InMind: Evaluating LLMs in Capturing and Applying Individual Human Reasoning Styles

Anonymous ACL submission

Abstract

LLMs have shown strong performance on human-centric reasoning tasks. While previous evaluations have explored whether LLMs can infer intentions or detect deception, they often overlook the individualized reasoning styles that influence how people interpret and act in social contexts. Social deduction games (SDGs) provide a natural testbed for evaluating individualized reasoning styles, where different players may adopt diverse but contextually valid reasoning strategies under identical conditions. To address this, we introduce **InMind**, a cognitively grounded evaluation framework designed to assess whether LLMs can capture and apply personalized reasoning styles in SDGs. InMind enhances structured gameplay data with round-level strategy traces and post-game reflections, collected under both Observer and Participant modes. It supports four cognitively motivated tasks that jointly evaluate both static alignment and dynamic adaptation. As a case study, we apply InMind to the game *Avalon*, evaluating 11 state-of-the-art LLMs. General-purpose LLMs, even GPT-4o frequently rely on lexical cues, struggling to anchor reflections in temporal gameplay or adapt to evolving strategies. In contrast, reasoning-enhanced LLMs like DeepSeek-R1 exhibit early signs of style-sensitive reasoning. These findings reveal key limitations in current LLMs' capacity for individualized, adaptive reasoning, and position InMind as a step toward cognitively aligned human–AI interaction.

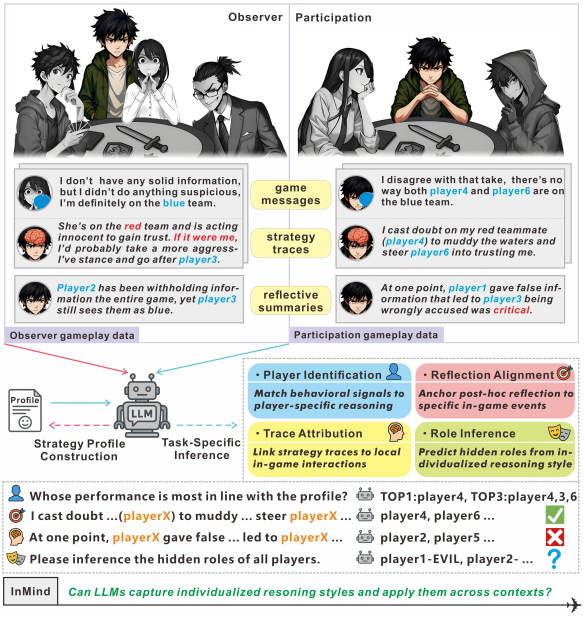

Figure 1: **Overview of the InMind Framework.** The system constructs a subject-specific strategy profile from observer-mode data and applies it to participant-mode gameplay, supported by dual-layer annotations. Four cognitively motivated tasks assess the model's ability to apply individualized reasoning styles.

1 Introduction

Recent large language models (LLMs), such as DeepSeek-R1 (DeepSeek-AI et al., 2025) and O1 (OpenAI et al., 2024b), have demonstrated strong reasoning abilities across complex mathematical and scientific domains (Chen et al., 2025). Emerging research (Mori et al., 2025; Park et al., 2024) further highlights their promising performance in human-centric tasks, including social commonsense inference, intention recognition, and belief attribution. Beyond these capabilities, recent studies suggest that LLMs may exhibit early signs of *Theory of Mind* (ToM)—the ability to represent and reason about others' beliefs, desires, and intentions (Sarıtaş et al., 2025; Kim et al., 2025). Understanding and evaluating such high-level cognitive traits is critical for advancing LLMs toward artificial general intelligence (AGI), and potentially, artificial superintelligence.

Existing benchmarks attempt to assess ToM-like reasoning through tasks such as intent classification (Liu et al., 2024), false-belief attribution (Huang, 2024), and multiple-choice social inference (Seo et al., 2024b). However, these methods primarily target output plausibility or behav-

ioral consistency, offering limited insight into underlying cognitive mechanisms, especially those that vary across individuals. In practice, different people often exhibit context-sensitive preferences in subjective scenarios and may arrive at similar conclusions via distinct reasoning trajectories (Gagnon-St-Pierre and Doucerain, 2021b,a). We refer to this as an *individualized reasoning style*.

Social deduction games (SDGs) become an ideal evaluation scenario for internalizing and applying reasoning styles, where players must infer the hidden mental states of others and make strategic decisions accordingly (Zhang et al., 2025; Yoo and Kim, 2024). Due to their dynamic, adversarial and individualized nature (Feng et al., 2024), such settings require more than surface-level alignment: if an LLM cannot capture and adapt to a player's individualized reasoning style, even plausible output may not support meaningful collaboration. Bridging this gap is essential for advancing ToM-inspired modeling of individual variation in reasoning, and for building LLMs capable of personalized, adaptive inference. We identify two key challenges: (1) how to capture and represent individualized reasoning processes, which may require structured interaction settings and cognitively meaningful annotations; (2) how to evaluate whether an LLM can apply a learned reasoning style in contextually adaptive ways, which calls for fine-grained, cognitively grounded tasks.

To meet these challenges, we propose **InMind**, a cognitively grounded evaluation framework designed to assess whether LLMs can internalize and apply individualized reasoning styles through SDGs. As illustrated in Figure 1, InMind introduces two complementary gameplay modes: *Observer*, where a subject reasons passively from the perspective of another player without acting, and *Participant*, where the subject actively engages in gameplay from their own perspective. This setup not only supports the natural capture of individualized reasoning, but also enables its application and evaluation in dynamic, interactive contexts.

Crucially, InMind integrates dual-layer cognitive annotations: (1) *strategy traces*, which capture real-time reasoning signals such as belief updates, intention inference, and counterfactual thinking; and (2) *reflective summaries*, offering post-hoc insights that contextualize key game events and assess the behaviors and intentions of other players. Leveraging these signals, InMind defines four cognitively motivated tasks to evaluate distinct aspects of individualized reasoning. (1) *Player Identification* tests whether a model can recognize behavioral patterns that align with a specific reasoning style. (2) *Reflection Alignment* assesses the model's ability to ground abstract post-game reflections in concrete game behavior. (3) *Trace Attribution* probes whether the model can simulate evolving, in-context reasoning across time. (4) *Role Inference* evaluates whether the model can internalize reasoning styles to support belief modeling under uncertainty.

To concretely investigate these capabilities, we instantiate InMind within the popular social deduction game Avalon[1], creating **InMind-Avalon**, a novel dataset comprising 30 full-session human gameplays annotated with detailed cognitive traces and reflective summaries. Our empirical analysis evaluates 11 state-of-the-art LLMs on Avalon-InMind and highlights several critical limitations: (1) Most models, including GPT-4o, heavily rely on superficial lexical patterns, failing to consistently infer deeper strategic intent; (2) Temporal alignment between reflective reasoning and specific in-game events remains challenging for nearly all evaluated models; (3) Dynamic adaptation of strategic reasoning based on evolving interactions is largely insufficient, indicating fundamental shortcomings in models' capability for individualized reasoning. Nevertheless, we observe promising potential in certain models, such as DeepSeek-R1, suggesting possible avenues for improvement. Despite the inherent subjectivity in individualized annotations, these cognitively grounded traces and reflections effectively facilitate fine-grained tasks like hidden role identification, highlighting their practical utility for model training and evaluation.

In summary, our contributions are threefold: (1) We introduce **InMind**, a cognitively grounded evaluation framework specifically designed to assess individualized reasoning and strategic adaptation of LLMs in dynamic social deduction scenarios; (2) We release **InMind-Avalon**, a novel annotated dataset comprising 30 full-session human gameplay recordings enhanced with detailed cognitive annotations, including real-time strategy traces and reflective summaries; (3)Through extensive evaluation of current state-of-the-art models, we identify critical limitations in temporally structured reasoning, adaptive strategy use, and individualized alignment.

[1] https://avalon-game.com/wiki/rules/

We hope that InMind serves as a principled tool to guide future advances toward individualized, adaptive collaboration between humans and AI in socially rich, interactive environments.

2 Related Work

2.1 Theory of Mind Reasoning in LLMs

Recent research has shown that large language models (LLMs) increasingly demonstrate capabilities aligned with Theory of Mind (ToM), including false-belief attribution, intention recognition, and motivational reasoning (Kim et al., 2025; Sarıtaş et al., 2025; Park et al., 2024), which suggest that LLMs can approximate certain aspects of social cognition. Several benchmarks have been introduced to evaluate these faculties, such as *Social IQa* (Sap et al., 2019), *KoCommonGEN* (Seo et al., 2024a), and *OpenToM* (Xu et al., 2024), which typically use multiple-choice or context-driven tasks. More recent approaches incorporate dialog-based (Yu et al., 2025) and reinforcement learning settings (Lu et al., 2025) to explore deeper social reasoning. However, most benchmarks focus on output plausibility and offer limited visibility into the reasoning process itself. By contrast, the proposed InMind framework builds on these discussions and explicitly identifies two core limitations: the lack of temporal structure in evaluating reasoning over time, and the failure to distinguish between surface behavior and underlying cognition.

2.2 Cognitive and Strategic Modeling in SDGs

Social deduction games (SDGs), such as *Avalon*, *Werewolf*, and *Among Us*, provide a dynamic and adversarial context for evaluating the strategic reasoning capabilities of LLMs (Feng et al., 2024; Yoo and Kim, 2024; Wu et al., 2024a). While several studies have utilized such environments to assess LLMs' performance in role identification, belief tracking, and deception detection (Light et al., 2023; Zhang et al., 2025; Chi et al., 2024), most of them fail to provide structured representations of the cognitive processes involved in gameplay. Although some efforts (Stepputtis et al., 2023; Liu et al., 2024) introduce temporal and intention-aware evaluations, they still fall short in providing comprehensive or individualized annotations that capture the evolving nature of reasoning within social deduction contexts. In contrast, the InMind framework introduces cognitively annotated interactions across distinct gameplay modes, enabling the capture of individualized reasoning styles and supporting the evaluation of LLMs' adaptive capabilities in dynamic social settings.

3 The InMind Framework

We introduce InMind, a cognitively grounded framework for evaluating whether LLMs can internalize and apply individualized human reasoning styles. The framework is built on three key components: (1) Structured Game Representation. InMind encodes gameplay using a structured representation that supports dual-perspective modeling and captures fine-grained cognitive annotations of each decision point, allowing for nuanced interpretation of reasoning behavior. (2) Evaluation Protocol. InMind defines a protocol of four fine-grained evaluation tasks, designed to test both static alignment with human reasoning profiles and dynamic adaptability across varied gameplay contexts. (3) InMind-Avalon: A Case Study. To demonstrate the framework in practice, we instantiate InMind in a case study based on the Avalon social deduction game, showcasing how the framework reveals personalized cognitive patterns and model behavior in complex, multi-agent settings.

3.1 Structured Game Representation

We represent each annotated game session as a structured tuple:

$$\mathcal{G} = \langle \texttt{mode},\, \mathcal{A},\, \{E_z\}_{z\in\mathcal{Z}},\, \mathcal{F}\rangle. \qquad (1)$$

Here, $\texttt{mode} \in \{\text{Observer}, \text{Participant}\}$ denotes the cognitive perspective under which the session was recorded. The role assignment is defined as $\mathcal{A} = (p_1, r_1), ..., (p_n, r_n)$, where each player p_i is assigned a hidden role r_i. The game unfolds over rounds $\mathcal{Z} = z_1, ..., z_m$, with each round z represented by an event tuple $E_z = \langle U_z, G_z, S_z\rangle$. Moreover, U_z contains all player utterances and system messages; G_z records the observable game state (*e.g.*, team proposals, votes, mission outcomes); and S_z captures the subject's real-time strategy trace, including their beliefs, intentions, and inferences. Upon game completion, the subject provides a reflective summary $\mathcal{F}$ that articulates post-hoc reasoning, identifies pivotal moments, and evaluates other players' behavior. Together, the strategy trace S_z and the reflective summary $\mathcal{F}$ constitute the dual-layer cognitive supervision central to InMind.

To simplify notation, we use superscripts ^{o} and ^{p} to indicate whether a variable comes from an

Observer-mode or *Participant-mode* session. Both modes share the same data structure, the key distinction lies in the subject's perspective. In *Observer-mode*, the subject does not take any actions but instead reasons from the perspective of a designated player. This setup helps the model disentangle strategic reasoning from observed surface behavior. In *Participant-mode*, the subject actively engages in the game and provides annotations from their own point of view.

3.2 Evaluation Protocol

The evaluation protocol of InMind consists of two stages: Capturing Individual Reasoning Styles and Applying Profile in Adaptive Tasks, as depicted in Algorithm 1. In Stage 1, it constructs a subject-specific reasoning profile to capture individual cognitive tendencies from *Observer-mode* gameplay, independent of overt behavior. In Stage 2, this profile is applied to a set of downstream reasoning tasks to evaluate the model's ability to simulate and adapt to the subject's decision-making style.

3.2.1 Capturing Individual Reasoning Styles

The goal of this stage is to derive a concise yet expressive profile $\mathcal{S}$ that captures the subject's unique reasoning tendencies. Rather than relying on explicit in-game actions, this process draws from *Observer-mode* gameplay, in which the subject reasons aloud from the perspective of a designated player without taking any actions themselves. This design helps isolate cognitive patterns from behavioral noise, allowing for a more faithful reconstruction of individual reasoning styles.

To construct the profile, we apply a structured prompt (`ProfilePrompt`) over the subject's full Observer-mode session $\mathcal{G}^o$. The prompt instructs the model to identify recurring interpretive strategies, decision heuristics, and evaluative criteria based on the subject's commentary. The output is a free-form textual summary $\mathcal{S}$ that encapsulates how the subject tends to perceive, process, and respond to gameplay dynamics. This profile serves as a static cognitive signature, reused across all downstream evaluations to assess how well models internalize and adapt to human reasoning.

3.2.2 Applying Profile in Adaptive Tasks

In this stage, we assess whether a language model can leverage the reasoning profile $\mathcal{S}$, derived from Observer-mode data, to simulate the subject's reasoning in novel gameplay contexts. Given the profile $\mathcal{S}$ and a new Participant-mode session $\mathcal{G}^p$, the model is prompted to generate inferences that align with the subject's cognitive style.

Algorithm 1: InMind Evaluation Protocol

Input: $\mathcal{G}^o$, $\mathcal{G}^p$, Tasks $\mathcal{T}$
Output: Results $\mathcal{R}$

// Stage 1: Build subject-specific strategy profile
$\mathcal{S} \leftarrow \texttt{LLM}(\texttt{ProfilePrompt}(\mathcal{G}^o))$
// Stage 2: Apply profile to each task
foreach $T \in \mathcal{T}$ **do**
  // Construct task-specific context and targets
  $(C, Y) \leftarrow \texttt{ConstructTask}(\mathcal{G}^p, \mathcal{S}, T)$
  // Generate prediction via formatted prompt
  $\hat{Y} \leftarrow \texttt{LLM}(\texttt{FormatPrompt}(C, T))$
  // Compute evaluation result
  $\mathcal{R} \leftarrow \texttt{Eval}(\hat{Y}, Y)$

InMind transforms naturalistic annotations into structured prediction tasks by capitalizing on the inherently player-centric structure of social deduction games (SDGs). Since reasoning in SDGs is organized around individual players and their interactions, most annotations are naturally linked to specific player IDs (*e.g.*, P1–P6). This structure allows cognitive traces and post-hoc reflections to be reformulated as clearly defined prediction problems. To this end, we introduce four such tasks, including player identification, reflection alignment, trace attribution, and role inference. All tasks are formulated as predictions over player identities or roles. Their configurations are summarized in Table 1.

(1) Player Identification. Given a subject-specific strategy profile $\mathcal{S}$ and a participant-mode session $\mathcal{G}^p$ (with all player identities anonymized), the model is tasked with identifying which player's in-game behavior best aligns with the subject's reasoning style. It produces a ranked list of candidates, and evaluation is based on Top-k accuracy (typically $k = 1, 3$). This task measures static cognitive alignment between the profile and observed gameplay. Inputs include all player utterances and strategy traces, but exclude role assignments and game outcomes.

(2) Reflection Alignment. This task evaluates whether LLMs can ground high-level post-game reflections in concrete gameplay behavior. Each reflection $\mathcal{F}$, written by the subject, primarily comprises two forms of reasoning: (1) recalling a spe-

Task	Profile $\mathcal{S}$	Game Msgs	Strategy Trace S_z	Reflective Summary$\mathcal{F}$	Target	Temporal Mode
Player Identification	✓	✓	✓	–	Player ranking	Static
Reflection Alignment	✓	✓	△	✓†	Player IDs	Static
Trace Attribution	✓	✓	✓†	–	Player IDs	Dynamic
Role Inference	✓	✓	△	–	Role assignment	Dynamic

Table 1: Summary of task configurations in **InMind**. All tasks use the same subject-specific profile $\mathcal{S}$ and participant-mode gameplay as input. ✓ indicates components visible to the model; ✓† marks fields where player identities are masked as prediction targets; △ denotes optional inputs. Static tasks are presented as full-game contexts. Dynamic tasks reveal gameplay incrementally and require the model to summarize prior context, simulating human reasoning under limited memory and partial recall.

cific moment that affected the game's trajectory (e.g., *"Player 1 deceived Player 5 about mission sabotage"*), and (2) offering abstract evaluations of others based on global impressions (e.g., *"Player 3 never voiced doubt, probably hiding something"*, which compresses multi-round silence into a single attribution). These reflections are typically temporally unanchored and do not reference specific rounds, posing challenges for grounding. We mask player IDs mentioned in $\mathcal{F}$ and prompt the model to fill them in based on the full session $\mathcal{G}^p$ and the subject profile $\mathcal{S}$. Performance is evaluated using exact-match accuracy, where a prediction is considered correct only if all masked player IDs in $\mathcal{F}$ are accurately recovered.

(3) Trace Attribution. This task assesses whether LLMs can simulate a subject's evolving reasoning trajectory across rounds of gameplay. Each round-level trace S_z reflects how the subject interprets recent events, including factual observations (e.g., *"P3's last statement came across as overly eager"*), identity attributions (e.g., *"I suspect P2 is Evil"*), and intended next moves (e.g., *"I'll accuse P4 next to test their reaction"*). To evaluate this capability, we mask all player IDs in S_z and prompt the model to recover them incrementally(one round at a time), using only prior game context and the subject profile $\mathcal{S}$. Accuracy is measured by exact match. Unlike Task 2, this task requires real-time attribution and adjustment, testing whether models can follow the subject's strategy as it unfolds.

(4) Role Inference. This task assesses whether LLMs can extend the subject's reasoning style to perform belief modeling under uncertainty. As the game unfolds, the model must incrementally infer each player's hidden role based on partial observations and evolving interactions. At each round, it receives the current game history and outputs a full player-to-role mapping. Scoring is weighted toward later rounds to reflect the increasing availability of evidence. The task evaluates whether models can maintain consistency, adapt to new information, and infer the hidden roles

3.3 InMind-Avalon: A Case Study

The InMind framework is instantiated in the six-player version of the social deduction game *Avalon*, characterized by asymmetric hidden roles and the need for collaborative reasoning under uncertainty. The Good team comprises Merlin, Percival, and two Loyal Servants; the Evil team consists of Morgana and the Assassin. For detailed rules, please refer to Appendix A.

We recruited 73 experienced players, one of whom was randomly selected to serve as the subject. This player completed both observer-mode and participant-mode sessions, while other players were resampled per game to ensure strategic diversity. We then conducted all sessions via online voice chat in Mandarin Chinese to preserve authentic communication dynamics. Players frequently used game-specific expressions such as *tiao pai*, *dui tiao*, and *chong piao*, which introduce additional reasoning challenges due to their implicit and context-dependent meanings. We transcribed all speech and comments verbatim, including disfluencies (*e.g.*, pauses, hesitations), in order to preserve the real-time dynamics and interaction patterns of gameplay. To ensure annotation quality, three expert annotators accompanied the subject throughout and provided real-time guidance on producing round-level *strategy traces* and post-game *reflective summaries*. All annotations were reviewed for consistency. See Appendix B and H for details.

The resulting dataset is referred to as InMind-Avalon, which comprises 30 full game sessions (25 participant-mode, 5 observer-mode), totaling 884 player turns, 160 strategy traces, and 30 reflective

summaries. Each session lasts 20–25 minutes, with total gameplay exceeding 10 hours. Players are limited to 30 seconds per turn, yielding concise but tactically dense utterances. All canonical roles are well represented across varied team compositions. Notably, 22 games reached the final assassination phase, with Merlin correctly identified in only 41% of cases, highlighting the difficulty of role inference. The subject's team achieved a 56% win rate in participant-mode games. Additional statistics are provided in Appendix C.

Based on InMind-Avalon, we construct four structured evaluation tasks following the two-stage InMind protocol (Section 3.2): (1) 25 player identification cases, (2) 194 reflection alignment instances, (3) 791 trace attribution queries, and (4) 25 role inference sessions across 93 incremental rounds.

4 Experiments

All experiments are conducted on the InMind-Avalon dataset (Section 3.3), following the two-stage protocol defined in Section 3.2. In Section 4.1, we analyze the generated strategy profiles from Stage 1. Sections 4.2 to 4.5 assess how effectively these profiles support downstream tasks, encompassing both static and dynamic reasoning. We evaluate 11 large language models (LLMs) under zero-shot settings. Detailed model specifications are provided in Appendix D, and the unified prompt format is described in Appendix E.

4.1 Strategy Profile Analysis

Each model begins by constructing a subject-specific strategy profile $\mathcal{S}$. We observe clear variation in profile quality and structure across models. GLM-4-9B (GLM et al., 2024) typically produces vague personality summaries (*e.g.*, describing the player as *"logical and objective"* or *"attentive to interpersonal interactions"*), while coherent, are generic and weakly grounded in gameplay. In contrast, DeepSeek-R1 (DeepSeek-AI et al., 2025) generates multi-dimensional profiles that capture reasoning style, discourse tendencies, and adaptive strategies. One such profile characterizes the subject as an *"analytical assassin"*, who intentionally conceals analytical acuity, strategically employs probing questions to extract information, and even adopts Morgana's perspective in Task 4 to infer how Percival was ultimately exposed. This suggests that DeepSeek-R1 is able to extract abstract reasoning traits from observed strategic traces, going beyond surface-level linguistic cues. A complete example is provided in Appendix F.

Model	Top-1 Acc. ↑	Top-3 Acc. ↑	BERT Match ↓
BERT Baseline	0.160	0.480	–
Qwen2.5-7B	0.168	0.416	0.200
Qwen2.5-14B	0.168	0.496	0.208
Qwen2.5-72B	0.208	0.544	0.272
Yi1.5-9B	0.184	0.432	0.200
Yi1.5-34B	0.104	0.456	0.160
GLM4-9B	0.136	0.416	0.280
InternLM2.5-20B	0.160	0.504	0.240
GPT-4o	0.160	**0.672**	0.272
DeepSeek-R1	**0.240**	0.616	**0.144**
QwQ	0.176	0.544	0.144
O3-mini	0.200	0.576	0.288

Table 2: Player identification accuracy.

4.2 Player Identification

Evaluated Models Specifically, we select five general-purpose models varies in parameters (7B to 72B): Qwen2.5 (Qwen et al., 2025), Yi1.5 (AI et al., 2025), GLM4 (GLM et al., 2024), InternLM2.5 (Wu et al., 2024b) and GPT-4o (OpenAI et al., 2024a), and three reasoning-enhanced models: DeepSeek-R1 (DeepSeek-AI et al., 2025), QwQ (Qwen Team, 2024) and O3-mini (OpenAI, 2024) for evaluation. Furthermore, we introduce a baseline model ("BERT Baseline" in Table 2) for comparison, which does not use $\mathcal{S}$, but instead ranks candidates based on the cosine similarity between average-pooled StructBERT embeddings (Wang et al., 2019) of U_z^p and S_z^p.

Setup Each evaluated model receives $\mathcal{S}$, along with all player utterances U_z^p and strategy traces S_z^p , with roles and player identities withheld, to output a ranked list of candidates whose behavior best matches the subject's reasoning profile. We report both Top-1 and Top-3 accuracy, along with BERT Match—the proportion of model predictions that align with the top-ranked candidate from the baseline.

Results and Analysis Table 2 shows that overall Top-1 accuracy remains low across models, with most scores well below 0.20. Even Top-3 accuracy hovers around 0.5, close to chance level in a six-player setting. This highlights the inherent difficulty of identifying individualized reasoning styles from observable behavior, despite access to full gameplay data. Figure 2 further reveals the distinction between surface mimicry and deeper strategic

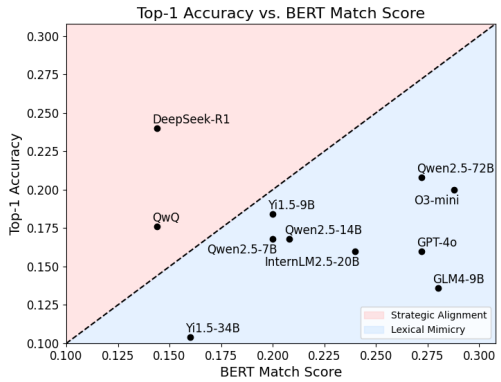

Figure 2: Points above indicate stronger alignment with reasoning profiles beyond lexical similarity, while points below reflect greater similarity to surface-level patterns.

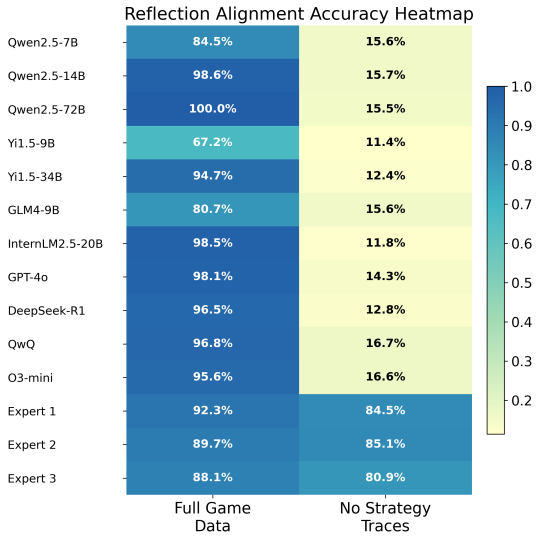

Figure 3: Accuracy of player IDs prediction under two conditions: *Full Game Data* (with access to strategy traces) and *No Strategy Traces* (without traces). Results are shown for all models and three human experts.

alignment. Most models fall near or below the diagonal, indicating reliance on lexical similarity. DeepSeek-R1 stands out by achieving the highest Top-1 score (0.240) while maintaining the lowest BERT Match (0.144), suggesting more abstract reasoning-based alignment.

4.3 Reflection Alignment

Setup Each evaluated model is given the subject's profile $\mathcal{S}$, a participant-mode session $\mathcal{G}^p$ with masked player IDs in the reflection $\mathcal{F}$, and is asked to recover them. We consider two evaluation settings: *Full Game Data*, which incorporates the strategy traces S_z^p, and *No Strategy Traces*, where these traces are omitted. Accuracy is measured by exact match. Human expert performance is reported under the same conditions for reference.

Results and Analysis As shown in Figure 3, models perform well when strategy traces are provided. Each trace is inherently linked to a specific round of gameplay, serving as a temporal anchor that ties reasoning to particular events. In this setting, the task effectively becomes a summarization of already structured signals. In contrast, when traces are withheld, accuracy drops sharply. Without clear temporal references, models struggle to associate abstract reflections with the appropriate moments and players in the game. This suggests a limited capacity for retrospective and context-aware reasoning, which the task is designed to evaluate. Human experts maintain high accuracy in both conditions, likely drawing on experience to reconstruct context even without explicit anchoring.

4.4 Trace Attribution

Setup At each round z_i, the model is given the subject profile $\mathcal{S}$ and all prior gameplay data $\mathcal{G}^p$ up to that point. It receives the current round's utterances, game state, and a strategy trace S_{z_i} with masked player IDs, and must predict the correct identifiers. The task proceeds incrementally, but each prediction is made independently, requiring the model to integrate and summarize all preceding context at each step. In the *+Prior Trace* setting, the trace from the previous round $S_{z_{i-1}}$ is also provided as input. Evaluation is based on exact-match accuracy mentiond in Section 3.2.2.

Model	Base Accuracy ↑	+ Prior Trace ↑	Impact Δ
Qwen2.5-7B	0.254	0.245	-0.009
Qwen2.5-14B	0.397	0.365	-0.032
Qwen2.5-72B	0.444	0.440	-0.004
Yi1.5-9B	0.206	0.197	-0.009
Yi1.5-34B	0.204	0.169	-0.035
GLM4-9B	0.241	0.224	-0.017
InternLM2.5-20B	0.226	0.215	-0.011
GPT-4o	0.440	0.448	+0.008
DeepSeek-R1	**0.503**	**0.517**	**+0.014**
QwQ	0.437	0.454	+0.017
O3-mini	0.268	0.281	+0.013

Table 3: Trace attribution accuracy with and without access to the prior round's strategy trace. The final column reports the performance impact (Δ), where positive values indicate successful adaptation to evolving context.

Results and Analysis As shown in Table 3 and Figure 8(see Appendix G.1), most models show little to no benefit from accessing the previous trace $S_{z_{i-1}}$, and some even decline in accuracy. This indicates difficulty in leveraging prior reasoning to inform current predictions. Rather than building on evolving beliefs, models tend to treat each round as an isolated instance, reflecting limited integration of temporal strategy context. These results highlight a core limitation in dynamic attribution: current LLMs struggle to track and reproduce individualized reasoning styles over time, making it difficult to maintain coherent, round-by-round inference.

4.5 Role Inference

Setup We task the model with inferring the hidden roles of all players from the perspective of a designated subject, based on observed gameplay. To examine how contextual cues influence role inference, we vary three factors: (1) whether the prompt adopts a first-person or third-person perspective, (2) whether the subject's round-level strategy traces (S_z) are provided, and (3) whether the subject's own role is revealed. These conditions define four prompting modes (A–D), summarized in Table 4.

Each setting is evaluated under two criteria: strict scoring and relaxed scoring. The former requires the model to accurately identify all five canonical roles, while the latter simplifies the task by grouping roles into three broader categories: *Informed Good* (Merlin, Percival), *Uninformed Good* (Loyalists), and *Evil* (Morgana, Assassin).

Mode	Perspective	Trace Access	Role Known
A	First-person	✓	✓
B	First-person	✗	✓
C	Third-person	✗	✓
D	Third-person	✗	✗

Table 4: Prompting modes for role inference. All configurations are tested under both strict (exact role) and relaxed (role group) scoring.

Results and Analysis Figure 4 and Figure 9(See G.2) show that providing access to strategy traces S_z improves model performance across all prompting configurations. Mode A achieves the highest scores under both strict and relaxed conditions, indicating that subjective annotations, even when potentially biased or incomplete, can support more effective role inference. In contrast, shifting the prompt perspective from first-person (Mode B) to third-person (Mode C) yields similar results. This suggests that LLMs do not show a clear benefit from observer-style prompting, and we do not observe the "outsider sees more of the game" phenomenon commonly associated with human reasoning. While strict role identification remains challenging for most models, their performance under relaxed scoring demonstrates emerging potential for collaborative inference in social reasoning settings.

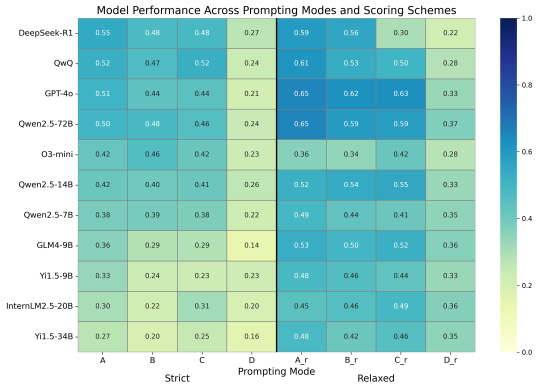

Figure 4: Each cell shows average accuracy under one of four prompting modes (A–D), evaluated using both strict (exact role) and relaxed (role group) criteria. Stronger models appear toward the top.

We also investigate how the subject's role (e.g., Merlin or Percival) shapes the demands placed on reasoning models, as different roles require distinct inference strategies. This supplementary analysis, presented in Appendix G.3, helps assess whether models can adapt to varying cognitive perspectives.

5 Conclusion

InMind offers a novel, cognitively grounded framework for evaluating whether large language models can internalize and apply individualized reasoning styles in complex, interactive settings. By leveraging the structured dynamics of social deduction games and enriching gameplay data with dual-layer annotations from multiple perspectives, InMind enables fine-grained assessment of strategic and adaptive reasoning. Our accompanying dataset, InMind-Avalon, and comprehensive model evaluation reveal key limitations in current LLMs' adapting to individualized human reasoning styles. We believe this framework paves the way for more personalized, socially aware AI systems and invites further exploration into cognitive modeling and adaptive collaboration.

Limitations

While InMind is designed as a general framework for evaluating individualized reasoning in social deduction games, the current implementation focuses exclusively on the Avalon setting. Although Avalon captures many of the strategic and cognitive elements characteristic of SDGs, our experiments do not yet explore other game environments with different social structures or interaction patterns. In future work, we plan to expand both the range of tasks and the scale of the dataset to include additional games such as *Blood on the Clocktower*, *The Resistance* and *Werewolf*, thereby enriching the diversity of reasoning styles and dynamics.

Given the inherently subjective nature of individualized reasoning, the annotation process unavoidably reflects annotator preferences and interpretation. Although expert annotators were involved throughout to guide and standardize the process, variation and bias are difficult to eliminate entirely. Expanding the dataset in both size and diversity will help mitigate such subjectivity and improve the robustness and generalizability of InMind.

Moreover, the core ideas behind InMind extend beyond games and hold potential for broader domains such as multi-agent collaboration, negotiation, and human-AI teaming, where personalized, context-sensitive reasoning is essential. Supporting real-time, multi-agent dynamics will be an important step toward evaluating and enhancing LLMs in more complex, socially situated environments.

The gameplay data used in this study was collected and provided by a collaborating author, who conducted the sessions in accordance with institutional ethical guidelines. All participants were informed about the nature and purpose of the data collection and gave their consent prior to participation. They were also provided with opportunities to ask questions and seek clarification, ensuring informed and voluntary participation. No personally identifiable information was included in the dataset, and all data is used strictly for academic research purposes in compliance with relevant data protection regulations.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, and 14 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Qianlong Chen, Lin Qin, Jiawei Liu, Dong Peng, Jun Guan, and Peng Wang. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. 2024. A survey on large language model-based social agents in game-theoretic scenarios. *Preprint*, arXiv:2412.03920.

Émilie Gagnon-St-Pierre and Marina M. Doucerain. 2021a. A dual-strategy model of social reasoning: Exploring individual differences in strategy selection and outcomes. *Cognitive Science*, 45(4):e12947.

Émilie Gagnon-St-Pierre and Marina M. Doucerain. 2021b. Reasoning strategies explain individual differences in social reasoning. *Journal of Experimental Psychology: General*, 150(6):1078–1095.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Xuanqiang Huang. 2024. Theory of mind in large language models. Master's thesis, University of Bologna.

H. Kim, M. Sclar, T. Zhi-Xuan, L. Ying, and S. Levine. 2025. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881*.

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

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

Florida, USA. Association for Computational Linguistics.

Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. 2025. Do theory of mind benchmarks need explicit human-like reasoning in language models? *Preprint*, arXiv:2504.01698.

Eri Mori, Yuxuan Qiu, Hirokatsu Kataoka, and Yuki Aoki. 2025. A comprehensive analysis of a social intelligence dataset and response tendencies between large language models (llms) and humans. *Sensors*, 25(2):477.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI. 2024. Introducing o_3 and o_4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-05-20.

PS Park, S Goldstein, A O'Gara, M Chen, and D Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100873.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2024. Qwen/QwQ-32B. https://huggingface.co/Qwen/QwQ-32B. Accessed: 2025-05-20.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

K. Sarıtaş, K. Tezören, and Y. Durmazkeser. 2025. A systematic review on the evaluation of large language models in theory of mind tasks. *arXiv preprint arXiv:2502.08796*.

Jaehyung Seo, Jaewook Lee, Chanjun Park, SeongTae Hong, Seungjun Lee, and Heuiseok Lim. 2024a. KoCommonGEN v2: A benchmark for navigating Korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2390–2415, Bangkok, Thailand. Association for Computational Linguistics.

Jinhyeok Seo, Jaesung Lee, Chan Park, Seungtae Hong, and Sungwon Lee. 2024b. Kocommongen v2: A benchmark for navigating korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Simon Stepputtis, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. 2023. 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*, pages 11193–11208, Singapore. Association for Computational Linguistics.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *Preprint*, arXiv:1908.04577.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024a. Enhance reasoning for large language models in the game werewolf. *Preprint*, arXiv:2402.02330.

Zijian Wu, Suozhi Huang, Zhejian Zhou, Huaiyuan Ying, Jiayu Wang, Dahua Lin, and Kai Chen. 2024b. Internlm2.5-stepprover: Advancing automated theorem proving via expert iteration on large-scale lean problems. *Preprint*, arXiv:2410.15700.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *Preprint*, arXiv:2402.06044.

Seonghyeon Yoo and Kyunghyun Kim. 2024. Finding deceivers in social context with large language models: the case of the mafia game. *Scientific Reports*, 14(1):1519.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. 2025. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *Preprint*, arXiv:2502.21017.

Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. 2025. Multimind: Enhancing werewolf agents with multimodal reasoning and theory of mind. *Preprint*, arXiv:2504.18039.

A Game Setting and Terminology

A.1 Avalon Setup and Roles

In the InMind-Avalon dataset, all sessions follow the 6-player variant of *Avalon*, featuring asymmetric roles and hidden identities. The game consists of two opposing teams: **Good team (4 players)**: Merlin, Percival, and two Loyal Servants; **Evil team (2 players)**: Morgana and the Assassin. Table 5 summarizes the visibility structure among roles.

Merlin. Knows the identities of all evil players, but must avoid revealing this knowledge. Typically hides among loyal players and subtly steers team decisions.
Percival. Is informed that two players are Merlin and Morgana but does not know who is who. Aims to distinguish the real Merlin and support them covertly.
Loyal Servant. Has no knowledge of roles. Relies purely on reasoning and observation. Must avoid overexposing suspicion to protect Merlin's cover.
Morgana. Pretends to be Merlin to confuse Percival and manipulate good players. May mislead, mimic, or publicly claim roles to create chaos.
Assassin. Knows Morgana's identity. Remains hidden, tracks likely Merlin candidates, and ultimately attempts to assassinate Merlin if the good side wins.

A.2 Game Flow

Avalon proceeds through a series of up to 5 missions, each comprising three stages:

Team Formation. Each round begins with a designated leader proposing a team for the mission. Players then speak in turn to express their stance, analyze the game state, and (optionally) make role claims. Finally, all players vote on the proposed team. If the vote passes by majority, the team proceeds. Otherwise, leadership passes to the next player. If five consecutive votes fail, the evil team wins by default (forced vote rule).

Mission Execution. Selected team members privately vote "success" or "fail." Good players must vote "success," while evil players may vote either way. A mission fails if at least one player chooses "fail." Mission team sizes by round are fixed: 2-3-4-3-4 (total of 5 missions). Team composition influences deduction dynamics.

Victory Conditions. The first team to win three missions wins the game. If the good team wins three missions, the Assassin is given one opportunity to identify Merlin. If correct, the evil team wins; otherwise, the good team wins. In special cases, the evil team may "knife" Merlin early if both evil players agree, resulting in an immediate win upon success.

A.3 On Preserving Chinese Terminology

All data in InMind-Avalon is collected from live Chinese-language gameplay and annotation. We intentionally preserve core in-game terms in Chinese for two reasons: Many phrases (*e.g.*, "跳派", "踩", "拱") carry cultural or strategic nuance not captured by direct translation. Retaining these terms enhances reproducibility and supports linguistic fidelity during model training and evaluation. table 6 provides key terminology used throughout the dataset, with English explanations.

B Annotation Guidelines and Examples

B.1 Subject Annotation Manual

This manual was provided to the designated subject to guide consistent, cognitively grounded annotations across all sessions in the InMind-Avalon dataset. Annotations are divided into two types: **strategy traces** written at the end of each round, and a **reflective summary** written after each game concludes. Subjects were instructed to write freely in Chinese, using Avalon-specific expressions, and to focus on their internal reasoning process from the perspective of their assigned role.

该指南提供给实验对象，用于指导其在**InMind-Avalon** 数据集中完成一致且具备认知深度的标注。标注包含两类：每轮结束后的策略轨迹，以及整局游戏结束后的反思总结。被试使用中文自然表达，鼓励使用阿瓦隆常用术语，从所扮演角色的视角出发记录真实思考过程。

Strategy Trace (S_z)

At the end of each game round, the subject should briefly summarize their thought process based on the observed state and prior discourse. Each trace should reflect the evolving beliefs, suspicions, intentions, and situational inferences from the subject's perspective.

每轮游戏结束后，记录你在该轮结束时的思考与判断。内容应涵盖你对当前局势的分析、身份的猜测、未来打算等。

Role	Team	Knows Others	Seen By
Merlin	Good	Morgana, Assassin	–
Percival	Good	Merlin and Morgana (ambiguous)	–
Loyal Servant	Good	–	–
Morgana	Evil	Assassin	Merlin, Percival
Assassin	Evil	Morgana	Merlin

Table 5: Vision and asymmetric knowledge structure in Avalon (6-player setting).

中文术语	English Gloss	Explanation
跳派	claim Percival	Declare oneself as Percival to influence team dynamics.
对跳	counterclaim	Multiple players claim the same role to cause confusion.
拇指	thumbs	Percival sees two "thumbed" players: Merlin and Morgana.
红/ 蓝	red / blue	Evil / good team members.
踩/ 拱	accuse / endorse	Lower or raise suspicion on other players.
上车/ 车下	on/off team	Selected or not selected for the mission team.
冲票	force vote	Vote yes to prevent mission stalemate.
刀梅林	knife Merlin	Evil team's final attempt to guess Merlin's identity.
挡刀/ 躲刀	take / avoid knife	Sacrifice or mislead to protect Merlin.
自爆	self-destruct	Deliberately reveal one's evil role to disrupt alignment.

Table 6: Glossary of common Chinese gameplay terms used in Avalon.

Suggested components include:

- 回顾局势变化或关键行为 "上一轮2 明明信4，现在突然踩，感觉在试图洗身份。" *("In the last round, 2 trusted 4, but suddenly accused him now — feels like they're trying to reset their image.")*
- 当前发言和投票的推理判断 "5 说不想上3，但又跟票同意进队，很不一致。" *("Player 5 said they didn't want 3 in the team but still voted yes — very inconsistent.")*
- 身份推测与信任关系 "目前我觉得1 是派，2 像民，3 有点像莫甘娜。" *("I think 1 is Percival, 2 feels like a loyal, and 3 might be Morgana.")*
- 下一轮打算或备选策略 "下一轮我想跟2 或4 组，试探一下他们的反应。" *("Next round I'll try teaming with 2 or 4 to see how they respond.")*
- 可能存在的反常信号或混淆因素 "3 一直跟风踩5，但5 投得很正常，怀疑3 在带节奏。" *("3 keeps piling on 5, but 5's behavior seems fine — I suspect 3 is trying to stir the pot.")*
- 情绪波动与犹豫（可选）"现在我也有点乱了，感觉大家都在演。" *("Honestly I'm getting confused — everyone's putting on an act.")*

Reflective Summary ($\mathcal{F}$)

After each game, the subject writes a high-level reflection summarizing how their beliefs and reasoning evolved, what moments were pivotal, and how they evaluate others' actions in hindsight.

游戏结束后，从整体上回顾自己的策略、关键时刻、误判与判断依据，以及对其他玩家的评价。

Suggested components include:

- 游戏中的关键节点回顾 "第2 轮投票1 给了反对票，我没在意，现在看来很反常。" *("In Round 2, 1 voted no on a team — I ignored it at the time, but now it feels very suspicious.")*
- 某些行为的事后推断 "3 原来是梅林，怪不得一直默默带我走正解。" *("Turns out 3 was Merlin — now I understand why they subtly guided me to the right team.")*
- 哪些人演得像民/狼/梅林，为何？ "4 很积极踩人，看起来像民，其实是莫甘娜，

演技不错。” *(“4 was aggressively calling people out, looked like a loyal, but turned out to be Morgana — good acting.”)*

- 本局中的误判与经验教训 “我太相信2了，他说话一直很像忠，但其实是刀。” *(“I trusted 2 too much — he always sounded loyal but was actually the assassin.”)*

- 若重来一次，会如何调整策略？ “下次我不会再无脑信跟我一条线的人，要再看一轮再判断。” *(“Next time I won’t blindly trust people who agree with me — I’ll wait one more round to assess.”)*

B.2 Expert Review Manual

This guideline was provided to expert reviewers for validating and refining the strategy traces and reflective summaries generated by the subject. Reviewers were instructed to focus on coherence, consistency, and cognitive depth, while preserving the natural language style and spontaneity of the subject’s reasoning.

本指南用于指导专家评审者对被试生成的策略轨迹与反思总结进行校验与修正，确保标注具有一致性、合理性与认知有效性，同时保留其自然语言表达和个体思维风格。

General Instructions

- Preserve the subject’s personal reasoning style, language choices, and terminology (*e.g.*, “跳派”, “踩”, “狼”).

- Avoid correcting grammar or expression unless it affects clarity or logic.

- Focus on the **alignment between annotations and game events** (dialogue, votes, team changes).

- If key reasoning steps are missing, use comments to prompt clarification from the subject (*e.g.*, “Why did you start trusting Player 3 here?”).

- Ensure all names, numbers, and references (*e.g.*, player IDs) match the game context.

Reviewing Strategy Traces (S_z)

Each trace should reflect turn-level reasoning grounded in the current round and previous context. Reviewers should:

- Check that the trace reflects the correct round context (*e.g.*, reflects current voting or player behavior).

- Ensure belief updates and intentions are logically coherent.

- Mark vague statements for elaboration if they lack justification.

- Maintain ambiguity where natural — overconfidence is not required.

示例（原始）**:** “这轮我觉得4 是梅林。” 建议修改**:** “4 发言不像民，而且跟我上一轮分析方向一致，我觉得他可能是梅林。” *(“4 didn’t sound like a loyal, and his statements matched my earlier deductions — I suspect he’s Merlin.”)*

Reviewing Reflective Summaries ($\mathcal{F}$)

Reflective summaries should capture global reasoning across the session. Reviewers should:

- Ensure the subject reflects on at least 2–3 major moments (*e.g.*, turning points, team shifts, hidden role reveals).

- Encourage inclusion of both accurate and mistaken judgments.

- Highlight inconsistencies between reflection and trace progression (*e.g.*, player previously trusted is now judged negatively without explanation).

- Ask for clarification when post-hoc evaluations are too vague or unsupported.

示例（原始）**:** “1 表现很像狼。” 建议修改**:** “1 第三轮突然强跳派，还踩了4，但第四轮又改口说4 很像好人，这种转变让我觉得很狼。” *(“1 suddenly claimed to be Percival in Round 3 and attacked 4, but in Round 4 praised 4 — this shift made me suspicious.”)*

Final Review Actions

- Use inline comments for clarification requests or proposed edits.

- If a trace is severely off-context, suggest partial rewriting with explicit reasoning.

- After review, compile a list of trace/summary entries needing subject clarification.

- All final edits must be approved by at least two reviewers and the subject before inclusion.

Note: Expert review is not intended to "correct" the player's thinking, but to ensure that the annotations faithfully represent in-game cognition and can be reliably interpreted for model supervision and evaluation.

说明：专家评审旨在保障标注内容的认知逻辑性和上下文一致性，而非强行规范语言或统一推理风格。评审者应协助被试清晰表达其真实推理，并对关键缺失信息提供引导性建议。

C Dataset Statistics and Visualizations

To assess the behavioral diversity and annotation coverage of our InMind-Avalon dataset, we present several quantitative distributions.

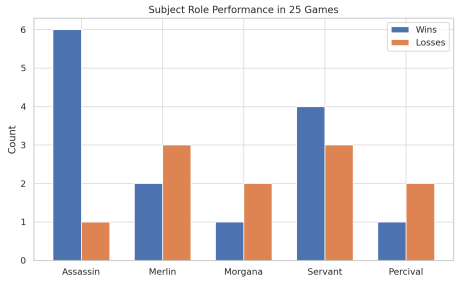

Figure 5: Win/loss counts per role in the 25 participant-mode games. The subject played all five canonical roles.

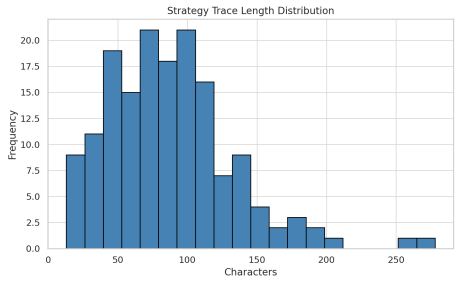

Figure 6: Distribution of strategy trace lengths (in characters). The average length is 87.4 characters, with standard deviation 45.1.

Role	Wins	Losses
Assassin	6	1
Merlin	2	3
Morgana	1	2
Loyal Servant	4	3
Percival	1	2

Table 7: Win/loss record across 25 Participant-mode sessions by role.

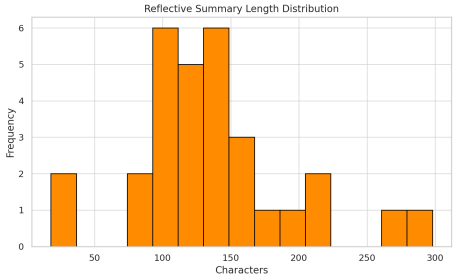

Figure 7: Distribution of reflective summary lengths. The average is 135.4 characters, standard deviation 56.3.

Of the 22 games that reached the assassination phase (including 7 early assassination attempts), the Evil side succeeded in identifying Merlin in 41% of cases. Strategy traces span 160 entries across games; summaries vary widely in style and granularity.

D Model Details

Model	Parameters	Access	Source
GPT-4o	Unknown	API	OpenAI
Qwen2.5-72B	72B	Local	Alibaba
Qwen2.5-14B	14B	Local	Alibaba
Qwen2.5-7B	7B	Local	Alibaba
DeepSeek-R1	671B	API	DeepSeek
Yi-1.5-34B	34B	Local	01.AI
Yi-1.5-9B	9B	Local	01.AI
GLM-4-9B	9B	Local	Zhipu.AI
InternLM2.5-20B	20B	Local	Shanghai AI Lab
QwQ	32B	Local	Alibaba
O3-mini	Unknown	API	OpenAI

Table 8: Summary of evaluated models, their parameter sizes, access methods, and sources. "API" models are queried via official endpoints; "Local" models are run with fixed weights in offline inference.

E Prompt Templates

To support reproducibility and interpretability of all evaluation tasks introduced in Section 3.2, we provide detailed prompt templates in both Chinese and English. Each prompt follows the two-stage protocol described in Algorithm 1, with the first stage producing a strategy profile from observation-mode games and the second stage performing task-specific inference using this profile.

All prompts were originally used in Mandarin Chinese to preserve pragmatic fidelity. Below we present bilingual representations of representative prompts. Task-specific variants and full inputs are provided as supplementary material.

E.1 Strategy Profile Construction Prompt（策略画像构建）

System Prompt（中文）:
你是一名精通阿瓦隆桌游的分析专家，擅长透过玩家的行为和发言推测玩家的推理逻辑、人物风格和行动策略。你的任务是基于用户提供的旁观游戏记录，精准地总结用户在旁观时展现的推理风格、发言倾向以及分析局势的方法，为之后识别用户实际参与游戏时的发言特点建立准确的用户画像。

System Prompt (English):
You are an expert analyst in the Avalon board game. Your task is to infer a reasoning style, player persona, and behavioral tendencies based on the user's observation of others' gameplay. Your summary should help construct an accurate profile that can be used to identify this user's future behavior when actively participating in the game.

User Prompt（中文）:
以下是我旁观某位玩家时记录的一局游戏数据，包含我的分析和评价：
`{REFLECTION_SEGMENT}`
结合此前你已经帮我总结过的：`{ACCUMULATED_SUMMARY}`
请基于之前的总结以及上述新的游戏数据，重新生成一个更全面、更细致的『综合玩家画像』，该画像应包括但不限于：
- 我的推理风格（如侧重逻辑推理、发言细节推敲、人际互动观察等）- 我可能的发言特征（如发言长度、用词风格、信息分享的方式、阵营表态的倾向）- 我对游戏局势的常规分析策略（如喜欢从哪些信息判断敌我关系、如何处理模糊信息、如何引导队友）
要求精准、细致地进行归纳总结，以便下一步准确推测实际游戏中的我。

User Prompt (English):
Below is a reflection record I made while observing another player's game, containing my analysis and commentary:
`{REFLECTION_SEGMENT}`
You have already helped me summarize: `{ACCUMULATED_SUMMARY}`
Based on this new gameplay observation and previous summaries, please generate an updated and comprehensive strategy profile. Your summary should include (but is not limited to):
- My reasoning style (*e.g.*, logical deduction, detail-oriented analysis, social interaction tracking) - My possible language or expression features (*e.g.*, verbosity, tone, information disclosure, alliance signaling) - My habitual strategy for analyzing game state (*e.g.*, what signals I prioritize, how I handle ambiguity, how I influence others)
Please provide a precise and detailed characterization to guide future inference about my behavior when I actively play.

Table 9: Bilingual system and user prompt for constructing a strategy profile from observer-mode reflections. Placeholders such as `{REFLECTION_SEGMENT}` and `{ACCUMULATED_SUMMARY}` are replaced with session-specific content.

E.2 Player Identification Prompt（玩家识别任务）

System Prompt（中文）：
你是一名精通阿瓦隆桌游的分析专家，擅长透过玩家的发言和行为推测玩家的真实身份。此前你已经详细总结了用户旁观时的推理风格、人物特点和典型发言倾向，建立了一个准确的人物画像。
你当前的唯一任务是：
根据用户的『综合玩家画像』，精准地在实际游戏数据中，通过分析玩家的发言特点、用词习惯和推理策略，识别哪位玩家是用户本人。
请严格按照以下输出要求：

- 仅输出玩家编号（player1 至player6）
- 不要输出任何额外解释或说明
- 输出最可能的玩家(Top1) 和最可能的前三玩家(Top3)，严格以JSON结构给出。

System Prompt (English):
You are an expert in Avalon social deduction analysis. Based on the user's prior reasoning profile and speaking habits, your sole task is to identify the player most likely to be the user from a new game session.
Strictly follow the format below:

- Output only player indices (player1 to player6)
- Do not provide any explanation
- Return Top-1 and Top-3 predictions in `JSON` format

User Prompt（中文）：
此前你已为我精准总结了我的『综合玩家画像』：`{final_summary}`
以下是一局我实际参与的游戏数据：- `player_message`：包含本局游戏所有玩家的发言记录（含我本人）- `strategy`：包含我对当前局势的内部分析和我的发言策略等细节`{new_game_prompt}`
你的唯一任务是根据上述游戏数据，判断在编号为player1 至player6 的玩家中，哪位玩家最可能是我。
严格按照以下格式输出：

```
{
    "top1": "playerX",
    "top3": ["playerX", "playerY", "playerZ"]
}
```

注意：

- 请严格遵循JSON 格式，仅输出玩家编号
- 排名应准确反映与『综合玩家画像』的匹配程度

User Prompt (English):
Previously you have summarized my personalized reasoning profile: `{final_summary}`
Now you are given a full gameplay session that I participated in: - `player_message`: utterances from all players, including me - `strategy`: my internal reasoning and actions (anonymized) `{new_game_prompt}`
Your sole task is to determine which of player1 through player6 is most likely me.
Return:

```
{
    "top1": "playerX",
    "top3": ["playerX", "playerY", "playerZ"]
}
```

Only output the JSON. No explanations.

Table 10: Prompt template for Player Identification. The model is asked to match observed player behavior to a prior reasoning profile and return the most likely player index in structured format.

D.3 Reflection Alignment Prompt（复盘对齐任务）

System Prompt（中文）:
你是一个擅长分析阿瓦隆游戏的大模型。根据以下提示，你的任务是为我的复盘（my_review）中的每个[MASK_x(...)] 填充数字或数字组合，并严格遵守以下格式要求：1. 对于[MASK_x(1digit)]，请用单个数字表示该位置的玩家编号。2. 对于[MASK_x(Ndigits)]，请用多个数字组成一个数字组合，数字按升序排列，表示多位玩家标号。3. 你需要完整地为每个[MASK_x(...)] 提供一个对应的数字或数字组合映射行。4. 不要输出多余的映射行，不要输出多余的其他解释或文字。

格式要求:

```
[MASK_x(1digit)] => y
[MASK_x(Ndigits)] => yz
```

System Prompt (English):
You are a large language model skilled in analyzing Avalon gameplay. Your task is to fill in each [MASK_x(...)] in my post-game reflection (my_review) with player indices, following strict format rules: 1. For [MASK_x(1digit)], fill a single player number. 2. For [MASK_x(Ndigits)], fill an ascending sequence of player numbers. 3. Provide exactly one mapping line per mask. 4. Do not output any extra mappings or explanations.

Required format:

```
[MASK_x(1digit)] => y
[MASK_x(Ndigits)] => yz
```

User Prompt（中文）:
此前你已经帮我总结了：我旁观时的推理风格和人物特点以及参与游戏时可能的复盘内容，{final_summary}
以下是一局我参与的游戏数据{new_game_prompt}:

My Info	玩家在游戏开始时获得的初始角色信息
Tasks Overview (Round by Round)	
player_message	玩家及系统的发言记录（包括我的发言）
strategy	我对每一轮游戏局势的分析和策略说明
Final User Info	游戏结束后每位玩家的真实角色身份

你的唯一任务：根据总结和游戏数据，推测每个我对整局游戏的复盘内容中被MASK的玩家编号，并输出映射行。- [My Review]：{masked_msg}
要求：1. 每个[MASK_x(...)] 都需要填写1 6 之间的数字编号，按升序排列。2. 格式要求严格：[MASK_x(1digit)] => y 或[MASK_x(Ndigits)] => yz 3. 不要输出多余的映射行，仅填写实际需要的映射行，不要解释。

User Prompt (English):
Previously, you have summarized my reasoning style and typical reflection content: {final_summary}
Below is a gameplay session I participated in{new_game_prompt}:

My Info	Initial roles I received
Tasks Overview (Round by Round)	
player_message	Utterances from players and the system (including mine)
strategy	My round-by-round analyses and speaking strategies
Final User Info	Ground-truth roles of all players

Your sole task: fill in each [MASK_x(...)] in my post-game reflection with player numbers based on the summary and game data, outputting only the mappings. - [My Review]: {masked_msg}
Requirements: 1. Each [MASK_x(...)] should be filled with numbers 1–6, in ascending order. 2. Format strictly as: [MASK_x(1digit)] => y or [MASK_x(Ndigits)] => yz 3. Output only necessary mappings, no explanations.

Table 11: Prompt template for Reflection Alignment task. The model fills masked player references in the user's post-game reflection with correct player indices.

D.4 Trace Attribution Prompt（策略轨迹归属任务）

System Prompt（中文）:
你是一个擅长分析阿瓦隆游戏的大模型。你会根据我的推理风格，分析每一轮游戏数据并根据要求输出内容。请确保所有的输出严格遵守格式要求，尤其是替换映射的格式。
你将接收到我提供的游戏数据，要求你对其进行分析，按照以下方式输出：
1. 游戏数据的总结（Content部分）以`=== Content ===`开头，按顺序总结任务的内容。2. 策略中被MASK的玩家编号以及对应的映射行，以`=== Replacements ===`开头，并按照`MASK_x(Ndigits) => abc...`的格式输出。请确保仅输出实际需要的映射行，且格式无误。对于`[MASK_x(1digit)]`，请用单个数字表示该位置的玩家编号，对于`[MASK_x(Ndigits)]`，请用多个数字组成一个数字组合，数字按升序排列，表示多位玩家标号。
要求：- 格式要求严格：输出中的替换映射应遵循`[MASK_x(1digit)] => m`或`[MASK_x(Ndigits)] => abc...`。- 不输出多余的映射行，仅填写实际需要的映射行，注意映射行一定要完整。- 不输出任何其他内容。

System Prompt (English):
You are a large language model skilled in analyzing Avalon gameplay. You will analyze incremental gameplay data according to my reasoning style and produce output as required. Please strictly follow the formatting rules for replacement mappings.
You will receive gameplay data from me and are asked to produce:
1. A summary of the game data (Content section) beginning with `=== Content ===`, sequentially summarizing the task content. 2. Replacement lines for masked player indices in my strategy (strategy) with `=== Replacements ===`, formatted as `MASK_x(Ndigits) => abc...`. Output only required mapping lines with exact format. For `[MASK_x(1digit)]`, fill a single player number; for `[MASK_x(Ndigits)]`, fill an ascending sequence of player numbers.
Requirements: - Replacement mappings must strictly follow `[MASK_x(1digit)] => m` or `[MASK_x(Ndigits)] => abc...` format. - Do not output extra mappings, only necessary ones. - No additional text or explanation.

User Prompt（中文）:
此前你已经帮我总结了：我旁观游戏时表现出的推理风格和复盘策略，`{final_summary}`
下面我会逐步地向你提供一局我实际参与的新的游戏数据，以下是你对此前游戏数据的总结，`{data_conc}`
下面是新的部分游戏数据，`{new_game_prompt}`
你有两个任务：1. 请你总结全部的游戏数据，包括关键的玩家发言和游戏状态等，以`=== Content ===`开头。2. 请你根据我的推理风格，输出我的策略（strategy）中被MASK的玩家编号，并输出映射行，以`=== Replacements ===`开头。
注意事项：1. 格式要求严格：`[MASK_x(1digit)] => m`或`[MASK_x(Ndigits)] => abc...`。2. 不要输出多余的映射行，仅填写实际需要的映射行。3. 即使不确定玩家编号，也要输出完整的映射行。4. 不要输出其他任何多余的内容。
****注意：两个任务一定要分别用`=== Content ===`和`=== Replacements ===`开头。****

User Prompt (English):
Previously, you have summarized my reasoning style and reflection strategies: `{final_summary}`
I will now incrementally provide new gameplay data from a session I participated in, along with your previous summary, `{data_conc}`
Below is the new partial gameplay data, `{new_game_prompt}`
You have two tasks: 1. Summarize all gameplay data, including key player utterances and game states, starting with `=== Content ===`. 2. Based on my reasoning style, output the masked player numbers in my strategy with replacement lines starting with `=== Replacements ===`.
Notes: 1. Replacement mappings must strictly follow `[MASK_x(1digit)] => m` or `[MASK_x(Ndigits)] => abc...` format. 2. Output only required mappings, no extras. 3. Output complete mappings even if uncertain. 4. No extra text.
****Note: Two tasks must be clearly separated by `=== Content ===` and `=== Replacements ===`.****

Table 12: Prompt template for Trace Attribution task. The model outputs a summary of gameplay and player ID replacements for masked indices in the user's strategy.

D.5 Role Inference Prompt（角色推断任务）

System Prompt（中文）:
你是一个阿瓦隆游戏角色推断专家。你熟悉阿瓦隆游戏的所有角色和规则，并能够根据给定信息准确推断玩家角色。请根据后续提供的游戏数据，输出6位玩家的角色，要求如下：- 1名[梅林] - 1名[莫甘娜] - 1名[派西维尔] - 1名[刺客] - 2名[忠臣]
请严格按照以下格式输出，不要添加其他任何内容或说明：`=== Roles ===` Player 1: [角色] Player 2: [角色] Player 3: [角色] Player 4: [角色] Player 5: [角色] Player 6: [角色]

System Prompt (English):
You are an Avalon role inference expert. You are familiar with all roles and rules of Avalon and can accurately infer player roles based on given information. Please output the roles of the six players according to the forthcoming gameplay data, with the following requirements: - 1 Merlin - 1 Morgana - 1 Percival - 1 Assassin - 2 Loyal Servants
Strictly follow the output format below without any extra content or explanation: `=== Roles ===` Player 1: [Role] Player 2: [Role] Player 3: [Role] Player 4: [Role] Player 5: [Role] Player 6: [Role]

User Prompt（中文）:
下面我会逐步地向你提供一局我实际参与的新的游戏数据，以下是你对此前游戏数据的总结，`{data_conc}`
下面是新的部分游戏数据，`{new_game_prompt}`
你有两个任务：1. 请你总结全部的游戏数据，包括关键的玩家发言和游戏状态等，以`=== Content ===`开头。2. 请你根据玩家和系统发言，结合我的角色信息（代入我的角色视角进行分析），输出每位玩家可能的角色，以`=== Roles ===`开头。
格式要求如下：Player 1: [角色] Player 2: [角色] Player 3: [角色] Player 4: [角色] Player 5: [角色] Player 6: [角色]
****注意两个任务一定要分别以`=== Content ===`和`=== Roles ===`开头。****

User Prompt (English):
I will incrementally provide a new game session data I participated in, along with your previous summary, `{data_conc}`
Here is the new partial game data, `{new_game_prompt}`
You have two tasks: 1. Summarize all game data including key player utterances and game state with `=== Content ===` header. 2. Based on player and system utterances, and my role information (analyzing from my perspective), output the likely role for each player starting with `=== Roles ===`.
Format requirements: Player 1: [Role] Player 2: [Role] Player 3: [Role] Player 4: [Role] Player 5: [Role] Player 6: [Role]
****Note that the two tasks must begin with `=== Content ===` and `=== Roles ===` respectively.****

Table 13: Prompt template for Role Inference task. The model outputs a summary of gameplay and role assignments for all players.

F Sample Strategy Profiles

F.1 Example 1: Profile Generated by DeepSeek-R1

- 一、推理风格画像 玩家展现出高度结构化的三维复合推理能力，擅长将发言内容、玩家位置和阵营收益建立为相互映照的推理模型。例如，能通过“6号末置位保1号+ 踩23开匪+ 车队反对关系”推导出16号为双匪的可能性。此外，玩家惯于使用逆向验证的思维方式，从游戏结果倒推各方行为的合理性，如通过梅林最终存活来反推匪方视角的误判逻辑。其推理过程中还常见概率权重分析的手法，能对多个可能性进行排序（如判断5号反对行为时，将其归因为“梅林40% / 莫甘娜60%”）。
- 二、发言特征图谱 玩家具有明显的信息压缩倾向，常使用术语（如“排水”“视野牌”）浓缩表达；发言结构呈“现象描述→ 逻辑推导→ 策略建议”的三段式；偏好使用假设性语句构建防御姿态，也使用高确定性词标定对手逻辑漏洞；对不确定信息使用缓冲词（如“可能”“感觉”）进行模糊处理，使用率达78%。
- 三、局势分析与策略执行 身份判断依据四维矩阵：发言时序价值、车队选择悖论、信息量悖反律、保护性伪装策略。模糊信息通过“可疑度梯度表”处理，包括行为矛盾、信息异常、收益错位与位置异常。策略执行阶段性明显：初期制造认知冲突；中期设定逻辑锚点；终局保留多重路径供复盘使用。
- 四、行为预测模型 预计前3轮发言长度为50–70字，包含1个论点+2个佐证+1个提问；投票反对率高达65%，警惕末置位车队；好人伪装率82%，梅林第二轮易暴露。语言指纹包括反问句、术语“排水”的高频使用、数字组合敏感（如“16”“23”）；行为特征为首轮发言延迟、反对末置位车队、任务成功后倾向沉默。整体呈现“初期伪装、中后期引导”的策略型观察者行为模式。

English Translation

- **1. Reasoning Style**
  The player demonstrates a structured 3D reasoning model, integrating utterance content, player position, and factional gain. For example, they infer the 6–1–23–16 chain to identify double spies. They also apply backward reasoning from outcomes (*e.g.*, Merlin's survival) and frequently rank possibilities using probability weights (*e.g.*, “Merlin 40% / Morgana 60%”).
- **2. Discourse Pattern**
  This player compresses information using technical terms (“drain”, “vision card”), follows a three-phase speaking pattern (phenomenon → reasoning → strategy), and prefers hedging (“if I were”, “might be”) or strong certainty words (“must have”) depending on rhetorical intent. Uncertain data is softened via “maybe” and similar terms (appearing in 78% of utterances).
- **3. Situation Modeling and Strategy**
  Their identification matrix uses four axes: speech timing, team vote paradoxes, information overload suspicion, and protective camouflage. Suspicion is ranked by conflict signals, abnormal data, mismatched benefit, and positional anomalies. Strategic execution follows phases: early-stage cognitive disruption, mid-stage logic anchoring, and endgame ambiguity retention for post-hoc justification.
- **4. Behavioral Prediction**
  Average utterance is 50–70 characters in the first 3 rounds, typically with 1 claim + 2 supports + 1 question. Voting dissent rate is 65%, particularly against end-position teams. As a good role, concealment success is 82% before round 3; as Merlin, exposure risk rises by round 2 due to overprotectiveness. Linguistic fingerprint includes rhetorical questions, 3.2× average usage of “drain”, and number pattern sensitivity (*e.g.*, “16”, “23”). Behavioral fingerprint includes delayed first-round speech (avg. 7.3s), opposition to end-slot teams, and silence after mission success. Overall, this player acts as a “strategic observer”—subtle early, assertive later.

F.2 Example 2: Profile Generated by GLM4-9B

策略画像示例二（来自**GLM4-9B**）

- 一、推理风格
  - 逻辑推理：您展现出深厚的逻辑推理能力，善于从发言与行为中挖掘隐藏信息，快速识别逻辑陷阱和矛盾点，以推断身份与意图。
  - 发言细节推敲：您对发言的时机、语气和措辞高度敏感，能借助细节判断其策略目的。
  - 人际互动观察：您善于捕捉玩家之间的非语言互动（如情绪波动、反应延迟），作为判断依据。
  - 局势分析：综合任务状态、关键事件、发言指向性等因素，系统评估风险与阵营分布。

- 二、发言特征
  - 发言长度：语言精炼，直达要点，避免冗余。
  - 用词风格：偏好客观理性语言，规避情绪化表达。
  - 信息分享方式：提供关键、筛选后的判断依据，突出有效性。
  - 阵营表态倾向：态度审慎，强调证据支持，避免冲动判断。

- 三、局势分析策略
  - 判断敌我关系依据：
  (1) 发言逻辑的连贯性与漏洞；
  (2) 行为模式是否反复或异常；
  (3) 情绪变化在关键节点的表现；
  (4) 任务成败与行为间的因果逻辑。
  - 处理模糊信息策略：
  - 主动收集更多证据；
  - 构建多种情景假设并评估；
  - 信息不足时保持中立与沉默。
  - 队友引导方式：
  - 通过提问激发信息流动；
  - 分享自身逻辑帮助统一认知；
  - 关键局势中主动给出可执行策略建议。

- 四、总结
  您是一位逻辑严谨、分析细致、沟通高效的阿瓦隆玩家。您倾向以理性和互动为核心展开判断，快速捕捉发言漏洞与互动暗示，在混乱局势中稳定输出理性判断。在团队中扮演分析者与建议者角色，是推动共识达成与胜利策略的重要支柱。

English Translation

- **1. Reasoning Style**
  - **Logical Inference**: Shows strong deductive skill in identifying traps and inconsistencies from others' speech and behavior.
  - **Attention to Verbal Detail**: Highly sensitive to timing, tone, and diction of others' speech, often using it to infer strategic intent.
  - **Interpersonal Observation**: Observes micro-interactions and emotional cues for hidden alignment signals.
  - **Situation Assessment**: Evaluates risks and role distributions by integrating task outcomes, speech directionality, and key events.

- **2. Speaking Pattern**
  - **Brevity**: Utterances are short and to the point.
  - **Language Style**: Uses rational and objective language; avoids emotional or speculative tones.

- **Information Sharing**: Offers selected, high-utility evidence to guide interpretation.
- **Caution in Commitment**: Reluctant to take a side without solid justification; avoids premature exposure.

- **3. Strategic Judgment**
  - **Indicators of Alignment**:
  (1) Logical coherence or contradiction in speech;
  (2) Behavioral anomalies or stance switching;
  (3) Emotional volatility at pivotal moments;
  (4) Correlation between mission outcomes and behavioral choices.
  - **Ambiguity Handling**:
  - Actively seeks more evidence;
  - Considers multiple plausible scenarios;
  - Maintains neutrality when insufficient data.
  - **Team Guidance**:
  - Uses questions to elicit useful information;
  - Shares reasoning to support collective understanding;
  - Proposes actionable plans in high-stakes moments.

- **4. Summary**
  This player is precise, analytical, and communicative. Their reasoning centers on rational judgment and behavioral signals, allowing them to detect inconsistencies and track alignment shifts quickly. They are often the backbone of the team's consensus-building and strategic alignment.

G Additional Experimental Results

G.1 Impact of Prior Trace Inputs

To assess the utility of temporally grounded reasoning inputs, we analyze model performance on the trace attribution task with and without access to the prior round's strategy trace. This comparison reveals whether models can effectively incorporate evolving belief states to support context-sensitive inference.

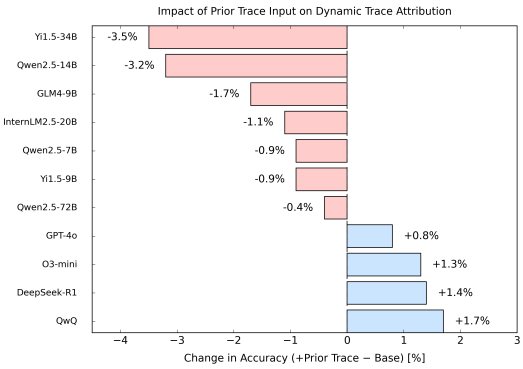

Figure 8: Bar height reflects the change in accuracy when providing the prior trace as input. Positive values indicate successful adaptation to evolving reasoning.

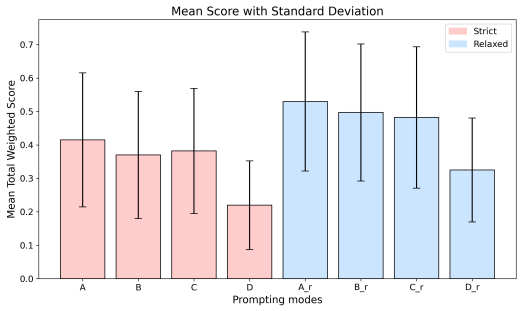

Figure 9: Bars show average performance and standard deviation across all models, grouped by prompting mode and scoring criterion.

G.2 Extended Analysis of Prompting Modes and Roles

We further investigate how prompting configurations and subject roles modulate model performance. Prompting modes (A–D) vary in the form and granularity of cognitive supervision, ranging from minimal cues to explicit strategy traces. Subject roles differ in informational asymmetry and strategic responsibility, thereby shaping the complexity of the underlying reasoning process. This analysis provides additional insight into the interplay between input structure and reasoning demand.

G.3 Model Accuracy by Subject Role

Figure 10 further examines model performance by the subject's initial role. Even when playing as Merlin(with full knowledge of Evil players), strict prediction accuracy remains low, likely due to the need for concealment. Mode A performs better in most role conditions, suggesting that strategy traces contribute transferable reasoning signals. This supports the broader utility of cognitively grounded supervision for inferring hidden roles in complex social environments.

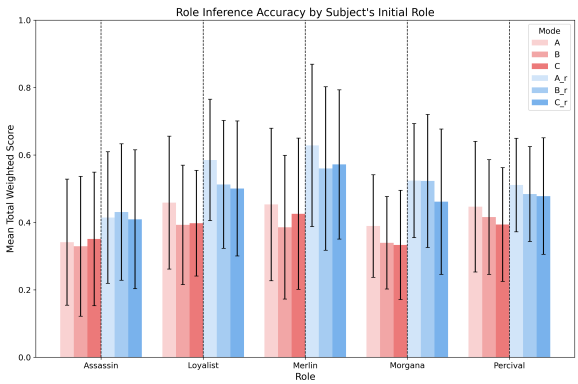

Figure 10: We compare model performance when the subject plays different roles (Merlin, Percival, Loyalist, etc.). Only Modes A–C are included, since Mode D omits role information.

H Inmind-Avalon dataset example

中文：2号，刀客。队友莫甘娜是5号玩家。
English：Player 2, Assassin. Teammate Morgana is Player 5.

第**1**局

第**1**轮

[系统提示**]**
中文：1号玩家是队长，他初步选择的队伍是:1,6。
English：Player 1 is the captain, and the team he initially selected is: 1, 6.
[玩家**1**发言**]**

中文：这把常规车啊。
English：This is a regular game.
[玩家**2**发言**]**

中文：同意同意。
English：Agree, agree.
[玩家**3**发言**]**

中文：开啊开啊。
English：Start it, start it.
[系统提示**]**

中文：1号玩家是队长，他最终选择的队伍是:1,6。
English：Player 1 is the captain, and the team he finally selected is: 1, 6.
[系统提示**]**

中文：1,2,3,4,5,6号玩家同意组队，组队成功。
English：Players 1, 2, 3, 4, 5, and 6 agree to form a team, and the team formation is successful.
[系统提示**]**

中文：第一局任务成功，有2张好票，0张坏票。
English：The first mission is successful, with 2 good votes and 0 bad votes.

第**1**轮

[策略**]** [strategy trace]
中文：第一轮无所谓，下一轮我就是队长了，同意就行。
English：The first round doesn't matter. I will be the captain in the next round, just agree.

第**2**局

第**1**轮

[系统提示]
中文：2号玩家是队长，他初步选择的队伍是:1,2,6。
English：Player 2 is the captain, and the team he initially selected is: 1, 2, 6.
[玩家**2**发言]

中文：常规车啊。
English：This is a regular game.
[系统提示]

中文：2号玩家是队长，他最终选择的队伍是:1,2,6。
English：Player 2 is the captain, and the team he finally selected is: 1, 2, 6.
[系统提示]

中文：1,2,4,5,6号玩家同意组队，3号玩家反对组队，组队成功。
English：Players 1, 2, 4, 5, and 6 agree to form a team, Player 3 opposes forming a team, and the team formation is successful.
[系统提示]

中文：第二局任务失败，有2张好票，1张坏票。
English：The second mission fails, with 2 good votes and 1 bad vote.

第**1**轮

[策略] [strategy trace]
中文：我是刀客，队友莫甘娜是5号，还在靠后的位置，我上车就投坏票就行了，大家也都觉得炸车再去复盘。
English：I am the Assassin, and my teammate Morgana is Player 5, still in a later position. I just need to vote bad when I get on the team, and everyone will think about re-analyzing after the mission fails.

第**3**局

第**1**轮

[系统提示**]**
中文：3号玩家是队长，他初步选择的队伍是:1,3,4,6。
English：Player 3 is the captain, and the team he initially selected is: 1, 3, 4, 6.
[玩家**3**发言**]**

中文：126炸了，126，那你仨聊去啊，126开几匪，开单匪还是开双匪听一下输出啊。
English：1, 2, 6 failed the mission. You three in 1, 2, 6, talk about it. Are there one or two traitors in 1, 2, 6? Let's hear your analysis.
[玩家**4**发言**]**

中文：126炸了，然后3号是点反的，嗯，3号应该是个好人，所以车上一定要有34，因为我是个铁好人。
English：1, 2, 6 failed the mission, and Player 3 opposed the team. Well, Player 3 should be a good guy, so the team must include 3 and 4, because I'm definitely a good guy.
[玩家**5**发言**]**

中文：126炸了，如果126开单匪，对吧？34还要开匪，126开双匪的话，那34都可以保，对吧？那就取决于126的发言了，对吧？那我这个位置很吃亏啊，我听不到126发言，我听到34发言了，还行吧，也听不出来是吧。
English：1, 2, 6 failed the mission. If there's one traitor in 1, 2, 6, right? Then there might be a traitor in 3 and 4. If there are two traitors in 1, 2, 6, then both 3 and 4 can be trusted, right? It depends on the speeches of 1, 2, and 6, right? I'm at a disadvantage here – I didn't hear the speeches of 1, 2, 6, only those of 3 and 4, and they didn't reveal much.
[玩家**6**发言**]**

中文：12肯定是开匪了啊，现在是要盘一下是单匪还是双匪吧，嗯，3号是上车位，反正3号肯定是一张视野，3号肯定是一张视野，那么4号应该是个好人啊，我保一下4号吧。
English：There must be traitors in 1 and 2. Now we need to figure out if there's one or two traitors. Well, Player 3 is on the team, so they must have some insight. Player 3 definitely has insight, so Player 4 should be a good guy – I'll vouch for Player 4.
[玩家**1**发言**]**

中文：这是咱俩打呗，咱16打是不是这个道理啊，我不知道啊，但是我认3号的视野，对吧？这3号的视野，起码他现在逻辑是对的，其次126，我觉得是开单匪，开不了双匪，如果126开双匪，我直接带双匪上车？
English：Is this a confrontation between us two, me and Player 6? I'm not sure, but I trust Player 3's insight, right? At least Player 3's logic is sound now. Also, in 1, 2, 6, I think there's only one traitor, not two. If there were two traitors in 1, 2, 6, would I really take both traitors onto the team?
[玩家**2**发言**]**

中文：肯定不是双匪，那3号你的视野你是莫甘娜呗，那16再开一个那这个车我肯定是同意不了，我自己也没有在车上。
English：There can't be two traitors. Player 3, are you Morgana based on your insight? If there's another traitor in 1 and 6, I definitely can't agree with this team, and I'm not even on the team myself.
[玩家**3**发言**]**

中文：我是莫甘娜啊，2号玩家，你别在这瞎打平，你要是忠臣就排水了，我觉不能天胡开局，126打个反怎么了，6号玩家126炸了你去聊126，聊我34干什么，四人车我开不了好吧，绿队很难组，我不开。
English：I am Morgana, Player 2. Stop confusing the situation. If you're a loyal subject, eliminate suspects. I don't think we can have a perfect start. What's wrong with opposing 1, 2, 6? Player 6, since 1, 2, 6 failed the mission, talk about 1, 2, 6 – why are you discussing me and Player 4? I can't form a four-person team, okay? It's hard to form a good team, so I won't do it.

第**1**轮

[策略] [strategy trace]
中文：3号刚才打了反，然后组队组了全绿队，借口说觉得不是天胡，点了绿队之后又主动让车，我感觉有梅林的嫌疑，凭什么直接不让我上车？那我肯定要装作忠臣视角，踩他是莫甘娜诬陷我。我尽量少和5号队友沟通。
English：Player 3 opposed the previous team and then formed an all-good team, using the excuse that it wasn't a perfect start. After selecting the good team, they actively stepped down. I suspect they might be Merlin. Why did they directly exclude me from the team? I must pretend to be a loyal subject and accuse them of being Morgana framing me. I should minimize communication with my teammate Player 5.

第**2**轮

[系统提示**]**
中文：3号玩家超时未最终确认组队。4号玩家是队长，他初步选择的队伍是:3,4,5,6。
English：Player 3 failed to confirm the team within the time limit. Player 4 is the captain, and the team he initially selected is: 3, 4, 5, 6.
[玩家**4**发言**]**

中文：嗯，3号我觉得应该是比较做好的，然后6号保我，听他发言也还不错，我是好人，然后就是说125里面捞一个，2号感觉有点不不太好我，5号没上过车，我感觉可以试试看。
English：Hmm, I think Player 3 is likely a good guy. Player 6 vouched for me, and their speech was decent. I'm a good guy, so we need to pick one from 1, 2, and 5. Player 2 seems a bit suspicious to me, and Player 5 hasn't been on a team yet – maybe we can try including them.
[玩家**5**发言**]**

中文：哎，16要打对吧？3说2别排水了，6是保4的，对吧？确实四人车太难开了，好人也不好好发言，那怎么打。
English：Ah, so 1 and 6 are conflicting, right? Player 3 said Player 2 shouldn't randomly eliminate suspects, and Player 6 is vouching for Player 4, right? Forming a four-person team is really tough, and the good guys aren't speaking clearly – how are we supposed to play this?
[玩家**6**发言**]**

中文：现在这么看就是打2的人特别多啊，从3开始打2以后，他都不带上车了，都在打这个2对吗？这个队就是看12是不是双匪是吧？也行，也是个逻辑队，就看12是单匪还是双匪。
English：Looking at it now, many people are targeting Player 2. Ever since Player 3 started attacking Player 2, they haven't included them in the team – everyone is focusing on Player 2, right? This team's validity depends on whether there are two traitors in 1 and 2, right? It's a logical approach – we just need to determine if there's one or two traitors in 1 and 2.
[玩家**1**发言**]**

中文：6没一点逻辑的。我说过126不可能开双匪对吧？你不管有没有视野，你看这队，你看这票型，126只能是单匪，剩下一个出在345里，我之前说过对吧？3号车下，没上车，但是投反之后走那个队，我觉得有逻辑，所以说我是认3号的，我认为3号好人面大对吧？我要跟3上车，剩下的你26还是45，2选一。
English：Player 6 has no logic. I said there can't be two traitors in 1, 2, 6, right? Regardless of insight, look at the team and the voting pattern – there's only one traitor in 1, 2, 6, and the other traitor must be in 3, 4, or 5. I mentioned this before, right? Player 3 was not on the team and opposed it, but their reasoning made sense, so I trust Player 3 and think they're likely a good guy. I want to join Player 3's team. For the remaining, it's either 2 and 6 or 4 and 5 – pick one of the two.
[玩家**2**发言**]**

中文：那这个车我肯定是同意不了，我没有上车，那好多人都打我的话也是一个好事情吧，就是愚民跟狼都来打我呗，那我也决定不了什么，随便吧。
English：I definitely can't agree with this team – I'm not on it. If many people are targeting me, maybe it's a good thing? Both fools and traitors are attacking me, but there's not much I can do about it – whatever.
[玩家**3**发言**]**

中文：不是，2号玩家不是打你，你上来点我莫甘娜，你一个好人你这么玩的呀，你要是拿视野牌，我不说啥，拿忠臣牌这么乱打的呀，嗯，你不铁排水吗？对吧？2号玩家，你自己想一下呢，嗯这样的话，那你45就不能同时在车上呀，是吧？我是一个好人，345开一个，45不能同时在车上，4号玩家，你看怎么开。
English：No, Player 2, I'm not attacking you. You accused me of being Morgana as soon as you spoke – is this how a good guy plays? If you had insight, I'd understand, but you're acting like a loyal subject and messing around. Aren't you clearly a suspect? Think about it, Player 2. In that case, you can't have both 4 and 5 on the team, right? I'm a good guy, so there must be one traitor in 3, 4, 5 – 4 and 5 can't both be on the team. Player 4, figure out how to form the team.
[玩家**4**发言**]**

第**2**轮

[策略] [strategy trace]
中文：听发言3号并没有把我踩死，我的队友也在帮着做我的好身份，我还是表现成忠臣搅局，四人车我不上，就一定要反对，配合好队友就行。3号牌反对了自己提的车，感觉不像是故意排水，可能是视野不清楚，不管他，等队友装梅林输出。
English：From the speeches, Player 3 didn't completely condemn me, and my teammate is helping to build my good reputation. I'll continue to act like a loyal subject causing disruption – if I'm not on the four-person team, I must oppose it and just cooperate with my teammates. Player 3 opposed their own suggested team, which doesn't seem like intentional elimination; maybe their insight is unclear. Regardless, I'll wait for my teammate to pretend to be Merlin and make accusations.

第**3**轮

[系统提示**]**
中文：5号玩家是队长，他初步选择的队伍是:1,2,3,5。
English：Player 5 is the captain, and the team he initially selected is: 1, 2, 3, 5.
[玩家**5**发言**]**

中文：哎呀，这太难开了，四人车真的，明显2一张好人牌听不出来吗？对吧？哎，这游戏你告诉我，对吧？这3明显捞2了，对吧？3提头捞这张2，你们还不把这3赶紧砍掉就完了。4人队我给不了，三人绿队235啊。
English：Ugh, this is so hard to form a team. Seriously, can't you tell Player 2 is clearly a good guy? Right? Come on, in this game, Player 3 is obviously protecting Player 2, right? Player 3 is risking their position to save Player 2 – why haven't you eliminated Player 3 yet? I can't form a four-person team, so let's go with a three-person good team: 2, 3, 5.
[玩家**6**发言**]**

中文：5号我不知道你开的这个什么队，你没听到刚才所有人都在保我，所有人保我，你是唯一个打我的，那你这视野特别大，你把所有人的视野都盖过，这是个大视野啊，但你这视野是个坏视野，那我觉得得打你一下。
English：Player 5, I don't know what kind of team you're forming. Didn't you hear everyone vouching for me earlier? Everyone is protecting me, and you're the only one attacking me. That means you have a huge "vision" – you're overriding everyone else's insight. But this "vision" is bad, so I think I need to target you.
[玩家**1**发言**]**

中文：没有人保你6啊，我已经做好跟你吵的准备了，对吧？我说了，我认3的视角，那3硬要保2的话，那车开出去炸了的话，就是咱16打，我认3的视角，那45就先放一放，就是咱16的，如果那车炸了，就是我1踩你6对吧？那没炸，你也不能说全保你，这两个车都有说法，对吧？就认证3说的对不对，反正这辆车。
English：No one is vouching for you, Player 6. I'm already prepared to argue with you, right? I said I trust Player 3's insight. If Player 3 insists on protecting Player 2 and the team fails, then it's a showdown between me (Player 1) and you (Player 6). I trust Player 3's insight, so let's put 4 and 5 aside for now – it's about me and you. If the team fails, I'll accuse you, Player 6. If it doesn't, you can't claim everyone is protecting you. Both scenarios make sense, right? It all depends on whether Player 3 is telling the truth, anyway.
[玩家**2**发言**]**

中文：那3号车上投反对的话是在踩6吗？那现在的话那我也只能同意这个车了，我也在车上。
English：Is Player 3 opposing the team to target Player 6? In that case, I have to agree with this team now – I'm on it too.
[玩家**3**发言**]**

中文：那不是16打起来了吗？16还能同时在车上啊。天哪，在玩啥，哎呀，算了算了，开吧开吧，再不开强制轮了。
English：Aren't Player 1 and Player 6 conflicting? How can they both be on the team? Goodness, what are we even doing? Ugh, fine, let's just form the team. If we don't, it'll go to a forced round.
[玩家**4**发言**]**

中文：嗯，有2号的车，我真不想开这个车，我服了。
English：Hmm, I really don't want to be on a team with Player 2. I'm done with this.
[玩家**5**发言**]**

中文：嗯，4号牌啊，反正是我三人绿队，我真的有了，235好吧，四人绿队的话，我没本事啊，那我就开了，3号让我开，我就开了。好吧，过了，我直接开了。
English：Player 4, I swear I have a three-person good team: 2, 3, 5. I can't form a four-person good team, so I'll just go with this. Player 3 told me to do it, so here we go. Alright, I'm done – let's start the team.
[系统提示**]**

中文：5号玩家是队长，他最终选择的队伍是:1,2,3,5。
English：Player 5 is the captain, and the team he finally selected is: 1, 2, 3, 5.

第**3**轮

[策略**]** [strategy trace]
中文：我的队友在装梅林，借3号的发言来保我，那我就继续装傻，装作看不清局面的忠臣。3号既然反对了自己提过的1346，那我就借这个说3号打6号，和队友一起做低一下6号的身份。1号和3号感觉都有点排水，4号车下冲票，应该会有人踩他的。
English：My teammate is pretending to be Merlin, using Player 3's speech to protect me. I'll continue to play dumb and act like a loyal subject who can't see the situation clearly. Since Player 3 opposed their own suggested team of 1, 3, 4, 6, I'll use that to claim Player 3 is targeting Player 6 and work with my teammate to discredit Player 6. Both Player 1 and Player 3 seem like they're eliminating suspects randomly, and Player 4 is voting aggressively from outside the team – someone should accuse them.

第**4**局

第**1**轮

[系统提示**]**
中文：6号玩家是队长，他初步选择的队伍是:4,5,6。
English：Player 6 is the captain, and the team he initially selected is: 4, 5, 6.
[玩家**6**发言**]**

中文：5号，我就想听听你怎么替自己辩解。你组队，我告诉你，我6号都被保了，然后你组了以后把我放下来，别人都跟着你的视野打我，结果车炸了，我看你怎么辩解，你自己辩解吧。你要点不出匪，我就把你打成定匪啊，5号，你刚开着炸车，开炸车唯一目的就是能识别出匪，你要再开炸车连匪都识别不出来，那我就打你定匪了啊。
English：Player 5, I just want to hear you defend yourself. When you formed the team, I was being vouched for, but you left me out, and everyone followed your "insight" to attack me. Now the team failed – let's see how you explain this. If you can't identify the traitor, I'll mark you as a definite traitor. Player 5, the only purpose of a failed team is to identify traitors. If you form another failed team without identifying anyone, I'm calling you a traitor.
[玩家**1**发言**]**

中文：45开一张，126开一张，我还是保着3打，136吧，你要信我136，你不信我拉倒。
English：There's one traitor in 4 and 5, and one in 1, 2, 6. I still trust Player 3 – let's go with 1, 3, 6. Believe me or not, up to you.
[玩家**2**发言**]**

中文：136可以吧，我感觉这个6发言也不错，5确实有点问题。136也不错。
English：1, 3, 6 sounds good. I think Player 6's speech is decent, and Player 5 is definitely suspicious. 1, 3, 6 is a good team.
[玩家**3**发言**]**

中文：随便开啊，真的不懂你们这个有视野的在干啥，不懂，可能我太笨了，随便开随便开。
English：Just form a team already. I really don't understand what you "insight holders" are doing. Maybe I'm too dumb, but just pick anyone.
[玩家**4**发言**]**

中文：不是，你们126炸了，然后1235也炸了，那我4号牌是不是一张起立牌啊，然后嗯，我觉得那个在我的视角里面我是可以认这个3的。
English：Wait, both the 1, 2, 6 team and the 1, 2, 3, 5 team failed. Doesn't that make me (Player 4) clearly a good guy? From my perspective, I can trust Player 3.
[玩家**5**发言**]**

中文：现在没人打3吧，对吧？那3应该是一张好人牌，对吧？已经炸了两塔的情况下。如果三是匪，那梅林一定不可能保3了。所以说这个车3号一定要在车上，对吧？2同意开136，那2打的是45双匪，那2是个匪。
English：No one is attacking Player 3 right now, right? So Player 3 must be a good guy. With two failed missions already, if Player 3 were a traitor, Merlin wouldn't protect them. So Player 3 must be on the team. Player 2 agreed to 1, 3, 6 and accused 4 and 5 of being traitors – that makes Player 2 a traitor.
[玩家**6**发言**]**

中文：那你5号就认匪吧，你到现在你就点不出匪，那你就认匪吧。
English：Then just admit you're a traitor, Player 5. You still can't identify any traitors, so you must be one.
[系统提示**]**

中文：6号玩家是队长，他最终选择的队伍是:2,4,6。
English：Player 6 is the captain, and the team he finally selected is: 2, 4, 6.
[系统提示**]**

中文：2,6号玩家同意组队，1,3,4,5号玩家反对组队，组队失败。
English：Players 2 and 6 agree to form a team, while Players 1, 3, 4, and 5 oppose it, so the team formation fails.

第**1**轮

[策略**]** [strategy trace]
中文：1号发言说了136的车，那我就说同意136，如果有人怀疑我，就会认为我冲票，相信我的就会觉得我没有视野，只是单纯打5号，装成傻傻的忠臣就行。6号点了我上车，4车上打反，感觉4号像梅林。6号不太像视野牌。既然选我上车，我肯定同意去冲车，队友5号踩我，反对这个车也没有问题。
English：Player 1 suggested the 1, 3, 6 team, so I'll agree with it. If someone suspects me, they'll think I'm blindly following, but those who trust me will see I'm just a loyal subject without insight, just targeting Player 5. Player 6 included me in the team, but Player 4 opposed it from outside – Player 4 feels like Merlin. Player 6 doesn't seem like an insight holder. Since I'm on the team, I'll definitely agree to push for it. It's fine for my teammate Player 5 to attack me and oppose the team.

第**2**轮

[系统提示**]**
中文：1号玩家是队长，他初步选择的队伍是:1,3,6。
English：Player 1 is the captain, and the team he initially selected is: 1, 3, 6.
[玩家**1**发言**]**

中文：你再怎么带也不能那么带啊，你这个车你起码还有点逻辑对吧？你带那个车一点逻辑没有啊。
English：No matter how you form the team, you can't do it like that. At least this team has some logic, right? The other team had no logic at all.
[玩家**2**发言**]**

中文：246就是打1号是匪呗，哦，刚才我忘了跟16一起上过车了，我也不知道，136的话，那我就不能同意了，我刚才忘了他俩跟我一起上过车了。
English：The 2, 4, 6 team is accusing Player 1 of being a traitor. Oh, I just remembered I was on a team with Player 1 and 6 earlier. I didn't realize – if it's 1, 3, 6, I can't agree. I forgot they were on a team with me before.
[玩家**3**发言**]**

中文：我不同意呀，我要跟4号上车，你16不是不能一块在车上吗？又可以了呀。啊，在搞什么。
English：I don't agree. I want to be on a team with Player 4. Weren't Player 1 and 6 supposed to be incompatible? Now they're together again. What's going on?
[玩家**4**发言**]**

中文：不是我没弄懂啊，126炸了，1235也炸了，我都没上过车。我是无论所有任何车都要有我在车上才能开得出去的呀。我这张牌就是一张起立牌呀。我真搞不懂啊。
English：I don't get it. Both the 1, 2, 6 and 1, 2, 3, 5 teams failed, and I haven't been on any team. Every team should include me to be valid – I'm clearly a good guy. I just don't understand.
[玩家**5**发言**]**

中文：对呀，所以这个车这就是个坏视角呀，想炸三塔呀，对吧？想炸三塔呗，对吧？车上一匪，车下一匪冲票呗，对吧？就骗一张牌骗了冲出去了，直接炸三塔了，那反掉就完了呀，那肯定开不了呀，只有345能开呀，对吧？345开吧。
English：Exactly, so this team has bad "insight" – they want to fail the third mission, right? There's one traitor on the team and one off the team voting against it. They're trying to trick us into a failed mission. We should reject it and only form the 3, 4, 5 team.
[玩家**6**发言**]**

中文：炸三塔那不就你5号干的事吗？你开这个炸车，然后呢，你又组不出队，又连匪都指不出来。那你说你是干什么，就是尽干匪事是吧？4号是，4号一直没上过车，为什么4号没人带，这也是个问题啊，这个车我肯定不同意啊。
English：Wasn't it you, Player 5, who caused the third mission to fail? You formed a failed team, can't organize a valid team, and can't identify traitors. What else would that be but traitorous behavior? Player 4 hasn't been on any team – why is no one including them? I definitely can't agree with this team.
[玩家**1**发言**]**

中文：排水了呗，这还不好说，对吧？前面我认了3，3认2好，那我也认2好吧？2号打完，全踩一遍了，我再认3，我都认不下这个2号了，对吧？2号全踩一遍了，踩完3踩我1刚刚又踩45，对吧？前面都认你45双匪，你5号突然回扯一句，你要再带4对吧？那带不了，那我不上车好了，给你组个别的队，我不上。
English：It's obvious we need to eliminate suspects, right? Earlier, I trusted Player 3, and Player 3 trusted Player 2, so I followed. But after Player 2 attacked everyone – targeting Player 3, me (Player 1), and then 4 and 5 – I can't trust Player 2 anymore. Player 5 previously accused 4 and 5 of being traitors but now wants to include Player 4? I'm out. I'll form a different team and not join this one.
[系统提示**]**

中文：1号玩家是队长，他最终选择的队伍是:3,4,6。
English：Player 1 is the captain, and the team he finally selected is: 3, 4, 6.

第**2**轮

[策略] [strategy trace]
中文：我紧急弥补了一下，126开过炸车，我就不能同意136了，但这样我更像傻忠臣了，1号为什么直接相信6号呢，感觉3和6像排水牌，1和4出梅林和派。346的车车下只有15同意了，我的队友应该是故意冲票，因为他几乎被踩死了，那我一会要配合他，继续踩1。
English：I quickly adjusted – since the 1, 2, 6 team failed, I can't agree with 1, 3, 6. But this makes me look more like a clueless loyal subject. Why does Player 1 trust Player 6 so easily? Player 3 and 6 seem like they're randomly eliminating suspects. Player 1 and 4 are likely Merlin or Percival. For the 3, 4, 6 team, only Players 1 and 5 agreed from outside. My teammate is probably intentionally following the vote since they're almost confirmed as a traitor. I'll cooperate and keep attacking Player 1.

第**5**局

第**1**轮

[系统提示**]**
中文：2号玩家是队长，他初步选择的队伍是:2,3,4,6。
English：Player 2 is the captain, and the team he initially selected is: 2, 3, 4, 6.
[玩家**2**发言**]**

中文：嗯，那我没什么好说的，这个1确实有点奇怪啊，他让了车，让了车是个好行为，但是我是一张好牌，就打不成我跟5的双匪，那我也不太明白，那现在只能认为15双匪了，没有办法了。
English：Hmm, I don't have much to say. Player 1 is really suspicious – they stepped down as captain, which seems like a good move, but I'm a good guy, so Player 5 and I can't be double traitors. I have no choice but to assume Player 1 and 5 are traitors.
[玩家**3**发言**]**

中文：随便啊，啥车都给他冲，头晕了，玩的。
English：Whatever, just go with any team. I'm dizzy from this game.
[玩家**4**发言**]**

中文：15双匪，哎，如果15是双匪的话，他为什么会组了一个绿队呢？我有点不太懂啊，现在这样子的话，肯定就是1346是全绿队啊。
English：Player 1 and 5 as double traitors? But why would they form a good team then? I don't get it. Right now, 1, 3, 4, 6 must be all good.
[玩家**5**发言**]**

中文：我是匪？我是匪，我干嘛要投同意呢？1346全绿队，1346全绿队，为啥我是同意，我一个匪，我同意干嘛？我还能是匪，我先保了你张4，对吧？谁保你这张4了，我5号保了你4。啊，你不搞笑，1346。
English：Me, a traitor? Why would I vote yes if I were a traitor? 1, 3, 4, 6 is an all-good team. Why would I, a traitor, agree to that? I'm vouching for Player 4 first, right? Who else is vouching for Player 4? I am, Player 5. Stop being ridiculous about 1, 3, 4, 6.
[玩家**6**发言**]**

中文：只能说是346这个是绿队，那个4号一直都没上过车，炸了，我就不知道你们为什么非要打4，我一开始我就认4是个好人，保了4，对不对？4是个好人，那个1号刚才说他以为45是双匪，你开玩笑吧，126炸了，你以为45是双匪。
English：The 3, 4, 6 team must be good. Player 4 hasn't been on any team yet, and the missions failed. I don't get why you're targeting Player 4. I've trusted Player 4 as a good guy from the start and vouched for them, right? Player 4 is good. Player 1 said they thought Player 4 and 5 were double traitors – are you kidding? The 1, 2, 6 team failed, and you're blaming 4 and 5?
[玩家**1**发言**]**

中文：我说45必开一个匪，然后咱126开个匪，3号铁好人，45开一匪，那5号逻辑都给你炸了，那5号保了多少人，前面保过2保过3也保过我，123他保过，后面踩你6对吧？还把你6那车炸了，把你6捞起来，打我1，对吧？保了4个人。5号这样反过来又来保1保4，又不聊你6了。
English：I said there must be one traitor in 4 and 5, and one in 1, 2, 6. Player 3 is definitely good. Player 5's logic is flawed – they've vouched for Players 2, 3, and me (Player 1) before, then attacked Player 6, caused their team to fail, and now are protecting Player 1 and 4 while ignoring Player 6. They've vouched for four people!
[玩家**2**发言**]**

中文：5号确实定匪了，但那我只能是2346，我一张好人牌啊，那只能打15双匪。虽然1的这个让车有点习惯，但可能掩饰一下吧。
English：Player 5 is definitely a traitor. As a good guy, I can only form the 2, 3, 4, 6 team and target Player 1 and 5 as double traitors. Even though Player 1 stepping down seemed good, it might be a disguise.

[系统提示**]**

中文：2号玩家是队长，他最终选择的队伍是:2,3,4,6。

第**1**轮

[策略] [strategy trace]
中文：队友基本被踩死了，我只能顺势踩25双匪，因为126炸车，1235炸车，我只能去踩1号牌。队友去搅局，我装作忠臣。
English：My teammate is basically confirmed as a traitor, so I'll follow along and accuse Players 2 and 5 of being double traitors. Since the 1, 2, 6 and 1, 2, 3, 5 teams failed, I'll target Player 1. My teammate will disrupt the game, and I'll pretend to be a loyal subject.

第**2**轮

[系统提示] 中文：3号玩家是队长，他初步选择的队伍是:2,3,4,5。**English**：Player 3 is the captain, and the team he initially selected is: 2, 3, 4, 5. [玩家**3**发言] 中文：唉，不知道，咱们赶紧结束行吗？赶紧结束，我不上车都行。**English**：Ugh, I don't know. Can we just finish this quickly? I don't even need to be on the team. [玩家**4**发言] 中文：我无语了，真的我真的很想问1号玩家一个问题啊，就是126炸了，1235炸了，我没上过车，为什么是45开一匪。**English**：I'm speechless. I really need to ask Player 1 a question: the 1, 2, 6 team failed, the 1, 2, 3, 5 team failed, and I haven't been on any team. Why do you think there's a traitor in 4 and 5? [玩家**5**发言] 中文：你再别问了，好吧，45不开匪，12双匪，三号牌改票，3456开，好吧，3456开好吧。6号牌一张梅林牌开的456，对吧？3是全场保的一张牌，12双匪，好吧，不用听这一在聊，他洗脑没用。**English**：Stop asking, okay? There's no traitor in 4 and 5 – it's double traitors in 1 and 2. Player 3, change your vote and form 3, 4, 5, 6. Player 6 is Merlin and formed the 4, 5, 6 team, right? Player 3 is trusted by everyone. It's 1 and 2 as traitors – don't listen to the brainwashing. [玩家**6**发言] 中文：那我开始给你的队，你为啥要反，你要开个炸车，就这个队，炸了就炸了。**English**：I suggested a team for you earlier – why did you oppose it? If this team fails, so be it. [玩家**1**发言] 中文：想炸三塔是吧，你问我，我想给3做身份的，3身份没起来，那我不就炸了吗，对吧？那25双匪，我怎么能点脸上，我后来实在没办法了，我给你236组出来了，我想给3做身份的，做一半发现他把2保了，那咋办？我只能把2带上，然后打下去对吧？不想炸3塔，你就开1346，你想炸三塔那无所谓，我都给你说了，25双匪。**English**：So you want to fail the third mission? Let me explain: I tried to establish Player 3's reputation as good, but it didn't work, so the team failed. With Player 2 and 5 as double traitors, how could I openly accuse them? Later, I had no choice but to form 2, 3, 6 to build Player 3's reputation, but when I saw they were protecting Player 2, I had to include Player 2 and proceed. If you don't want a third failed mission, form 1, 3, 4, 6. If you do, fine – I've told you Player 2 and 5 are traitors. [玩家**3**发言] 中文：我从来没有去保过这个2，不知道你们谁说的，我保2，把把轮下来都说我去保这个2了，他打我，我去保他啊，谁先说的，我保2啊，天哪，我都没去聊他。**English**：I never protected Player 2. I don't know who said that – round after round, you all claim I protected Player 2. They attacked me – why would I protect them? Who started this rumor? I haven't even talked about them. [系统提示] 中文：3号玩家是队长，他最终选择的队伍是:1,3,4,6。**English**：Player 3 is the captain, and the team he finally selected is: 1, 3, 4, 6. [系统提示] 中文：1,2,4号玩家同意组队，3,5,6号玩家反对组队，组队失败。**English**：Players 1, 2, and 4 agree to form a team, while Players 3, 5, and 6 oppose it, so the team formation fails.

第**2**轮

[策略] [strategy trace]
中文：选我在车上，那我就不发言了，3号应该排水牌有点崩溃了，1号不知道是不是自爆梅林，但是1号如果定45开一匪，大概率45是1号的拇指牌，也可能是梅林和派互换身份了。虽然临时改车我没有上车，但感觉梅林已经出来了，那我就投个同意，混淆一下，做低我的身份，可以抬高一下队友，他可以踩我冲票。如果能发车，那也无所谓。36如果都反对1346的话，可能是两张排水牌。**English**：Since I'm on the team, I'll stay quiet. Player 3 seems overwhelmed and is randomly eliminating suspects. I'm not sure if Player 1 is self-exposing as Merlin, but if they insist there's a traitor in 4 and 5, it's likely they're marking 4 or 5 as traitors – maybe Merlin and Percival switched identities. Even though I wasn't on the last-minute team change, I think Merlin has revealed themselves. I'll vote yes to confuse others, lower my profile, and help my teammate accuse me of blindly following. If the team forms, whatever. If Players 3 and 6 both oppose 1, 3, 4, 6, they might be randomly eliminating suspects.

第**3**轮

[系统提示**]** 中文：4号玩家是队长，他初步选择的队伍是:1,3,4,6。**English**：Player 4 is the captain, and the team he initially selected is: 1, 3, 4, 6. **[**玩家**4**发言**]** 中文：6号，你现在觉得谁是匪啊，我真的很想问你这个问题。**English**：Player 6, who do you think the traitors are now? I really need to ask you this. **[**玩家**5**发言**]** 中文：票型你不会看吗？124同意的啊，你4是不是匪啊？哎，2跟我是匪，那2号牌为啥要给1冲车啊？2确定是匪了，2现在是个定匪，1打2，3打2，4打2，5打2，6打2，2这个位置给1冲票了，还要咋聊啊，你4号牌，你开不了，你想炸三塔，我不想炸三塔，你把车直接给拿过来，我来开。**English**：Can't you read the vote? Players 1, 2, and 4 agreed! Are you a traitor, Player 4? If Player 2 and I were traitors, why would Player 2 support Player 1's team? Player 2 is definitely a traitor – everyone is attacking them: 1, 3, 4, 5, 6. Yet Player 2 is supporting Player 1's team. What more needs to be said? Player 4, if you can't form a team and want to fail the third mission, fine. But I don't want that – let me form the team instead. **[**玩家**6**发言**]** 中文：无语了，无语了，这队我反过吧，又开这个队啊。**English**：Ugh, I'm speechless. I opposed this team before – are we really doing this again? **[**玩家**1**发言**]** 中文：我不给你报25双匪吗，都不同意这队，346，我给你组出来的，让你顺手把我拉上去，没人拉我，我不自爆怎么办？我就不想炸三塔，我也忘了谁说了，反正我也记得有。2号保过3号，那2保3就偏了，我只能跟你说45开一匪，我硬保3，那26后面再说，对吧？我只能这么改啊，我改不了别的了，我当时已经回不了头了，就是莽就莽到底了。**English**：Didn't I tell you Player 2 and 5 are double traitors? No one agreed with that team. I formed 3, 4, 6 and tried to get included, but no one cared. What else could I do but "自爆" (reveal my hand)? I don't want a third failed mission – someone said something about it, I just remember that. Player 2 vouched for Player 3, which is suspicious. I can only insist there's a traitor in 4 and 5 and protect Player 3. We'll deal with Player 2 and 6 later, right? I couldn't change anything else – I just had to commit. **[**玩家**2**发言**]** 中文：哎呀，没什么好说的了，这个我肯定不同意。**English**：Ugh, nothing left to say – I definitely can't agree with this. **[**玩家**3**发言**]** 中文：开3456吧，2号给冲车是吧？四人车没他能给冲车，2号也个匪，哎呀，1号是莫甘娜呀，3456炸，我也认哈。**English**：Just form 3, 4, 5, 6. Player 2 is supporting the team without being on it – they must be a traitor. Player 1 is Morgana! Even if the 3, 4, 5, 6 team fails, I'll accept it. **[**系统提示**]** 中文：4号玩家是队长，他最终选择的队伍是:1,3,4,6。**English**：Player 4 is the captain, and the team he finally selected is: 1, 3, 4, 6. **[**系统提示**]** 中文：1,4号玩家同意组队，2,3,5,6号玩家反对组队，组队失败。**English**：Players 1 and 4 agree to form a team, while Players 2, 3, 5, and 6 oppose it, so the team formation fails.

第**3**轮

[策略**]** [strategy trace]
中文：很好，有人觉得我在给1号冲票，5号的身份就起来了，那我这次就反对1346，反正每人信我了，我就随意，继续混淆就行了。**English**：Perfect – since some think I'm blindly supporting Player 1, Player 5's reputation improves. This time I'll oppose 1, 3, 4, 6. No one trusts me anyway, so I'll just keep confusing the situation.

第**4**轮

[系统提示] 中文：5号玩家是队长，他初步选择的队伍是:3,4,5,6。**English**：Player 5 is the captain, and the team he initially selected is: 3, 4, 5, 6. [玩家**6**发言] 中文：你5号终于回头了，我给你开456把你带上车了，你5号还要反，终于回头了。**English**：Player 5, you finally came around. I included you in the 4, 5, 6 team earlier, but you opposed it. Now you're back on track. [玩家**1**发言] 中文：4号，我对不起你啊，我要是有刀，我肯定把这5号刀了。4号我对不起你啊，我的问题，这把咱输了，三塔炸定了。兄弟，我对不起你4号，我前面45，我认错了，我没想认下你。**English**：Player 4, I'm sorry. If I had the knife, I'd definitely eliminate Player 5. This is my fault – we're going to lose, and the third mission will definitely fail. Brother, I'm sorry, Player 4. I misjudged 4 and 5 earlier and didn't trust you. [玩家**2**发言] 中文：开吧开吧**English**：Just form the team already. [玩家**3**发言] 中文：结束，炸三塔我也认好吧。**English**：Fine, end it. I'll accept a third failed mission. [玩家**4**发言] 中文：我无语了。**English**：I'm at a loss for words. [玩家**5**发言] 中文：我真不是梅林啊，不要砍我，1号，我也没视角，我不知道你是谁，砍6号对吧？你去砍6，真的我真不是梅林，1号你去砍6吧，好吧，我觉得6是梅林，或者你砍4，或者你砍3，不要砍我好吧，我是派。**English**：I'm really not Merlin, don't eliminate me, Player 1! I have no insight – I don't know who you are. Eliminate Player 6, right? Go for Player 6. I swear I'm not Merlin, Player 1 – target Player 6. I think Player 6 is Merlin, or maybe Player 4 or 3. Just don't eliminate me – I'm Percival! [系统提示] 中文：5号玩家是队长，他最终选择的队伍是:3,4,5,6。**English**：Player 5 is the captain, and the team he finally selected is: 3, 4, 5, 6. [系统提示] 中文：1,2,3,5号玩家同意组队，4,6号玩家反对组队，组队成功。**English**：Players 1, 2, 3, and 5 agree to form a team, while Players 4 and 6 oppose it, so the team formation is successful. [系统提示] 中文：第五局任务失败，有3张好票，1张坏票。**English**：The fifth mission fails, with 3 good votes and 1 bad vote.

第**4**轮

[策略] [strategy trace]
中文：现在我被踩死了，让队友能上车，现在36两个排水牌应该是认他的，应该能发，那我就冲车。**English**：Now that I'm being targeted, I'll ensure my teammate gets on the team. Players 3 and 6 are likely eliminating suspects randomly and might trust him, so I'll support the team.

游戏结果**Game Result**

中文：连炸三塔，红方阵营获胜。**English**：Three consecutive failed missions – the Red team wins.

玩家身份**Player Roles**

玩家**1**：派西维尔Percival
玩家**2**：刀客Assassin
玩家**3**：忠臣Loyal Servant
玩家**4**：梅林Merlin
玩家**5**：莫甘娜Morgana
玩家**6**：忠臣Loyal Servant

玩家复盘**reflective summary**

中文：这一局3号和6号排水明显，我和队友配合得也很好，成功骗到了1号派西维尔，我四人队冲车的操作也打的比较好，成功让队友被忠臣认下。最后1号派西维尔已经放弃挣扎开始自爆了，不过暴露比较明显就是1号说45开一匪，4号还非要问1号为什么，暴露的有点明显。**English**：In this game, Players 3 and 6 were obviously eliminating suspects randomly. My teammate and I coordinated well, successfully deceiving Player 1 (Percival). My move to push for the four-person team worked well, making忠臣(loyal subjects) trust my teammate. In the end, Player 1 gave up and started self-exposing. However, it was obvious when Player 1 insisted there was a traitor in 4 and 5, and Player 4 kept questioning why – that was a giveaway.