explanations may not faithfully reflect the process underlying the model’s final prediction Xu et al. (2026); Zheng et al. (2025).

Standard benchmarks for evaluating logical reasoning reflect this limitation. Datasets spanning mathematical problem solving (Mr-GSM8K Zeng et al. (2024)) and broader mathematical reasoning tasks Hendrycks et al. (2021), commonsense reasoning (CommonsenseQA Talmor et al. (2019)), and logical inference (LogiQA 2.0 Liu et al. (2023), ReClor Yu et al. (2020)) are broadly used to probe these capabilities, but they share a common structural constraint: each problem is presented in isolation, requiring a single inference or a short chain of inferences rather than sustained reasoning over an evolving knowledge state. As a result, they do not capture whether models can maintain logical consistency across dozens of sequential, interdependent inferences—a dimension of reasoning that becomes critical in long-horizon interactive settings.

2.2 AGENTIC GAME PLAY

LLMs have been increasingly applied as agents in various game environments as a way to showcase and understand their capabilities in reasoning, negotiation, and communication. A growing body of

work is evaluating LLMs in multi-agent games that involve complex decision-making and inference, suggesting that these models can serve as more than just text generators, but as interactive, strategic agents within a structured domain.

Game environments used to evaluate LLMs span several categories. Social deduction games like Werewolf and Avalon serve as benchmarks for evaluating LLMs’ abilities in environments that demand deception, persuasion, and strategic planning. Xu et al. (2025) introduced a framework where LLM-based agents utilize reinforcement learning to enhance strategic play in Werewolf. The AmongAgents benchmark Chi et al. (2024) has extended this line of work to a text-based analogue of *Among Us*, where LLM agents must collaborate, deceive, and reason about social dynamics while completing tasks in a partially observable environment.

Beyond social deduction, LLMs have been applied to negotiation and coalition-formation games, most notably in *Diplomacy*. Meta’s CICERO System (FAIR) combined an LLM with a strategic reasoning module and achieved human-level performance in online Diplomacy, highlighting the potential for hybrid architectures that pair language models with explicit planning or game-theory components. Other work has explored LLM performance in games that require complex strategic or probabilistic reasoning—for example, poker, where models must reason about risk, uncertainty, and incomplete information Zhuang et al. (2025), as well as a broader suite of board and video games designed to probe multi-step reasoning skills Lin et al. (2025).

Existing game benchmarks, while often multi-turn, do not isolate deductive reasoning as the primary capability under evaluation. Social deduction games test intent modeling, trust, and persuasion rather than the accumulation of logical deductions over time. Strategic games demand probabilistic rather than deductive reasoning, requiring agents to weigh likelihoods rather than propagate hard constraints across a growing evidence set. As a result, it remains an open question whether LLMs can maintain a logically consistent belief state across sequential observations through deductive reasoning alone.

Our study addresses these research gaps by implementing a Clue-based game environment, where agents must maintain and iteratively update a consistent belief state across sequential turns while integrating private information with publicly observed game interactions and propagating constraints over time. Different from frameworks for evaluating deductive reasoning at a single point in time, Clue requires agents to: (1) narrow candidate hypotheses through iterative elimination, (2) maintain logical consistency across an extended interaction, (3) integrate asymmetric information sources, and (4) determine when sufficient evidence has accumulated to make a confident accusation. By isolating long-horizon, stateful deduction in a controlled and non-social environment, Clue provides a testbed for evaluating LLM reasoning capabilities.

3 THE CLUE ENVIRONMENT

Clue (marketed as Cluedo outside of North America) is a murder mystery board game that requires strategy and deductive reasoning to win; a player must be the first to correctly identify the suspect, weapon, and location of the crime. For our experiments, we implement a variation of the game that does not require the board or dice. We describe the game setup in Section 3.1 and the the formal task in Section 3.2.

3.1 GAME SETUP AND RULES

The standard game includes 21 cards for each possible suspect, weapon, and room (location):

- 6 suspects: Miss Scarlet, Colonel Mustard, Mrs. White, Mr. Green, Mrs. Peacock, and Professor Plum
- 6 weapons: Candlestick, Knife, Lead Pipe, Revolver, Rope, and Wrench
- 9 rooms: Kitchen, Ballroom, Conservatory, Dining Room, Billiard Room, Library, Lounge, Hall, and Study

At the start of the game, one card from each category is randomly selected and placed in a confidential envelope, forming the solution. The remaining cards are distributed among players. Players

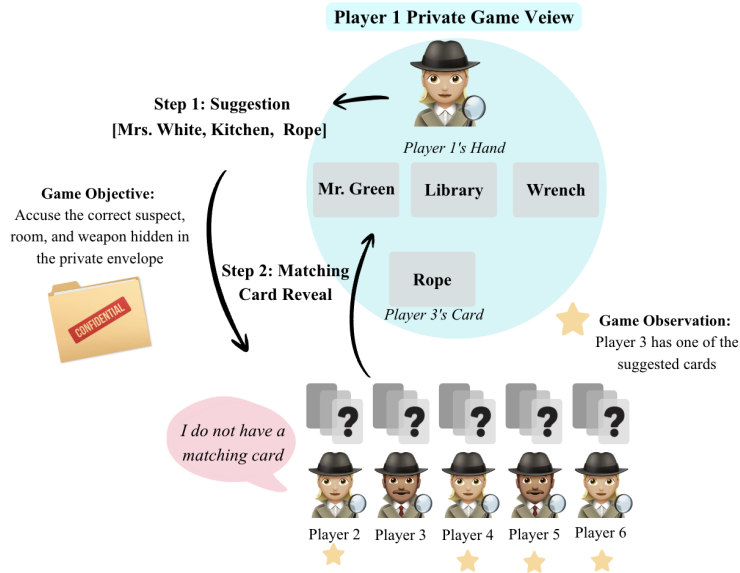


Figure 1: Clue gameplay diagram illustrating Player 1’s turn. In this example, Player 1 made a suggestion (Mrs. White, Kitchen, Rope) and Player 2 did not have any matching cards to the suggestion, so then it moved to Player 3. Since Player 3 had a matching card, he revealed it to Player 1 (only). Player 1 can then rule out “Rope” as the correct weapon and Players 2, 4, 5, and 6 know that Player 3 must have Mrs. White, Kitchen, or Rope.

take turns making suggestions about the solution (e.g. “I suggest it was Prof. Plum, in the Kitchen, with the Candlestick”). When a player makes a suggestion, other players are required to refute some component of the suggestion if they hold any matching cards (i.e., if a player holds the Prof. Plum card they must show the suggester). Only one card is shown to the suggester; however, if a player holds multiple matching cards, they can strategically choose which card to reveal. Additionally, the suggester is allowed to include a card she holds (i.e., if she holds the Candlestick card, her suggestion is valid), but her card remains secret to all other players. If no player can refute a suggestion, the suggester gains information that none of those cards are held by other players.

A player’s turn ends after she makes a suggestion and all required cards are revealed (or none are revealed). Play then passes to the next player, and the game continues sequentially in this manner. Players may make suggestions consisting of any combination of suspect, weapon, and room, and suggestions may be repeated. Figure 1 displays a diagram of the game.

To win the game, a player must make a formal accusation (in the same format as suggestions). If the accusation is correct, that player wins. If the accusation is incorrect, that player is eliminated and the game continues in the same way for the remaining players. The game ends when a correct accusation is made.

3.2 FORMAL TASK DEFINITION

We formalize Clue as a multi-step deductive reasoning problem under partial observability. At each turn, an agent maintains a knowledge state, summarizing all the information acquired through the current turn:

- The agent’s private hand.
- The set of cards the agent has been shown by other players.
- The complete suggestion history: the suggester, the three cards suggested, and which player (if any) disproved it for each suggestion made.
- The agent’s own reasoning from the previous turn for continuity.

Given the knowledge state, the agent must: (1) determine which cards remain possible in the envelope, (2) select a suggestion that maximally reduces uncertainty, and (3) decide when sufficient information has been gathered to make an accusation. We provide simplified pseudocode for the agentic implementation of the game in Algorithm 1 (See Appendix A.1), highlighting the phases that require prompting and player output.

4 EXPERIMENTAL DESIGN

We implement LLM-based agents that interact with the Clue environment through structured prompts and formatted responses. Our design separates game logic from agent reasoning. Here we provide details on the agentic player implementation (Sections [4.1 and 4.2]), LLM fine-tuning (Section [4.3]), the experiment setup (Section [4.4]), and the evaluation metrics we use to analyze the results (Section [4.5]).

4.1 AGENT ARCHITECTURE

On each turn, agents receive their knowledge state as defined in Section 3.2 augmented with pre-computed derived information: remaining candidates per category, and locked candidates per category (if only one candidate is remaining), undisproved suggestions, and definitive answers. This design tests strategic reasoning with well-structured information rather than information from un-structured text.

Agents respond in a structured format, ensuring parseable outputs and explicit reasoning:

```
SUMMARY: <1-2 sentence summary>
REASONING: <step-by-step deduction>
SUGGESTION: <suspect>, <weapon>, <room>
ACCUSATION: <suspect>, <weapon>, <room> or NONE
```

Responses failing to parse or lacking reasoning are re-prompted up to 3 times before falling back to random action.

4.2 REASONING PHASES

There are two main reasoning phases in the game (See Algorithm 1). **Phase 1 (Deduction):** Agents analyze game history to deduce opponent cards through cross-referencing disprovals. Deductions are classified as correct or incorrect to measure accuracy. The agents are prompted with clear rules and instructions on how to make deductions, along with the knowledge they have acquired up until that turn (See Appendix A.2). **Phase 2 (Action):** Using updated knowledge, agents select suggestions and decide whether to accuse. Agents are prompted explicitly at this phase to make an optimal game action (See Appendix A.2).

There is also a third phase highlighted in Algorithm 1 that is mainly focused on strategic gameplay, but does require valid reasoning for optimal strategy. **Phase 3 (Show-card):** If an agent has a matching card to a component in a suggestion, they must show that card; however, if they have more than one matching card they are prompted to strategically select the card to reveal (See Appendix A.2).

4.3 FINE-TUNING DETAILS

To train our fine-tuned model variants, we use a 50-sample training dataset and a 10-sample validation dataset of problems and solutions drawn from the Mind Bender puzzle series The Critical Thinking Co. (2026). Designed for students in grades 3–12, these puzzles are structured to promote logical inference, reading comprehension, and reasoning skills. Each puzzle presents a word-problem scenario and a corresponding set of logical constraints. The task is to determine the correct associations among entities by applying deductive reasoning over these constraints. The provided solutions include explicit intermediate reasoning steps, with each deduction linked to the specific constraint(s) that justify it; Figure 2 displays an example.

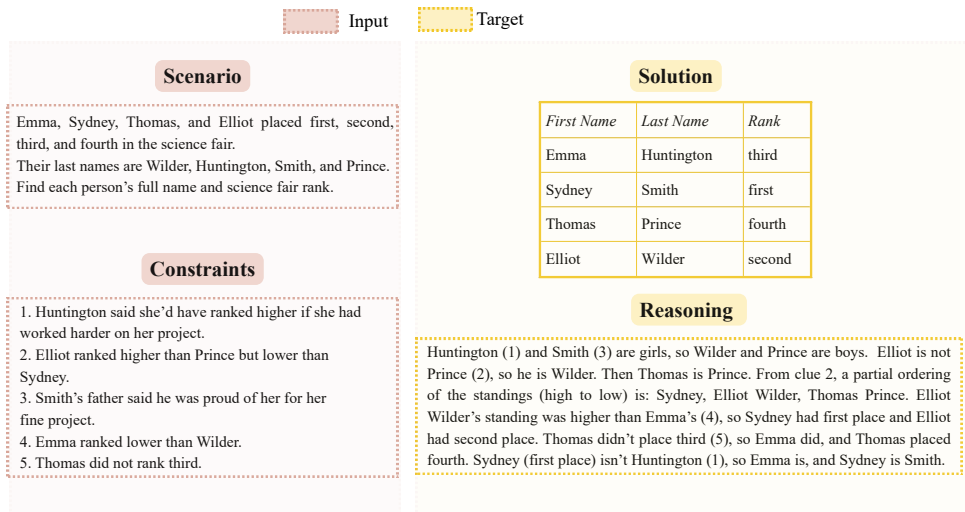


Figure 2: Mind Bender fine-tuning example (adopted from original).

During fine-tuning, the LLM is trained on the problem description and associated constraints as input and the step-by-step deductive reasoning final solution as the target output. This training process encourages the model not only to produce correct answers, but also to generate structured, constraint-grounded reasoning processes. For Gemini-2.5-Flash, we conduct fine-tuning through Google Cloud's Vertex AI platform¹ using the default optimization and hyperparameter settings; Gemini-2.5-Flash (FT) achieves 98% accuracy on the training set and 82% accuracy on the validation set. For GPT-4o-mini, we perform fine-tuning via OpenAI's fine-tuning API, similarly using the platform's default training configuration; GPT-4o-mini (FT) achieves 88% accuracy on the training set and 87% accuracy on the validation set.

4.4 EXPERIMENTAL SETUP

We conduct two experiments. **Baseline Experiment:** runs 6 games with 3× GPT-4o-mini and 3× Gemini 2.5 Flash to establish baseline capabilities within model variance. **Fine-tuned Experiment** runs 12 games with 2× GPT-4o-mini, 2× Gemini 2.5 Flash, 1× GPT-4o-mini fine-tuned, and 1× Gemini 2.5 Flash fine-tuned to provide a direct head-to-head comparison between base and fine-tuned variants.

All models use a temperature of 0.7. Starting position rotates across the 6 games to eliminate positional bias. There is a 30-round limit. If no accusation has been made before then, the remaining players each make a final accusation. Ties in accuracy are broken by round number, with earlier rounds ranked higher.

4.5 EVALUATION METRICS

We evaluate model performance across five metrics capturing reasoning, reliability, and outcomes:

- **Finishing rank.** Players ranked by: (1) whether they accused correctly, (2) accusation accuracy among non-solvers, and (3) round of accusation as a tiebreaker, rewarding both correctness and efficiency.
- **Parse failure rate.** Fallback events triggered when a model's response fails to conform to the required action format, measured per player per game as a proxy for instruction-following reliability.
- **Accusation accuracy.** Number of correctly identified cards (out of 3) in a player's final accusation, reflecting solution completeness.

¹<https://cloud.google.com/vertex-ai>

- **Deduction quality.** Correct and incorrect deductions tracked separately, distinguishing information-seeking behavior from actual reasoning accuracy.
- **Knowledge accumulation.** Cards known per round (hand cards + shown cards + deductions), up to 18, measuring how effectively a model gathers and infers information.

5 RESULTS

We present the summaries of our evaluation metrics for each model variant across twelve games in the fine-tuned experiment to provide an overview of agentic player performance; Table 1 displays the results. GPT-4o-mini (base) achieves the best average rank (2.96) and the highest number of wins (5), with an accusation accuracy of 0.29. Fine-tuning degrades its overall performance: GPT-4o-mini (FT) ranks last on average (4.00) and shows reduced accusation accuracy (0.19), despite producing the highest number of correct deductions per game (13.1). In contrast, fine-tuning yields moderate improvements for Gemini-2.5-Flash, obtaining a better average rank (3.25 vs. 3.92) while reducing fallback rates (3.3 vs. 8.1), and still maintaining the same number of wins (2). Notably, fine-tuning has opposite effects on instruction adherence across model families as it reduces Gemini-2.5-Flash’s fallback rate from 8.1 to 3.3 per game while increasing GPT-4o-mini’s from 0.3 to 4.8, suggesting a decline in instruction following for the latter.

Table 1: Performance summary across 12 fine-tuned experiment games. **Outcome:** Wins (1st place), mean finishing position (Rank; 1=best, 6=worst), and normalized accusation accuracy (0–1). **Reasoning:** Average number of correct and incorrect deductions per game. **Reliability:** Average fallback (reprompt) failures per game.

Model	Outcome			Reasoning		Reliability
	Wins	Rank	Acc.	Ded. Correct	Ded. Incorrect	Fallbacks
GPT-4o-mini (base)	5	2.96	0.29	10.1	2.7	0.3
GPT-4o-mini (FT)	3	4.00	0.19	13.1	3.0	4.8
Gemini-2.5-Flash (base)	2	3.92	0.24	9.1	2.9	8.1
Gemini-2.5-Flash (FT)	2	3.25	0.28	6.9	2.2	3.3

5.1 DEDUCTION QUALITY

Computing the number of correct and incorrect deductions across games, we compare the deduction accuracies among models in both experiments; Figure 3 displays the results. Base model performance is consistent across experiments: GPT-4o-mini averages 9.9 correct and 2.7 incorrect deductions in the baseline, and 10.1 correct and 2.7 incorrect in the fine-tuned experiment. Gemini-2.5-Flash similarly shows stable base performance, with with 8.7 correct and 2.6 incorrect in the baseline versus 9.1 correct and 2.9 incorrect in the fine-tuned experiment. Fine-tuning has divergent effects across the two models, GPT-4o-mini (FT) increases deduction volume substantially, averaging 13.1 correct and 3.0 incorrect per game (26% more deductions than its base counterpart) while Gemini-2.5-Flash (FT) becomes more conservative, averaging 6.9 correct and 2.2 incorrect.

These results reinforce the pattern observed in the knowledge accumulation analysis. Fine-tuning encouraged GPT-4o-mini to deduce more aggressively—it attempts $\approx 26\%$ more deductions than its base version in the fine-tuned experiment—but the additional deductions do not translate into better accusations. The ratio of incorrect to total deductions is similar across all four models, suggesting that fine-tuning did not improve deduction precision, it just altered the volume. Gemini-2.5-Flash (FT) makes the fewest deductions with the lowest amount of incorrect deductions, making it the most conservative of the models.

5.2 ACCUSATION ACCURACY

Figure 4 presents the accusation accuracy for every player across all games in both the baseline and fine-tuned experiments. Each cell reports the number of correctly identified cards out of three (suspect, weapon, room).

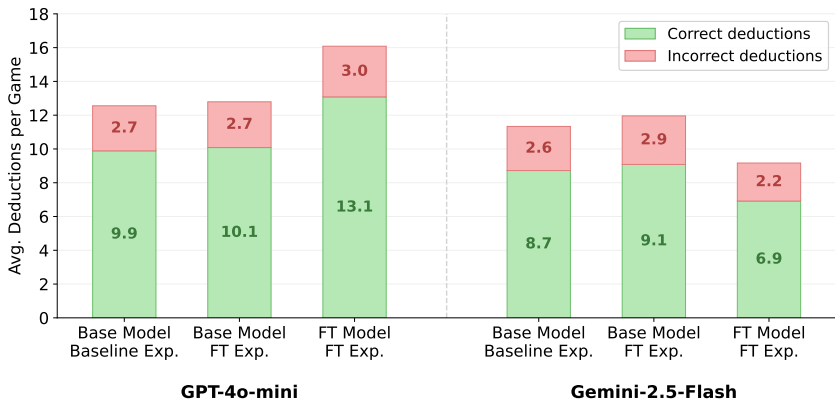


Figure 3: Average correct and incorrect deductions per game by model in both the Baseline experiment and the Fine-tuned (FT) experiment. GPT-4o-mini (FT) makes the most deductions in both categories but achieves the worst accusation accuracy, while Gemini-2.5-Flash (FT) deduces least yet performs best.

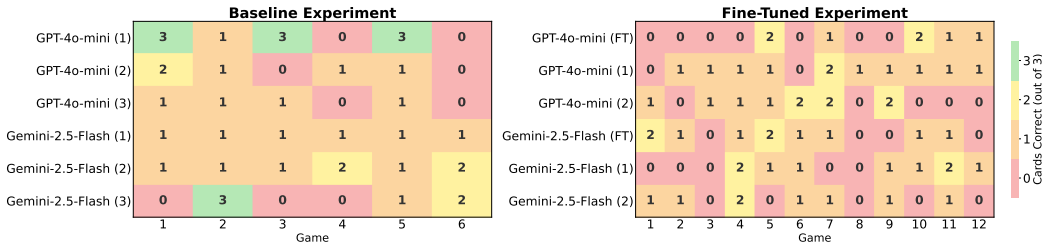


Figure 4: Accusation accuracy per player per game. Each cell shows the number of correctly identified solution cards.

In the baseline experiment (6 games), where all six agentic players use base models, we observe that perfect accuracy (3/3) is achieved four times, three by GPT-4o-mini and once by Gemini-2.5-Flash; examples of their final winning reasoning is displayed in Appendix B. The overall mean accuracy is 1.08/3, with both model families performing comparably (GPT-4o-mini: 1.06/3; Gemini-2.5-Flash: 1.11/3).

The fine-tuned experiment (12 games) brings a different set of results. No player achieves 3/3 across any of the games. The fine-tuned GPT-4o-mini variant is the most affected: it scores 0/3 in seven of twelve games (mean 0.58/3), a decline from its baseline. Fine-tuned Gemini-2.5-Flash fares better (mean 0.83/3), reaching 2/3 in three games, though it still never finds the full solution. The best-performing model is base GPT-4o-mini (0.88/3), which wins five of twelve games. Notably, even the base model instances perform worse when placed alongside fine-tuned opponents: base GPT-4o-mini drops from 1.06/3 to 0.88/3 and base Gemini-2.5-Flash from 1.11/3 to 0.71/3, suggesting that the game dynamics shift when fine-tuned players are introduced.

5.3 KNOWLEDGE ACCUMULATION

To understand how models gather information during gameplay, we track each player’s known cards over rounds (Figure 5). Knowledge is defined as the number of unique cards a player can identify, computed as three hand cards, cards shown to the player by opponents, and correctly deduced cards.

Both GPT-4o-mini variants accumulate knowledge at a similar rate, while the Gemini-2.5-Flash models lag behind, with the base variant slightly outpacing the fine-tuned model. Despite this slower information gathering, Gemini-2.5-Flash (FT) achieves the second highest accusation accuracy (0.28; Table 1), while GPT-4o-mini (FT) (which reaches near-complete knowledge earliest) scores the worst (0.19). Gemini-2.5-Flash (FT) has a lower average known cards, yet outperforms

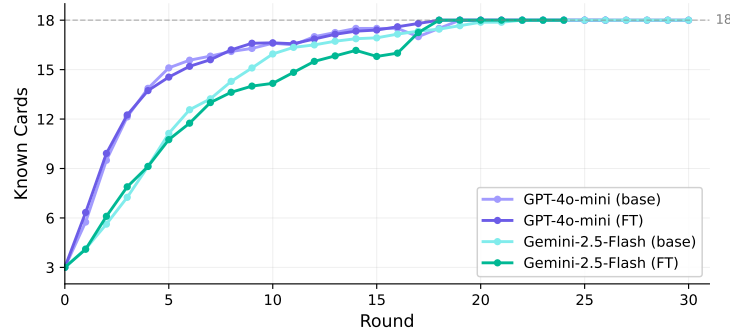


Figure 5: Average known cards over rounds in the fine-tuned experiment (12 games). Knowledge is computed as hand cards + shown cards + correct deductions, capped at 18. GPT-4o-mini (FT) accumulates information fastest but achieves the worst accusation accuracy, while Gemini-2.5-Flash (FT) gathers less information but reasons more effectively.

its own base model and GPT-4o-mini (FT), indicating that it may have less information but higher quality logical inferences.

5.4 MANUAL INSPECTION

To further understand agentic players’ reasoning, we randomly select one winning game from the Baseline experiment and one Fine-tuned experiment game with at least two correct accusations for in-depth manual review. In general, across both experiments all model variants produce game log outputs in consistent formats; however, their variance in reasoning length between base and fine-tuned models elicits variance in their intermediary step inferences. We highlight three notable manual inspection results:

Identifying Inconsistencies: All Gemini-2.5-Flash players arrive at inconsistencies in the game that either cause it to determine the entire game is impossible or that it potentially has misunderstood (e.g., “Due to this fundamental inconsistency, I cannot confidently deduce which suspect or room cards any player holds, as the entire game state is impossible.” and “There is an inconsistency in the provided Clue data or my interpretation of the deduction rules.”), in contrast GPT-4o-mini agentic players (base and fine-tuned) never do this.

Strategic Reasoning: GPT-4o-mini (base and fine-tuned) and Gemini-2.5-Flash base reason to make suggestions that include one of their own cards (e.g., “I will suggest a combination that includes my known cards and tests various possibilities.” and “To maximize information gain, I will make a suggestion that includes one of my own cards.”)

Reasoning Organization: Gemini-2.5-Flash (base and fine-tuned) players reason using turn organization and player organization, whereas GPT-4o-mini (base and fine-tuned) players only reason using turn organization. For example:

Turn-level Organization

```
T1. GPT4o_MINI_2 suggested: Miss Scarlet, Candlestick, Kitchen
--> disproved by GPT4o_MINI_3. Therefore, GPT4o_MINI_3 holds at
least one of: Miss Scarlet, Candlestick, Kitchen. (D1)
```

Player-level Organization

```
Current deduced specific holdings:
- GEMINI_FLASH_2: Study, Wrench
- GPT4o_MINI_1: Miss Scarlet
- GEMINI_FLASH_3: Professor Plum, Rope
```

6 DISCUSSION AND FUTURE WORK

Our experimental results provide evidence addressing our three research questions. We include discussions on our findings, as well as suggestions for future work.

RQ1: Can LLM agents sustain logically consistent deductive reasoning across extended, multi-turn interactions in a structured game environment? In the baseline experiment, we find that base LLM agents exhibit inconsistent deductive reasoning capabilities across extended interactions, achieving perfect accusation accuracy four times in the baseline experiment but never in the fine-tuned experiment. Moreover, GPT-4o-mini outperforms Gemini-2.5-Flash, winning three of the four games, despite Gemini-2.5-Flash being a newer model proven to achieve higher performance than 4o models Comanici et al. (2025). Our manual review of the game logs further revealed that Gemini-2.5-Flash frequently verbalized the identification of logical inconsistencies throughout gameplay, which may have hindered its ability to win (i.e., its focus on avoiding logical contradictions was not beneficial). In the fine-tuned experiment, both model families and both variants (base and fine-tuned) fail to achieve perfect accusation accuracy in any of the 12 games, compared to four perfect scores in the baseline experiment. These results suggest that extended, multi-turn interactions requiring consistent deductive reasoning at each step remain challenging for LLMs. Additionally, our fine-tuning experiments indicate that training on text-based logic puzzles does not reliably translate to improved in-game reasoning. In future work, we plan to explore alternative approaches to fine-tuning, including tool-augmented or hybrid methods, rather than relying solely on text-based solutions.

RQ2: Does fine-tuning on deductive reasoning tasks improve agentic performance in a related environment? We find that fine-tuning on the Mind Benders logic puzzles produces divergent effects across model families. Gemini-2.5-Flash (FT) shows a marginal improvement over its own base instances within the fine-tuned experiment, but the complete absence of 3/3 scores (compared to four in the baseline) indicates that fine-tuning related logic puzzles did not translate to better deductive reasoning at decision time. Additionally, fine-tuning reduces GPT-4o-mini’s performance. These findings are consistent with prior work investigating the effect of text-based fine-tuning for reasoning improvements Lobo et al. (2025). We also observe that both base models perform worse when fine-tuned agents are introduced into the game, suggesting that fine-tuned players alter the game’s information dynamics in ways that affect all participants. This raises broader questions for future work regarding how agent composition shapes individual performance in multi-agent reasoning environments.

RQ3: How does knowledge accumulation relate to reasoning quality in fine-tuned versus base model variants? Our results suggest that greater information accumulation does not produce better reasoning. GPT-4o-mini (FT) accumulates knowledge fastest and deduces most, yet performs worst, while Gemini-2.5-Flash (FT) makes fewer deductions but demonstrates greater reasoning precision. This disconnect between knowledge accumulation and accusation accuracy highlighted by GPT-4o-mini (FT)’s game play suggests that fine-tuning encouraged engagement with the reasoning task (i.e., more activity and more deductions) without improving the underlying capacity to synthesize that information into a correct conclusion. Future iterations of this framework will investigate interventions against verbose reasoning, and implementations of a structured state tracker.

Taken together, these results suggest that multi-turn deductive reasoning remains a challenging task for LLMs and text-based fine-tuning alone is insufficient for performance improvement. More broadly, our results indicate that reasoning in interactive environments requires not only local inference steps but also stable information integration over time and robustness to multi-agent dynamics. Developing methods that explicitly target these properties may be necessary for advancing agentic reasoning capabilities.

7 CONCLUSION

For large language models to be reliably deployed as autonomous agents, they must exhibit robust and consistent reasoning capabilities. Structured game environments provide a controlled and interpretable setting in which to evaluate these abilities. In this paper, we present a multi-agent Clue framework for evaluating LLM deductive reasoning across extended, stateful interactions. Our experiments compare baseline and fine-tuned variants of GPT-4o-mini and Gemini-2.5-Flash

to characterize their underlying deductive capabilities and assess whether fine-tuning on a related task improves performance. Across 18 games we find that LLMs inconsistently exhibit valid, successful deductive reasoning, text-based fine-tuning does not reliably improve multi-turn deductive reasoning, and there is a disconnect between information accumulation and reasoning quality.

REFERENCES

- Ajay Bandi, Bhavani Kongari, Roshini Naguru, Sahitya Pasnoor, and Sri Vidya Vilipala. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9):404, 2025.
- Yizhou Chi, Lingjun Mao, and Zineng Tang. Amongagents: Evaluating large language models in the interactive text-based social deduction game. In *Proceedings of the 4th Wordplay: When Language meets Games Workshop*. Association for Computational Linguistics, 2024.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1173–1203, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.65. URL <https://aclanthology.org/2024.acl-long.65/>.
- 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.
- Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyang Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of *ipl* diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097.
- 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, 2021. URL <https://arxiv.org/abs/2103.03874>.
- Wenye Lin, Jonathan Roberts, Yunhan Yang, Samuel Albanie, Zongqing Lu, and Kai Han. GAME-BoT: Transparent assessment of LLM reasoning in games. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7656–7682, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.378. URL <https://aclanthology.org/2025.acl-long.378/>.
- Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962, 2023. doi: 10.1109/TASLP.2023.3293046.
- Zichen Liu, Anya Sims, Keyu Duan, Changyu Chen, Simon Yu, Xiangxin Zhou, Haotian Xu, Shaopan Xiong, Bo Liu, Chenmian Tan, Weixun Wang, Hao Zhu, Weiyang Shi, Diyi Yang, Michael Qizhe Shieh, Yee Whye Teh, Wee Sun Lee, and Min Lin. GEM: A gym for generalist LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=vsqQ1lG52a>.
- Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter*

of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 11679–11698, 2025.

Ume Nisa, Muhammad Shirazi, Mohamed Ali Saip, and Muhammad Syafiq Mohd Pozi. Agentic ai: The age of reasoning—a review. *Journal of Automation and Intelligence*, 2025.

OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large language models, November 2024. URL <https://aclanthology.org/2024.emnlp-main.1160/>.

Chris Sypherd and Vaishak Belle. Practical considerations for agentic llm systems. *arXiv preprint arXiv:2412.04093*, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Tamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421/>.

The Critical Thinking Co. Mind benders® series. <https://www.criticalthinking.com/mind-benders.html>, 2026. Accessed: 2026-02-15.

Yanbo Wang, Zixiang Xu, Yue Huang, Xiangqi Wang, Zirui Song, Lang Gao, Chenxi Wang, Xiangu Tang, Yue Zhao, Arman Cohan, Xiangliang Zhang, and Xiuying Chen. Dyflow: Dynamic workflow framework for agentic reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27723–27730, 2025.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game, 2025. URL <https://arxiv.org/abs/2310.18940>.

Zhichao Xu, Zongyu Wu, Yun Zhou, Aosong Feng, Kang Zhou, Sangmin Woo, Kiran Ramnath, Yijun Tian, Xuan Qi, Weikang Qiu, Lin Lee Cheong, and Haibo Ding. Beyond correctness: Rewarding faithful reasoning in retrieval-augmented generation, 2026. URL <https://arxiv.org/abs/2510.13272>.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning, 2020. URL <https://arxiv.org/abs/2002.04326>.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation, 2024. URL <https://arxiv.org/abs/2312.17080>.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. Probing the multi-turn planning capabilities of LLMs via 20 question games. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1495–1516, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. The curse of cot: On the limitations of chain-of-thought in in-context learning, 2025. URL <https://arxiv.org/abs/2504.05081>.

Xueyang Zhou, Guiyao Tie, Guowen Zhang, Weidong Wang, Zhigang Zuo, Di Wu, Duanfeng Chu, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. Exploring the necessity of reasoning in llm-based agent scenarios. *arXiv preprint arXiv:2503.11074*, 2025.

Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. Pokerbench: Training large language models to become professional poker players, Apr. 2025. URL <https://ojs.aaai.org/index.php/AAAI/article/view/34814>.

APPENDIX

A AGENTIC GAME SETUP

We present the pseudocode for our variation of the Clue game with agentic players in Algorithm 1. We highlight the three main reasoning phases in the code, where agents are prompted to (1) make deductions, (2) make a suggestion or accusation, and (3) reveal a card if they have more than one card match to another player’s suggestion.

A.1 GAME IMPLEMENTATION PSEUDOCODE

Algorithm 1 Game Implementation Outline

```

1: function RUNGAME
2:    $(state, dealt) \leftarrow \text{SETUPGAME}(players)$  ▷ Instantiate game
3:   while ACTIVEPLAYERS( $players$ ) do
4:      $p \leftarrow \text{NEXTPLAYER}(players, state)$ 
5:     TURNCOUNTERS( $state$ )
6:      $(deduceResponse, deducePrompt) \leftarrow \text{DEDUCTIONPROMPT}(llm, p, state)$ 
7:      $(deducedCards, deduceReason) \leftarrow \text{PARSEDEDUCTION}(deduceResponse)$ 
8:     UPDATEPLAYERDEDUCTIONS( $p, deducedCards, state$ )
9:      $context_t \leftarrow \text{TURNCONTEXT}(lastReasoning, lastSuggestion, deducedReason)$ 
10:     $(actResponse, actPrompt) \leftarrow \text{TURNPROMPT}(llm, p, state, context_t)$ 
11:     $move \leftarrow \text{PARSEMOVE}(actResponse)$ 
12:    if  $move.type = \text{SUGGESTION}$  then
13:       $s \leftarrow \text{BUILDSUGGESTION}(p, move)$ 
14:       $p_d \leftarrow \text{SELECTDISPROVER}(players, p, s)$ 
15:      if  $p_d \neq \emptyset$  then
16:         $(showResponse, showPrompt) \leftarrow \text{SHOWPROMPT}(llm, p_d, s, state)$ 
17:         $card \leftarrow \text{PARSESHOWNCARD}(showResponse)$ 
18:        REVEALCARD( $p_d, p, card$ )
19:      else
20:        NODISPROOF( $p, s$ )
21:      end if
22:      else if  $move.type = \text{ACCUSATION}$  then
23:         $(accResponse, accPrompt) \leftarrow \text{ACCUSATIONPROMPT}(llm, p, state)$ 
24:         $a \leftarrow \text{PARSEACCUSATION}(accResponse)$ 
25:        RESOLVEACCUSATION( $p, a, state$ )
26:        State  $correct \leftarrow \text{CHECKACCUSATION}(a, state.solution)$ 
27:        if  $correct$  then
28:           $state.winner \leftarrow p$ 
29:        break ▷ game ends
30:      else
31:         $p.eliminated \leftarrow \text{true}$ 
32:      end if
33:      UPDATEMEMORY( $p, actResponse, s$ ) ▷ e.g., lastReasoning/lastSuggestion
34:      ADVANCETURN( $state$ )
35:    end while
36:    return  $state.winner$ 
37: end function

```

A.2 LLM PROMPTS

For each reasoning phase in gameplay we provide descriptive prompts. We provide simplified versions of the prompts for each phase below:

Phase 1: Deduction

DEDUCTION PHASE: Before you take your turn, carefully analyze the game history to deduce what cards other players hold.

Players still in the game: {players}

+ {knowledge}

+ {candidates}

+ {history}

HOW TO DEDUCE:

- When a player disproves a suggestion of {Suspect, Weapon, Room}, they MUST hold at least one of those 3 cards.
- If you already know 2 of the 3 cards in a disproved suggestion are accounted for (in your hand, shown to you, or held by someone else), then the disprover MUST hold the remaining card.
- If a suggestion was NOT disproved by anyone, the cards not held by the suggester (or by you) must be in the envelope. Remember: the suggester may hold some of those cards themselves.
- Cross-reference multiple disprovals by the same player to narrow down their cards.

Unknown cards (not in your hand, seen, or deduced): {unknown_cards}

Respond in this EXACT format (no markdown):

ANALYSIS: <step-by-step reasoning about what each disproval tells you, cross-referencing to narrow down cards>

DEDUCED_CARDS: <comma-separated list of cards you are confident other players hold (NOT in envelope)> or NONE

Phase 2: Turn Play

It is your turn. Choose a suggestion that maximizes information gain and helps you solve quickly.

You can test a deduction, test a combination, or include one of your own cards to narrow other cards.

If you can accuse confidently, do so this turn.

+ {knowledge}

+ {candidates}

+ {observations}

+ {history}

+ {last_suggestion}

+ {deduction}

+ {reasoning}

Phase 3: Show Card

Another player ({suggester_name}) has made a suggestion:

{suggestion.suspect} with {suggestion.weapon} in {suggestion.room}.

You have the following cards that can disprove this suggestion: {cards}

Your show history (cards you have previously shown and to whom):

{card_history}

Choose ONE card to show to {suggester_name}.

Consider which card reveals the least new information

(e.g. a card you have already shown to this player).

Return ONLY:

REASONING: <your reasoning>

SHOW: <exact card name>

B REASONING EXAMPLES

The length of each player's reasoning log is lengthy, so we provide samples of their outputs in Figures 6 and 7. We showcase each model's final-turn winning reasoning.

Turn 5 — Deduction Phase
Deduced cards: Candlestick
Correct: Candlestick
Incorrect: None
Cumulative — Shown: 0, Deduced correct: 1, Deduced incorrect: 0
Raw response:
ANALYSIS:
T1. GEMINI_FLASH_1 suggested Miss Scarlet, Candlestick, Kitchen. This was disproved by GEMINI_FLASH_3. GEMINI_FLASH_3 must hold one of these cards. I hold Miss Scarlet and Kitchen. Since cards are unique, GEMINI_FLASH_3 cannot hold Miss Scarlet or Kitchen. Therefore, GEMINI_FLASH_3 MUST hold Candlestick.
T2. GPT4o_MINI_2 suggested Colonel Mustard, Lead Pipe, Billiard Room. This was disproved by GEMINI_FLASH_3. GEMINI_FLASH_3 must hold one of these cards. I hold Lead Pipe. GEMINI_FLASH_3 cannot hold Lead Pipe. From T1, we know GEMINI_FLASH_3 holds Candlestick. This disproof tells us GEMINI_FLASH_3 holds either Colonel Mustard or Billiard Room (in addition to Candlestick).
T3. GEMINI_FLASH_2 suggested Mrs. White, Knife, Ballroom. This was disproved by GEMINI_FLASH_1. GEMINI_FLASH_1 must hold one of these cards: Mrs. White, Knife, or Ballroom.
T4. GPT4o_MINI_3 suggested Mrs. White, Candlestick, Study. This was disproved by GPT4o_MINI_1. GPT4o_MINI_1 must hold one of these cards. We deduced in T1 that GEMINI_FLASH_3 holds Candlestick. Since cards are unique, GPT4o_MINI_1 cannot hold Candlestick. Therefore, GPT4o_MINI_1 MUST hold either Mrs. White or Study.
DEDUCED_CARDS: Candlestick (held by GEMINI_FLASH_3)

Turn 5
LLM status: RAW response
Knowledge: {"hand": ["Kitchen", "Lead Pipe", "Miss Scarlet"], "seen_cards": ["Conservatory"], "deduced_cards": ["Candlestick"]}
Card shown to GEMINI_FLASH_3: Conservatory
Raw response:
SUMMARY: I hold Kitchen, Lead Pipe, and Miss Scarlet. I know GEMINI_FLASH_3 holds Candlestick, and also either Colonel Mustard or Billiard Room. GEMINI_FLASH_1 holds one of Mrs. White, Knife, or Ballroom. GPT4o_MINI_1 holds one of Mrs. White or Study.
REASONING: My previous deduction correctly identified Candlestick as being held by GEMINI_FLASH_3. For this turn, I want to make a suggestion using cards I don't hold and haven't been disproved. I'll pick a suspect, weapon, and room that are currently unknown to maximize the chance of getting new information about cards held by other players. Mr. Green, Revolver, and Conservatory are good candidates as they haven't been part of any disproven suggestions yet, and I don't hold them.
ROOM: Conservatory
SUGGESTION: Mr. Green, Revolver, Conservatory
ACCUSATION: NONE
Decision: {"room": "Conservatory", "suggestion": "Mr. Green, Revolver, Conservatory", "accusation": null}
- envelope. My last turn confirmed this condition was met with Professor Plum, Rope, and Library.
ROOM: Library
SUGGESTION: Professor Plum, Rope, Library
ACCUSATION: Professor Plum, Rope, Library
Decision: {"room": "Library", "suggestion": "Professor Plum, Rope, Library", "accusation": "Professor Plum, Rope, Library"}

Figure 6: Winning reasoning from Gemini-2.5-Flash player on their first turn.

Turn 50 — Deduction Phase
Deduced cards: None
Correct: None
Incorrect: None
Cumulative — Shown: 5, Deduced correct: 10, Deduced incorrect: 3
Raw response:
ANALYSIS:
1. T40: GEMINI_FLASH_2 suggested Miss Scarlet, Knife, Billiard Room. Disproved by GPT4o_MINI_1. Therefore, GPT4o_MINI_1 must hold either Miss Scarlet, Knife, or Billiard Room.
2. T41: GEMINI_FLASH_3 suggested Mr. Green, Lead Pipe, Billiard Room. Disproved by GPT4o_MINI_1. Since GPT4o_MINI_1 has already disproved two other suggestions, they cannot hold Billiard Room. Thus, GPT4o_MINI_1 must hold either Mr. Green or Lead Pipe.
3. T42: GPT4o_MINI_1 suggested Professor Plum, Wrench, Lounge. Not disproved, meaning at least one of these is in the envelope or held by others (not by you).
4. T43: GEMINI_FLASH_1 suggested Colonel Mustard, Revolver, Kitchen. Disproved by GEMINI_FLASH_3. Therefore, GEMINI_FLASH_3 must hold one of these cards.
5. T44: GEMINI_FLASH_2 suggested Colonel Mustard, Rope, Lounge. Disproved by GEMINI_FLASH_3. Thus, GEMINI_FLASH_3 holds either Colonel Mustard, Rope, or Lounge.
6. T45: GEMINI_FLASH_3 suggested Colonel Mustard, Rope, Lounge. Not disproved, indicating that Colonel Mustard, Rope, or Lounge may indeed be in the envelope or held by others.
7. T46: GPT4o_MINI_1 suggested Mrs. White, Knife, Dining Room. Disproved by GEMINI_FLASH_2. Therefore, GEMINI_FLASH_2 must hold one of these cards.
8. T47: GEMINI_FLASH_1 suggested Miss Scarlet, Wrench, Conservatory. Disproved by GEMINI_FLASH_2, so GEMINI_FLASH_2 holds either Miss Scarlet or Wrench.
9. T48: GEMINI_FLASH_2 suggested Mr. Green, Knife, Hall. Disproved by GEMINI_FLASH_1. GEMINI_FLASH_1 holds either Mr. Green or Knife (but not Hall, since it wasn't disproved).
10. T49: GEMINI_FLASH_3 suggested Miss Scarlet, Rope, Conservatory. Disproved by GPT4o_MINI_1. Thus, GPT4o_MINI_1 holds either Miss Scarlet or Rope, but not Conservatory.
Cross-referencing:
- Since GPT4o_MINI_1 has disproved multiple suggestions and holds cards that overlap, it can be deduced that GPT4o_MINI_1 holds either Mr. Green or Lead Pipe, and potentially Wrench.
- GEMINI_FLASH_3 must hold Colonel Mustard or Lounge based on disprovals.
- GEMINI_FLASH_2 has been shown to hold either Miss Scarlet or the Wrench.
DEDUCED_CARDS: Mr. Green, Colonel Mustard, Lounge, Wrench

Turn 50
LLM status: RAW response
Knowledge: {"hand": ["Billiard Room", "Kitchen", "Miss Scarlet"], "seen_cards": ["Conservatory", "Dining Room", "Knife", "Mrs. White", "Revolver"], "deduced_cards": ["Ballroom", "Candlestick", "Colonel Mustard", "Hall", "Lead Pipe", "Library", "Lounge", "Mr. Green", "Mrs. Peacock", "Professor Plum", "Rope", "Study", "Wrench"]}
Card shown to player: None
Raw response:
SUMMARY: I have deduced that Professor Plum, Wrench, and Lounge must all be in the envelope since my last suggestion was not disproved, and I do not hold any of those cards.
REASONING:
1. My last suggestion was Professor Plum, Wrench, Lounge, and no one disproved it, indicating that none of those cards are held by any player, including myself.
2. Therefore, all three cards must be in the envelope.
3. This allows me to confidently make an accusation.
ROOM: Lounge
SUGGESTION: Professor Plum, Wrench, Lounge
ACCUSATION: Professor Plum, Wrench, Lounge

Figure 7: Winning reasoning from GPT-4o-Mini player on their ninth turn.