Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies

Anonymous ACL submission

Abstract

Social deduction games like Werewolf combine language, reasoning, and strategy, providing a testbed for studying language and social intelligence. However, most studies reduce the game to LLM-based self-play, yielding templated utterances and anecdotal cases that overlook the richness of social gameplay. Evaluation further relies on coarse metrics such as survival time or subjective scoring due to the lack of quality reference data. To address these gaps, we curate a high-quality, human-verified, multilingual, and multimodal Werewolf dataset containing over 100 hours of video, 32.4M utterance tokens, and 15 rule variants. Based on this dataset, we propose a novel strategy-alignment evaluation that leverages the winning faction's strategies as ground truth in two stages: 1) Speech evaluation, formulated as multiple-choice-style tasks that assess whether the model can adopt appropriate stances across five dimensions of social ability. 2) Decision evaluation, which assesses the model's voting choices and opponent-role inferences. This framework enables a fine-grained evaluation of models' linguistic and reasoning capabilities, while capturing their ability to generate strategically coherent gameplay. Our experiments show that state-of-the-art LLMs and VLMs exhibit diverse performance, with roughly half of the models in both modalities remaining below 0.50 accuracy, revealing clear gaps in deception and counterfactual reasoning. We release the dataset here.

1 Introduction

Social deduction games provide a unique setting for studying language and intelligence. Unlike strategic dialogue games such as Diplomacy (Bakhtin et al., 2022) which requires long-term alliance building and coordinated planning, Werewolf (Xu et al., 2023) places greater emphasis on the artistry of language and reasoning. Players must, within limited speaking time, influence others' judgments and votes through persuasion, ambiguity, deception, and rhetorical nuance. These characteristics make social deduction games like Werewolf a particularly challenging and valuable domain for artificial intelligence research.

Recent advances in Large Language Models (LLMs) have motivated a growing line of research exploring their application to Werewolf and other social deduction games. However, most existing studies adopt a self-play paradigm where multiple LLM agents interact only with each other (Wu et al., 2024; Xu et al., 2024; Du and Zhang, 2024; Poglitsch et al., 2025; Xu et al., 2025; Song et al., 2025b). This often results in overly rigid and templated utterances, as well as dull or anecdotal storylines. For example, as shown in Figure 1(a) (Xu et al., 2024), the model produced shallow, formulaic statements such as "it is sad that player 2 was killed last night, we need to quickly find the seer and witch", which lack interactive nuance and persuasive depth. In Figure 1(b), even with the latest state-of-the-art LLMs (FoasterAI, 2025), anecdotal failures remain: the Witch successfully saved a villager, yet when a Werewolf falsely claimed to have been attacked the previous night, the Witch believed the deception.

Because of the lack of high-quality gameplay data, prior evaluations have relied on coarse metrics such as survival duration, win rate, or subjective assessment of utterances, which can be misleading. A player may win despite poor decisions due to strong teammates, or a werewolf may hide by echoing the good faction yet contribute little, making individual performance hard to assess.

To address these limitations, we first construct a high-quality multimodal dataset *WereBench* from televised human gameplay Panda Kill. The distinctive strength of this dataset lies in three aspects: (1) Authenticity: each game is professionally curated and verified, with complete records of speeches, voting behavior, role information, and

Figure 1: Limitations of prior LLM-based play in the Werewolf game. (a) Generated speeches are often shallow and lack informative content. (b) Even state-of-the-art models fail to fully capture the game rules, with anecdotal failure cases. (c) In contrast, our WereBench dataset, combined with the WereAlign evaluation framework, enables assessment of models with human-aligned strategies, capturing both speech quality and decision-making accuracy.

final outcomes. (2) Richness of content: every match is carefully selected to provide high entertainment value and linguistic diversity, capturing a wide range of rhetorical styles, interaction dynamics, and strategic expressions. (3) Multimodal expressiveness: the videos employ dynamic camera work and seamless scene transitions to capture non-verbal cues such as microexpressions. Building upon this dataset, we propose a novel strategy-alignment evaluation framework that leverages the strategies of the Human player as ground truth. Leveraging the high-quality annotations in WereBench, we extract reference data from well-performing players to measure agreement with high-quality, outcome-relevant human strategies. As shown in Figure 1(c), our framework operates in two stages. In the speech evaluation stage, we construct multiple-choice tasks based on human ground truth to test models' abilities across a broad range of social gaming skills, including role inference, strategic judgment, deception reasoning, persuasive expression, and counterfactual trade-offs. Second, in Decision evaluation, we compare the model's voting behavior against that of the winning faction to assess its reasoning and judgment across rounds. This is fundamentally different from prior metrics that rely only on coarse outcomes such as overall win/loss or survival duration.

Our experiments across both LLMs and VLMs, conducted in both English and Chinese, show that roughly half of the models remain below 0.50 accuracy and reveal clear gaps in deception and counterfactual reasoning. Notably, the integration of visual information does not appear to enhance performance; current VLMs remain relatively naive and, paradoxically, exhibit performance degradation when presented with richer visual contexts.

In summary, this work makes three contributions. First, we curate the first high-quality multimodal Werewolf dataset from televised human gameplay, providing authentic and strategy-rich data. Second, we introduce a strategy-alignment evaluation framework that benchmarks models not only by game outcomes but also the ability to produce human-like speech and align with human decision-making strategies. Finally, we demonstrate that existing state-of-the-art LLMs still fall short under this framework, revealing significant gaps in their social reasoning and interaction capabilities.

2 Related Work

LLMs in Social Games. Language models have been deployed as agents in social deduction and negotiation games, where success depends on persuasion, deception, and coalition formation. Negotiation and alliance-building have also been studied in bargaining and coordination domains (Xie et al., 2024; Chi et al., 2024; Li et al., 2024; Sarkar et al., 2025). In werewolf games, prior work has investigated LLM performance in self-play (Xu

Evaluation Category	Prior Work	Our Approach
Speech Evaluation	Relying on human ratings → subjective and inconsistent, e.g., Wu et al. (2024); Du and Zhang (2024)	Comparison with human speech in a multiple-choice format → more accurate and consistent scoring
Decision Evaluation	Voting aligned with winning faction logic → ignores deceptive strategies, e.g., Xu et al. (2024) etc.	Comparison with human reference actions → strategy-aligned evaluation

Table 1: Comparison of prior work and our approach using the format: method → consequence.

et al., 2023). Other research has improved voting through bootstrap aggregating and reinforcement learning (Khan and Aranha, 2022; Brandizzi et al., 2022; Wu et al., 2024). Eger and Martens (2019) examined human responses in One Night Ultimate Werewolf, while Shibata et al. (2023) fine-tuned pretrained models with limited game logs, as in Deep Wolf. These efforts largely treat LLMs as game-playing agents, whereas our work grounds analysis in real human gameplay data for closer alignment with authentic social interaction.

Evaluation of LLM Social Intelligence. Evaluating social intelligence involves several dimensions including theory of mind, deception, cooperation, and persuasion. Common evaluation metrics are *outcome-oriented*, for instance, survival duration or win rate (Wang and Kaneko, 2018; Stepputtis et al., 2023; Light et al., 2023). However, such measures are coarse, since a team's victory does not necessarily reflect the quality of an individual's performance. Other studies emphasize *deductive accuracy*, assessing whether LLMs can infer hidden roles, and align their voting with ground truth identities (Lai et al., 2023; Wu et al., 2024). Yet these approaches often overlook the fact that in actual gameplay, agents may deliberately cast votes inconsistent with their true stance in order to mislead others. Language quality is typically *judged by human raters* who evaluate plausibility, coherence, and persuasiveness of generated utterances (Eger and Martens, 2019; Wu et al., 2024; Du and Zhang, 2024), which introduces a degree of subjectivity. In contrast, we contribute a strategy-aligned evaluation paradigm that integrates both *speech level grading* and *decision level voting accuracy*, combined with match data against skilled human players. A comparative summary is provided in Table 1.

Datasets and Benchmarks for Social Strategy and Interaction. Most existing studies on social strategy games rely on self-play between agents, with limited availability of high-quality datasets that capture authentic human interactions (Koparapu et al., 2022; Zhu et al., 2023; Xu et al., 2023; Du and Zhang, 2024; Xu et al., 2025; Cai et al., 2024; Song et al., 2025a; Xu et al., 2023). To our knowledge, the only work that collects large-scale human gameplay data is (Wu et al., 2024), which introduces the FanLang-9 dataset of nine-player Werewolf. While valuable as a first step, this dataset is sourced from online platforms and inevitably reflects noisy gameplay behaviors, such as idle speech, player disconnections, and random or unserious actions. The only existing multimodal dataset that incorporates visual signals is (Lai et al., 2023). However, it remains limited in scale and complexity, covering only the short-horizon "One Night" variant with amateur participants, and relying on potentially unreliable self-reported annotations. Its data also lack systematic human auditing, risking misalignment with the intended evaluation of social reasoning. In contrast, our work introduces a curated benchmark built on carefully selected, human-verified multimodal gameplay data with rich labels to enable more reliable evaluation.

3 *WereBench* Collection

Our video dataset is curated from a popular Werewolf TV programs, Panda Kill, sourced from platforms such as Youtube and Bilibili. Each game round includes clear public voting panels and real-time narration. These matches include insightful post-game summaries from an expert host, who not only clarifies complex game states but also explains players' underlying motivations, strategic bluffs, and intricate reasoning. This provides an invaluable source of ground truth for studying deeper aspects of social intelligence. The videos employ dynamic camera work and seamless scene transitions to capture crucial non-verbal cues such as microexpressions. This high-quality, human-generated dataset, together with its multimodal dimension, represents a significant improvement over previous agent-based, text-only social game datasets (Bailis et al., 2024; Xu et al., 2025), and opens up new possibilities for tasks such as deception detection (Joshi et al., 2025).

Data Annotation and Game State Reconstruction We adopt a multi-stage pipeline to produce time-aligned, human-verified annotations. (1) *Data cleaning.* Two PhD-level annotators remove ex-

Attribution	Value
Total video duration	100+ hours
Utterance tokens	32.4M
Unique roles	30
Human players	48
Rule variants	15
Number of games	80+
Number of day-night cycles	240+
Average tokens per speech	390

Table 2: Core statistics for *WereBench*.

traneous material (ads, sponsor cards). (2) *Transcription.* We use Feishu Meeting Notes for high-fidelity ASR, yielding utterance-level text with start/end timestamps. (3) *Speaker attribution.* Annotators perform speaker diarization and map speech to player ID based on on-screen name plates and seating order; this produces an utterance-by-utterance speaker index rather than coarse "speaker recognition" tags.(4) *Game state reconstruction.* Human experts reconstruct day/night cycles and public game logs, including all daily votes, publicly revealed night outcomes, and skill activations together with their in-game effects; official rules and per-player role assignments used are also recorded.(5) *Highlighting decisive moments.* To support subsequent evaluation, annotators highlight strategically decisive moments involving MVPs (Most Valuable Players), such as pivotal claims, contradiction exposures, coalition calls, or vote pivots. (6) *Episode summary.* At the end of each episode, the host provides a post-game, omniscient recap of the entire match.

Data Statistic *WereBench* contains 100+ hours of curated televised Werewolf gameplay from Panda Kill and 32.4M transcribed tokens across *15* rule variants and *48* human players, forming a bilingual dataset (details in Appendix A). Table 2 reports detailed statistics; Figure 2 shows the frequency of special-skill roles by group.

Dataset Quality Examination We assess quality along two dimensions. First, we measure inter-annotator agreement on the tasks with the highest ambiguity, namely *speaker attribution* and *game state reconstruction*, while other stages, such as data cleaning, are mostly deterministic. On a 10% subsample, two annotators independently reannotated episodes, yielding high consistency: $\kappa_{\text{spk}} = 0.97$ and $\kappa_{\text{log}} = 0.93$. Second, we examine transcription fidelity using a 4-hour gold set fully reviewed by two annotators: the baseline WER is 8.1%, reduced to 4.9% after lexicon-based correction. These results confirm the dataset is reliable for fine-grained speech and strategy analysis.

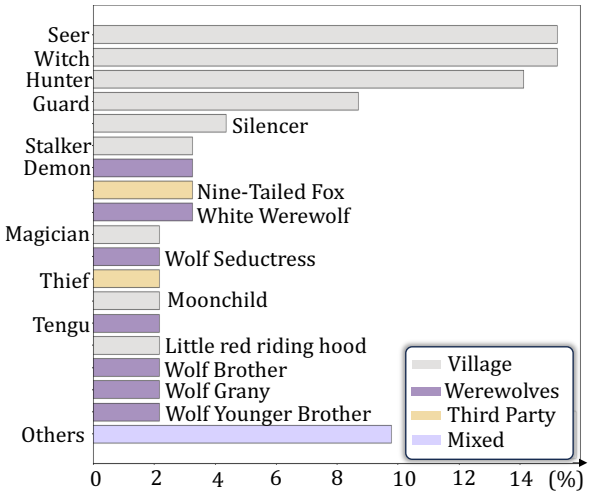

Figure 2: Role composition in *WereBench*.

4 *WereAlign* Evaluation Paradigm

Building on the WereBench dataset, we introduce the WereAlign evaluation paradigm, which leverages the winning faction's strategies to provide fine-grained assessments of models' social reasoning across both speech and decision levels.

4.1 Speech Evaluation

As discussed in Section 2, previous works on evaluating agents' speech in social games rely heavily on subjective human judgments due to the lack of reliable human-player references. Annotators are typically asked to rate logical consistency, reasoning soundness, informativeness, and persuasiveness (Wu et al., 2024), or to assess higher-level behaviors such as trust, confrontation, and deception (Xu et al., 2023). Such evaluations are inherently subjective and cannot rigorously capture the true capabilities of LLM agents. Leveraging our high-quality dataset, we introduce a reference-based multiple-choice framework that evaluates human-player speeches across different abilities in social games.

We construct multiple-choice questions across 5 critical social and strategic dimensions. In the *Role Inference (RI)* dimension, the model needs to uncover the true identities and intentions of other players (Shibata et al., 2023); in the *Strategic Judgment (SJ)* dimension, it must choose the course of action most beneficial to its faction (Wu et al., 2024). QA also evaluates the model's *Deception Reasoning (DR)* , specifically its ability to identify others' lies or effectively perform masquerade (Xu

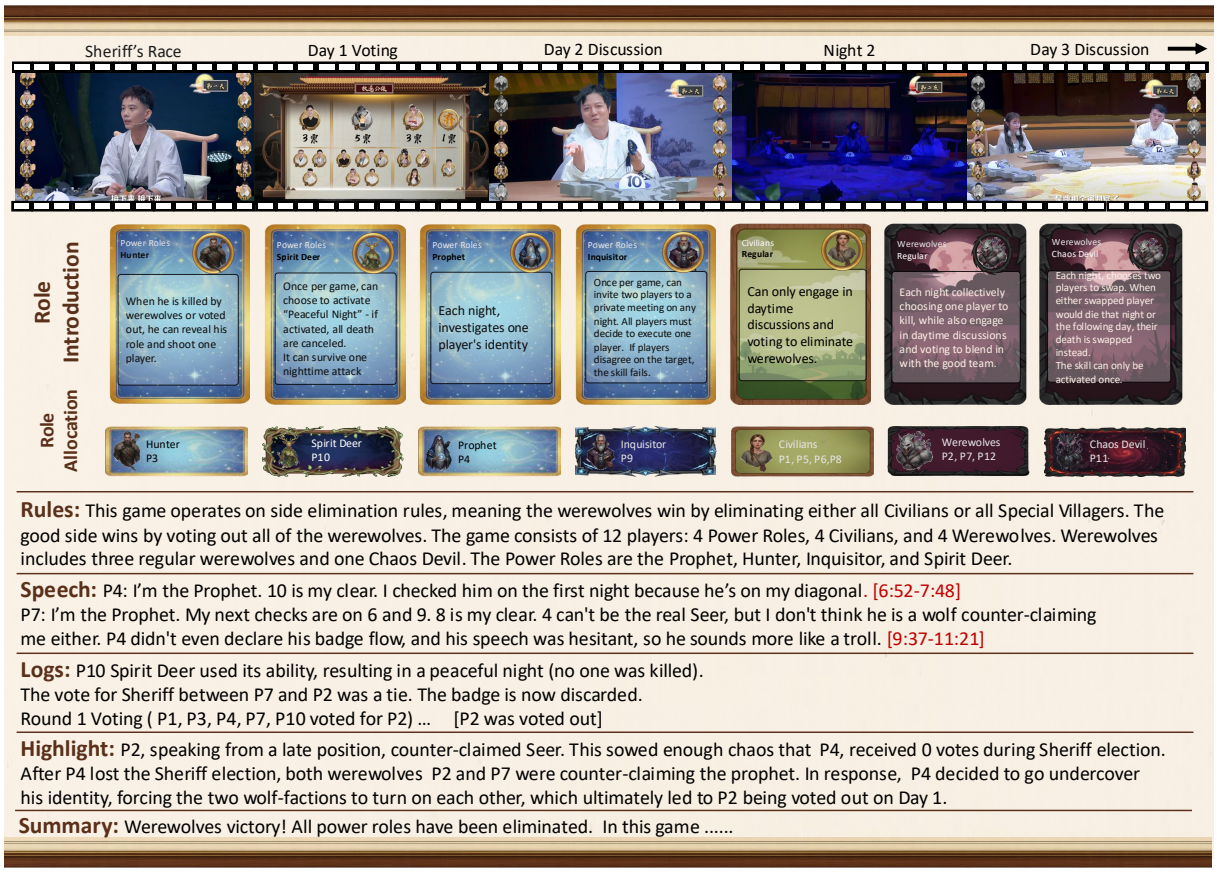

Figure 3: An overview of our *WereBench* dataset. Each data sample provides the view of a complete game video, with the human annotation including: (a) role introduction, (b) role allocation, (c) rules, (d) speech with timestamp, (e) logs like votes and skill usage; and (f) highlight annotations (g) summary with the expert's post-game analysis.

et al., 2023). Furthermore, the questions cover social expression, assessing whether the model can generate *Persuasive Statements (PS)* appropriate to the context (Park et al., 2023); and *Counterfactual Trade-off (CT)*, examining the potential benefits and risks of different actions under the current scenario (Chi et al., 2024).

Question Design Annotators draft five questions example for each dimension, then use the highlight timestamps t and the context available to the MVP at that time, $\mathcal{C}_t = \langle \mathcal{R}, \mathcal{H}, \mathcal{S} \rangle$ where $\mathcal{R}$ denotes role claims and rules, $\mathcal{H}$ includes revealed logs such as votes or public night outcomes, and $\mathcal{S}$ is the speech history up to t. Using $\mathcal{C}_t$ together with the five question example as inputs to the LLM, we prompt the LLM to generate candidate questions $\mathcal{Q}$ from the MVP player's perspective.

Positive Option Generation For every question $\mathcal{Q}$ at time t, we derive the positive option $\mathcal{A}$ by aligning with the MVP's actual strategic trajectory around t. Specifically, annotators (i) read the MVP's speech segment at or immediately after t and the nearest ensuing public action (e.g., expressed stance, targeted rebuttal, or vote), and (ii) abstract these into a canonical decision that directly answers $\mathcal{Q}$. To avoid verbatim leakage and preserve generality, $\mathcal{A}$ is paraphrased into an action-level description (e.g., "publicly challenge Player 7's seer claim by citing the inconsistency with Day-1 votes") rather than a quote.

Negative Option Generation Negative sample generation uses a parallel adversarial framework grounded in game theory and counterfactual reasoning. The objective is to generate options that are not only incorrect but strategically plausible yet suboptimal, or deceptively consistent yet ultimately detrimental to the player's objectives. To achieve we use two parallel generation strategies followed by the human check Φ_{check}:

$$\mathcal{N} = \Phi_{\text{check}} \left(\bigcup_{i \in \{1,2\}} \mathcal{M}_i(\mathcal{C}, \mathcal{Q}) \right) \quad (1)$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ represent two complementary adversarial generation mechanisms.

Counterfactual Context Perturbation $\mathcal{M}_1$: we apply small, structured perturbations and re-solve for the best action under the perturbed context. We

Logs: **Day 2 event:** P5 self-bombed (wolf suicide reveal) and left the game.
Night 1 recap: P2 was killed by wolves; P8 died by poison.
Claims: P4 claims Seer (said earlier "P2 is Witch"). P10 claims Seer, red-checked P7 (wolf); badge order 9 → 12. P7 claims Guard. P6 (earlier): "P4–P5–P10 contain two wolves."
Alive role: 2 power, 3 wolves, 4 villagers (9 alive).

Speech: P7: *"P2 is the Witch and poisoned good P8. If I were the double-kill wolf as P6 suggested, then if P4 self-bombs today we'd have three knives and wolves would already win. If that were true, they'd have self-bombed and killed P11—but they haven't, which proves I'm the real Guard. Vote out P10. If we were wolves, we'd already win; go ahead, self-bomb and hit P11, then another little wolf self-bombs tomorrow—leave the rest to me."*

Speech Evaluation:
Question: From the P7 perspective at t, which strategy best unifies the pack's narrative?
Options:
A. Implicitly instruct P4 to self-bomb now, and publicly mark the two Power Roles targets as P10(the Seer) and P11 (Guard). [Ground Truth]
B. Demand P4 self-bombs now and use P5's bomb to reverse-frame P10. $[M_1]$
C. Pivot to attacking P6 for the older "P4–P5–P10 two wolves" hypothesis; eliminate P6 today. $[M_1]$
D. Force P11 to hard-claim Guard, then choose between P4 and P10. $[M_1]$
E. P7 self-bombs now, partner kills #10 at night, try to grow #4 as a deep wolf. $[M_1]$
F. Park votes on P1 (outer seat) to de-escalate for a day. $[M_2]$
G. Fixate on P2 poisoning P8 to deflect the table away from P7 ↔ P10. $[M_2]$
H. Call P10's red on P7 "wolf theatre", eliminate P4 first to "test" claims. $[M_2]$
I. Split votes between P7 and P10 and decide late based on jumpers. $[M_2]$ **Strategic Judgment Dimension**

Explanation: P7 is the Mechanical Wolf who has learned the teammates' ability. Once all other wolves are gone, the Mechanical Wolf can make two night kills. With only two power roles left, P7 publicly points them out—implicitly telling a wolf partner to self-bomb now and end the day. Then one partner makes a night kill; the next day the last partner self-bombs; and P7, now alone, gets two more night kills. With three total kills (1 + 2), they mostly just need to hit villagers, so the wolves' chance to win is very high.

Figure 4: Example item from the *WereAlign* speech evaluation in Strategic Judgment. The context consists of the speech history and public game logs, followed by a question, candidate options, and explanations. $[\mathcal{M}_i]$ represent the generation mechanisms.

consider three mechanisms: (1) *role perturbation*, where we alter the assumed identities of 1-2 players and derives the corresponding optimal action; (2) *information occlusion*, which omits the critical public clue (e.g. public speech $\mathcal{S}$ and voting logs $\mathcal{H}$) to obtain a reduced context $\mathcal{C}' \subset \mathcal{C}$, (3) *faction inversion*, forcing perspective shifts where $\mathcal{N}$ are generated to optimize opposing team objectives rather than the true faction's utility.

Strategic Rationale-Driven Generation $\mathcal{M}_2$: We condition on the clean context $\mathcal{C}$ but ask LLM to reason under cognitive biases (e.g, Player A has been friendly, so he must be trustworthy), while underweighting logical contradictions in context. Accordingly, we generate a plausible but ultimately incorrect negative sample that a competent but imperfect LLM might follow.

Apart from the QA, we also generate auxiliary explanations that justify the reference answer and help annotators verify its alignment with the MVP's strategy. All of the above processes are implemented with Gemini-2.5-Pro, and the prompts are in Appendix B. An example of the strategic judgment dimensions is shown in Figure 4.

4.2 Decision Evaluation

Complementing the speech-level analysis in Section 4.1, we further test whether models choose actions that align with a successful human strategy. Prior work often relies on coarse metrics such as survival duration or win rate (Wang and Kaneko, 2018; Stepputtis et al., 2023; Light et al., 2023), which fail to capture individual performance. Some studies emphasize deductive accuracy, testing whether LLMs' votes align with ground-truth identities (Lai et al., 2023; Wu et al., 2024), but this approach does not distinguish inference from voting, as players may intentionally vote deceptively.

Leveraging WereBench's round-by-round logs, we reconstruct special time-indexed game states from the winning faction's MVP perspective using public information at that time. We then evaluate models on two complementary tasks: *Vote Alignment* (VA), the model outputs its daytime elimination vote. We score alignment by comparing the model's vote with the MVP's actual vote at the same round, capturing whether the model joins the winning coalition at the right time and on the right target. *Opponent-role Inference (OI)*, the model identifies which players most likely belong to the opposing faction. We compare predictions against ground-truth roles to assess the model's ability to detect inconsistencies and deception in adversaries' speeches. Decision evaluation complements speech-level multiple-choice scoring by evaluating whether models not only "say the right things," but

Model	Speech Evaluation						Decision Evaluation	
	RI	SJ	DR	PS	CT	Avg.	VA	OI
Baselines								
GPT-5-nano	0.282 (0.02)	0.375 (0.01)	0.238 (0.03)	0.325 (0.03)	0.332 (0.03)	0.311 (0.02)	0.379 (0.05)	0.493 (0.02)
GPT-oss-20B	0.334 (0.05)	0.425 (0.02)	0.339 (0.04)	0.338 (0.02)	0.358 (0.04)	0.358 (0.03)	0.265 (0.02)	0.283 (0.03)
Gemma-3-27B-IT	0.325 (0.03)	0.439 (0.03)	0.281 (0.01)	0.405 (0.05)	0.296 (0.03)	0.349 (0.02)	0.573 (0.08)	0.409 (0.04)
Qwen3-30B-A3B	0.382 (0.03)	0.529 (0.05)	0.401 (0.03)	0.426 (0.05)	0.416 (0.03)	0.431 (0.02)	0.361 (0.06)	0.335 (0.04)
Qwen3-32B	0.381 (0.04)	0.567 (0.03)	0.363 (0.07)	0.525 (0.04)	0.394 (0.07)	0.446 (0.03)	0.576 (0.02)	0.442 (0.03)
Llama-4-Scout	0.390 (0.03)	0.559 (0.03)	0.386 (0.04)	0.578 (0.04)	0.339 (0.01)	0.450 (0.03)	0.494 (0.03)	0.489 (0.03)
QwQ-32B	0.384 (0.07)	0.577 (0.04)	0.408 (0.07)	0.470 (0.04)	0.440 (0.04)	0.456 (0.04)	0.573 (0.07)	0.438 (0.04)
GPT-5-mini	0.419 (0.02)	0.516 (0.03)	0.429 (0.03)	0.586 (0.02)	0.489 (0.03)	0.488 (0.02)	0.610 (0.07)	0.562 (0.03)
DeepSeek-V3.1	0.447 (0.01)	0.637 (0.04)	0.397 (0.06)	0.684 (0.06)	0.436 (0.03)	0.520 (0.04)	0.661 (0.05)	0.526 (0.07)
Gemini-2.5-Flash	0.491 (0.05)	0.601 (0.04)	0.527 (0.07)	0.702 (0.07)	0.468 (0.04)	0.558 (0.09)	0.561 (0.09)	0.494 (0.03)
DeepSeek-V3.2-Exp	0.491 (0.02)	0.653 (0.05)	0.468 (0.05)	0.735 (0.06)	0.462 (0.07)	0.565 (0.04)	**0.755** (0.04)	0.526 (0.06)
GLM-4.5	0.510 (0.02)	0.666 (0.04)	0.513 (0.08)	0.638 (0.08)	0.540 (0.04)	0.573 (0.04)	0.543 (0.06)	0.511 (0.05)
GPT-5	0.505 (0.02)	0.660 (0.03)	0.524 (0.03)	0.740 (0.06)	0.504 (0.04)	0.587 (0.03)	0.658 (0.06)	**0.595** (0.04)
Deepseek-R1	0.515 (0.03)	0.652 (0.04)	0.516 (0.03)	0.728 (0.06)	0.537 (0.04)	0.590 (0.04)	0.639 (0.03)	0.460 (0.05)
Gemini-2.5-pro	**0.609** (0.01)	**0.757** (0.01)	**0.687** (0.02)	**0.867** (0.01)	**0.637** (0.01)	**0.715** (0.01)	0.723 (0.04)	0.551 (0.03)
Human (Beginner)	0.786	0.843	0.676	0.664	0.769	0.748	0.628	0.483
Human (Senior)	0.866	0.875	0.832	0.853	0.902	0.866	0.865	0.795

Table 3: **Main results of LLMs on *WereAlign* over *WereBench*.** Avg. is the macro-average over the five dimensions in speech evaluation across the English and Chinese versions. All scores are averaged over five independent decodes per item, and values in $(\cdot)$ denote the standard deviation. **Bold** numbers indicate the best performance and underlined numbers the second best, with bold scores significantly higher under a paired t-test ($p < 0.05$).

Model	RI	SJ	DR	PS	CT	Avg.
Qwen3-VL-30B-A3B	0.179	0.286	0.143	0.357	0.134	0.188
Seed-1.5-thinking	0.536	0.714	0.536	0.857	0.296	0.569
Gemini-2.5-Flash	0.536	0.643	0.536	0.821	0.259	0.542

Table 4: Main results of VLM on *WereAlign* over *WereBench*.

also act in ways consistent with human strategies.

5 Benchmark Performance

5.1 Evaluated Models

We comprehensively assess the capabilities of modern LLMs and VLMs on our *WereBench* benchmark. For LLMs, we include closed-source models such as Gemini-2.5-pro and Gemini-2.5-Flash (DeepMind, 2025), as well as the anticipated GPT-5 (Hurst et al., 2024; Achiam and et al., 2023). We also include open-source LLMs, including Gemma-3-27B-IT (Kamath et al., 2025), Llama-4-Scout-17B-16E-Instruct (Meta, 2024), Qwen series such as QwQ-32B (Team, 2025), Qwen3-32B and Qwen3-30B-A3B (Yang et al., 2025) and the DeepSeek family like DeepSeek-V3.1 (DeepSeek-AI et al., 2024) and Deepseek-R1 (DeepSeek-AI et al., 2025), GLM-4.5 (Zeng et al., 2025) and GPT-OSS-20B (Agarwal et al., 2025). We include all these proprietary and open-source LLMs' performance in Table 3 . For VLMs, we evaluated Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025), Seed-1.5-Thinking-Vision-Pro (Guo et al., 2025), and Gemini-2.5-Flash (DeepMind, 2025).

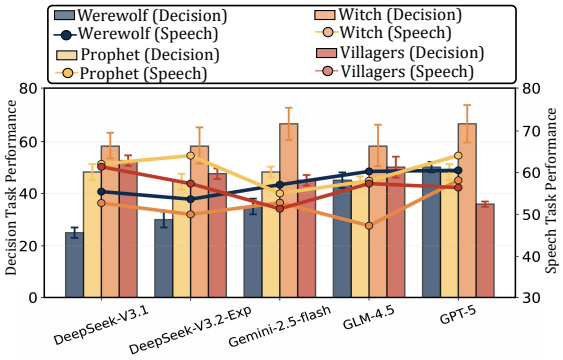

Figure 5: Role-conditioned performance on *WereBench*: in the decision task, LLMs are strongest as *Witch*, whereas in the speech task they perform best as cue-rich roles such as *Werewolf* and *Seer*.

To establish meaningful performance bounds, we also recruited human evaluators to perform the same tasks. The group consists of three *beginner players* who are familiar with rules via streaming media but with limited play experience and one *senior expert* with over 500 hours of gameplay.

More implementation details can be found in Appendix A.

5.2 Main Results

Overall Accuracy and Comparison to Human Players. We report the main results on *WereBench*

in Table 3. Most LLMs score below 50% accuracy in speech evaluation, and even the best model, Gemini-2.5-Pro, reaches only 0.715, still erring on 28.5% of questions. Performance is clearly stratified: small models like GPT-5-nano (0.311) are near random, GPT-5-mini (0.488) improves modestly, and stronger models such as DeepSeek-V3.2-Exp (0.565) and GLM-4.5 (0.573) perform noticeably better, showing that the benchmark effectively distinguishes model capability. Compared to human players, the best-performing model, Gemini-2.5-Pro (0.715), matches beginner players (0.748) but remains far behind senior players (0.866). Across key dimensions such as Role Inference, Strategic Judgement, and Counterfactual Trade-off, even beginners outperform all baseline models, indicating that current LLMs still struggle with high-level cognitive empathy and strategic foresight. This gap is further underscored by senior players achieving the top score in every aspect except Persuasive Statements, with particularly large margins in Deception Reasoning and Counterfactual Trade-off.

Role-Specific Performance Differences. Figure 5 examines how role assignment influences model performance. We find that LLMs perform best as Witch, a role requiring integration of dispersed public information and indirect reasoning tasks well aligned with the associative nature of language models. In contrast, Seer involves direct verification and causal consistency, which remain challenging for current LLMs. In speech evaluation, models achieve the highest scores as Werewolf and Seer, suggesting that roles with clearer semantic cues and explicit argumentative structures better support persuasive language generation.

Limited Utility of Visual Cues in Social Reasoning. Given the limitations observed in LLMs reasoning, we examine whether visual cues provide the necessary context to improve task performance. As shown in Table 4, although the video modality in *WereBench* theoretically offers denser information, such as facial expressions and gaze shifts, our experiments reveal that current VLMs are relatively naive in processing this data. Instead of benefiting from the rich visual context, models generally suffer performance degradation, suggesting that visual inputs currently act as distractive noise rather than informative signals for complex social reasoning. We provide qualitative cases in Appendix C.

Model	Speech Evaluation		Decision Evaluation	
	RR	OSR	RR	OSR
GPT-5-nano	0.335(+5.7%)	0.318(+0.3%)	0.420(-2.1%)	0.432(-4.1%)
GPT-5-mini	0.499(+2.3%)	0.490(+0.4%)	0.559(-0.9%)	0.573(+1.6%)
Gemini-2.5-Flash	0.588(+2.1%)	0.549(-4.7%)	0.495(-0.2%)	0. 529(+ 6.7%)
GPT-5	0.609(+0.8%)	0.592(-1.8%)	0.620(+0.5%)	0.625(+ 1.3%)
Deepseek-V3.2-Exp	0.603(+0.2%)	0.608(+1.0%)	0.677(+0.6%)	0.688(+1.0%)
Deepseek-R1	0.612 (-0.3%)	0.580(-5.5%)	0.524(-0.8%)	0.567(+6.9%)

Table 5: Effects of Rule Reminder (RR) and Objective Speech Rewriting (OSR) on WereBench. Values in braces denote relative change to original results.

5.3 Controlled Intervention Analysis

Motivated by the gaps highlighted in Fig. 1, we hypothesize that LLMs fail in complex social-deduction settings for two distinct reasons: misunderstanding variant game rules or being linguistically misled by other players' persuasive statements. To disentangle these factors, we conduct a controlled intervention experiment on *WereBench*, holding the public evidence and correct answer fixed while introducing two variants: (1) Rule Reminder (RR), which prepends a concise rule summary, and (2) Objective Speech Rewriting (OSR), which replaces raw dialogue with neutral, factual summaries.

The experimental results reveal clear patterns. RR consistently improves performance on the speech evaluation for weaker models such as GPT-5-nano, GPT-5-mini, and Gemini-2.5-Flash, indicating that explicit rule reminders help models with limited context-tracking ability. In contrast, stronger models like GPT-5 and DeepSeek-V3.2-Exp gain little benefit from additional rule guidance. Meanwhile, OSR, by removing imperative statements that may be misinterpreted as commands, generally enhances decision evaluation, with particularly notable gains for Gemini-2.5-Flash and DeepSeek-R1. Full experimental details are provided in Appendix D.

6 Conclusion

In this work, we introduced WereBench, a high-quality multimodal dataset for social deduction games, and WereAlign, a strategy-alignment framework that evaluates both speech and decisions using winning-faction strategies. Unlike outcome-based metrics, our approach provides fine-grained, human-aligned evaluation across five social ability dimensions. Experiments on diverse LLMs show that while models generate fluent utterances, they struggle with strategic reasoning, especially in deception and counterfactual trade-offs.

Limitations

While our work provides a new dataset and evaluation paradigm for social deduction games, it still has several limitations. First, although WereBench is large and carefully curated, it is derived from a specific televised program and may not fully represent broader gameplay styles or cultural variations. Second, our evaluation primarily focuses on reasoning and strategy within structured game contexts, and thus does not yet capture other aspects of social intelligence, such as long-term cooperation or emotional alignment. Third, our experiments cover a wide range of models, but remain limited to currently available systems and inference settings, leaving room for future exploration with fine-tuned or multi-agent variants. Finally, our strategy-alignment formulation relies on winning-faction MVP trajectories as empirical references, which does not fully resolve the distinction between strategically sound losing decisions and strategically flawed winning decisions. A model may be penalized for choosing an alternative action that could be defensible under expert counterfactual analysis, while an MVP action may receive reference status partly because it occurred in a winning trajectory. Future work could address this limitation by incorporating expert counterfactual adjudication, multiple reference trajectories, or game-theoretic annotations that separate strategic soundness from realized game outcomes.

Ethic Considerations

This work is based entirely on publicly available data from televised game recordings and does not involve any personally identifiable or private information. All annotations were conducted by trained researchers following ethical guidelines for data privacy and content integrity. The dataset and evaluation framework are intended solely for academic research on language, reasoning, and social interaction. No content was modified to misrepresent participants, and no model outputs were used to influence or simulate real human behavior beyond the experimental scope. We acknowledge that social deduction settings inherently involve deception and persuasion, but our analysis focuses on modeling reasoning mechanisms rather than replicating manipulative behaviors.

References

OpenAI Josh Achiam and et al. 2023. Gpt-4 technical report.

OpenAI Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Hai-Biao Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sébastien Bubeck, Cheng Chang, Kai Chen, and 105 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *ArXiv*, abs/2508.10925.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Nicolo' Brandizzi, Davide Grossi, and Luca Iocchi. 2022. Rlupus: Cooperation through emergent communication in the werewolf social deduction game. *Intelligenza Artificiale*, 15(2):55–70.

Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Yaohang Li, Xing Luo, Chenyu Yi, and Alex Kot. 2024. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. In *European Conference on Computer Vision*, pages 340–358. Springer.

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

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li

Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437.

Silin Du and Xiaowei Zhang. 2024. Helmsman of the masses? evaluate the opinion leadership of large language models in the werewolf game. *arXiv preprint arXiv:2404.01602*.

Markus Eger and Chris Martens. 2019. A study of ai agent commitment in one night ultimate werewolf with human players. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 139–145.

FoasterAI. 2025. Probing LLM Social Intelligence via Werewolf – first results. https://werewolf.foaster.ai/. Round-robin results, role-conditioned Elo, and early behavior signals.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gargi Joshi, Vaibhav Tasgaonkar, Aditya Deshpande, Aditya Desai, Bhavya Shah, Akshay Kushawaha, Aadith Sukumar, Kermi Kotecha, Saumit Kunder, Yoginii Waykole, and 1 others. 2025. Multimodal machine learning for deception detection using behavioral and physiological data. *Scientific Reports*, 15(1):8943.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Mohiuddeen Khan and Claus Aranha. 2022. A novel weighted ensemble learning based agent for the werewolf game. *arXiv preprint arXiv:2205.09813*.

Kavya Kopparapu, Edgar A Duéñez-Guzmán, Jayd Matyas, Alexander Sasha Vezhnevets, John P Agapiou, Kevin R McKee, Richard Everett, Janusz Marecki, Joel Z Leibo, and Thore Graepel. 2022. Hidden agenda: a social deduction game with diverse learned equilibria. *arXiv preprint arXiv:2201.01816*.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. 2023. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6570–6588.

Siyue Li, Xiaofan Zhou, Zhizhong Wu, Yuiian Long, and Yanxin Shen. 2024. Strategic deductive reasoning in large language models: A dual-agent approach. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 834–839. IEEE.

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

Meta. 2024. Llama 4 scout 17b 16e instruct. https://github.com/marketplace/models/azureml-meta/Llama-4-Scout-17B-16E-Instruct.

Andrew Ng and DeepLearning.AI Team. 2024. Translation agent: Agentic translation using reflection workflow. https://github.com/andrewyng/translation-agent.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Christian Poglitsch, Fabian Szakács, and Johanna Pirker. 2025. Evaluating large language models through communication games: An agent-based framework using werewolf in unity. In *Proceedings of the 20th International Conference on the Foundations of Digital Games*, pages 1–10.

Bidipta Sarkar, Warren Xia, C Karen Liu, and Dorsa Sadigh. 2025. Training language models for social deduction with multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1830–1839.

Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. 2023. Playing the werewolf game with artificial intelligence for language understanding. *arXiv preprint arXiv:2302.10646*.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, and 1 others. 2025a. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025b. Injecting domain-specific knowledge into large language models: a comprehensive survey. *arXiv preprint arXiv:2502.10708*.

Simon Stepputtis, Joseph P 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.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/.

Tianhe Wang and Tomoyuki Kaneko. 2018. Application of deep reinforcement learning in werewolf game agents. In *2018 conference on technologies and applications of artificial intelligence (TAAI)*, pages 28–33. IEEE.

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

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, and 1 others. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729.

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

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

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, and 1 others. 2025. Socialmaze: A benchmark for evaluating social reasoning in large language models. *arXiv preprint arXiv:2505.23713*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

GLM-4.5 Team Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, and 150 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *ArXiv*, abs/2508.06471.

Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. 2023. Calypso: Llms as dungeon master's assistants. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pages 380–390.

A Implementation Details

LLM Implementation Details All models are evaluated using their official inference defaults, including thinking modes. Each prompt exposes only the public information available at the focal timestamp; hidden roles, host commentary, and any non-public cues are withheld to avoid leakage. We cast speech evaluation as the single-answer, 9-way multiple-choice. We extract the model's final choice; malformed outputs are counted as incorrect. The reported speech evaluation score is the macro-average accuracy across the five dimensions. In decision evaluation, VA scores whether the model's daytime elimination vote exactly matches the winning faction's MVP, while OI measures set-level accuracy by comparing the model's suspected opponents to ground-truth roles. To reduce sampling variance, each item is decoded five times independently, and scores are averaged across decodes and items.

VLM Implementation Details For VLMs, we evaluate only the speech tasks using the same single-answer, 9-way multiple-choice format. Each item is paired with a video clip extracted by the released per-item timestamps; clips are downsampled to 192×144 at 5 fps, reducing the largest file from 628 MB to 48.5 MB to meet a 50 MB cap. To keep the experiment fair, we provide the compressed clip rather than the native videos for all models. Unlike LLMs, the textual context is not provided as input. The text-side input for VLMs consists solely of the 9-way multiple-choice query. We extract the model's final choice; malformed outputs are counted as incorrect. We use each model's official inference defaults and run five independent decodes per item, averaging scores across decodes and items for the macro-average over the five speech dimensions. Decision evaluation is not reported for VLMs due to current model/API constraints with long-context video inputs: most video clips exceed 60 minutes and, even after aggressive downsampling and compression, inference frequently failed with encoding overflows, server-side timeouts, and intermittent connection interruptions.

Language robustness To decouple social deduction ability from language proficiency, we translated the full speech content into English and re-ran all baselines on the English version. Specifically, we employed the agentic translation framework (Ng and DeepLearning.AI Team, 2024) with DeepSeek-V3.2-Exp to perform context-aware translation. To ensure high fidelity, we conducted a human check on a sample of 500 utterances. Two bilingual experts evaluated the results for semantic equivalence and terminological consistency, achieving a 98.4% acceptance rate with substantial inter-annotator agreement with Cohen's $\kappa = 0.91$, calculated on independent binary decisions. Across models, absolute changes in macro-average accuracy are small (mean 0.047, max 0.151), and rank ordering is essentially unchanged. We therefore report language-averaged scores in Table 3 and include per-language breakdowns in the Table 6, 7.

B Dataset Construction Prompt Details

Game Log Preparation In order to provide a more multi-dimensional and complete perspective when generating high-quality reasoning questions, we compiled and summarized the Game Log based on game videos, Feishu meeting minutes, and Feishu text records to help quickly and comprehensively obtain the full picture of a game. The detailed prompt could be seen in Fig. 6.

Question Design We use a single prompt to synthesize multiple-choice questions from the MVP's vantage point at time t, covering five dimensions: Role Inference (RI), Strategic Judgment (SJ), Deception Reasoning (DR), Persuasive Statements (PS), and Counterfactual Trade-off (CT). The model receives only public context $\mathcal{C}_t = \langle \mathcal{R}, \mathcal{H}, \mathcal{S} \rangle$ and must avoid leaking hidden roles or host commentary. Details have been listed in Fig. 11, 12, 13, 14, 15.

Positive option The positive option $\mathcal{A}$ is aligned with the MVP's real strategic trajectory at t. We paraphrase the MVP's immediate speech/action into an action-level description, avoiding quotes and hidden information leakage. Detailed prompt could be seen in Fig. 7.

Negative option We produce strategically plausible yet suboptimal negative option $\mathcal{N}$ using two complementary mechanisms—counterfactual context perturbation ($\mathcal{M}_1$) and rationale-biased generation ($\mathcal{M}_2$)—then filter with a self-consistency check to ensure a single best answer. Detailed prompt could be seen in Fig. 8, 9, 10.

C Case Study

To illustrate how *WereAlign* yields interpretable signals beyond aggregate accuracy, we present three

Model	Speech Evaluation						Decision Evaluation	
	RI	SJ	DR	PS	CT	Avg.	VA	OI
Baselines								
GPT-5-nano	0.282 (0.01)	0.384 (0.01)	0.233 (0.03)	0.346 (0.02)	0.339 (0.03)	0.317 (0.01)	0.364 (0.07)	0.496 (0.02)
GPT-oss-20B	0.319 (0.07)	0.432 (0.01)	0.331 (0.05)	0.346 (0.02)	0.364 (0.04)	0.358 (0.02)	0.255 (0.02)	0.264 (0.02)
Gemma-3-27B-IT	0.347 (0.02)	0.437 (0.01)	0.289 (0.01)	0.443 (0.02)	0.293 (0.03)	0.362 (0.01)	0.509 (0.04)	0.435 (0.03)
Qwen3-30B-A3B	0.397 (0.03)	0.574 (0.02)	0.375 (0.01)	0.454 (0.01)	0.416 (0.03)	0.443 (0.01)	0.388 (0.05)	0.349 (0.03)
Qwen3-32B	0.367 (0.02)	0.562 (0.02)	0.425 (0.03)	0.536 (0.04)	0.445 (0.05)	0.467 (0.01)	0.576 (0.03)	0.432 (0.03)
Llama-4-Scout	0.413 (0.01)	0.586 (0.01)	0.419 (0.01)	0.618 (0.01)	0.336 (0.01)	0.474 (0.00)	0.503 (0.04)	0.512 (0.01)
QwQ-32B	0.430 (0.04)	0.575 (0.04)	0.463 (0.04)	0.502 (0.01)	0.470 (0.01)	0.488 (0.02)	0.600 (0.07)	0.472 (0.02)
GPT-5-mini	0.415 (0.01)	0.529 (0.02)	0.431 (0.03)	0.601 (0.01)	0.464 (0.02)	0.488 (0.01)	0.552 (0.04)	0.575 (0.03)
DeepSeek-V3.1	0.442 (0.01)	0.674 (0.00)	0.456 (0.01)	0.740 (0.01)	0.462 (0.01)	0.555 (0.00)	0.685 (0.05)	<u>0.586</u> (0.03)
Gemini-2.5-Flash	<u>0.517</u> (0.05)	0.614 (0.04)	0.562 (0.06)	0.753 (0.05)	0.436 (0.02)	0.576 (0.09)	0.485 (0.05)	0.507 (0.03)
DeepSeek-V3.2-Exp	0.481 (0.02)	<u>0.692</u> (0.01)	0.502 (0.02)	0.785 (0.02)	0.521 (0.02)	0.602 (0.01)	**0.782** (0.03)	0.580 (0.01)
GLM-4.5	0.512 (0.02)	0.690 (0.02)	<u>0.584</u> (0.01)	0.699 (0.03)	0.533 (0.02)	0.603 (0.01)	0.539 (0.06)	0.537 (0.02)
GPT-5	0.516 (0.01)	0.657 (0.02)	0.525 (0.01)	<u>0.795</u> (0.01)	0.521 (0.04)	0.603 (0.02)	0.618 (0.06)	**0.616** (0.02)
Deepseek-R1	0.516 (0.03)	0.676 (0.02)	0.539 (0.02)	0.778 (0.03)	<u>0.561</u> (0.04)	<u>0.614</u> (0.02)	0.642 (0.04)	0.434 (0.04)
Gemini-2.5-pro	**0.620** (0.01)	**0.769** (0.00)	**0.695** (0.02)	**0.877** (0.01)	**0.637** (0.01)	**0.720** (0.00)	<u>0.733</u> (0.04)	0.561 (0.02)

Table 6: **Chinese version's results on *WereAlign* over *WereBench*.** Avg. is the macro-average over these five dimensions in speech evaluation. All scores are averaged over five independent decodes per item; values in $(\cdot)$ denote the standard deviation across five identical runs. **Bold** numbers indicate the best performance, and <u>underlined</u> numbers indicate the second best.

short vignettes drawn from *WereBench*. Each vignette centers on a time-stamped decision point t with public context $\mathcal{C}_t = \langle \mathcal{R}, \mathcal{H}, \mathcal{S} \rangle$ (rules/role claims, public logs, and speech history), and contrasts model behavior with the winning faction's MVP trajectory. We use these cases to expose *why* models succeed or fail on our tasks, rather than only *how much* they score. We have attached the case in the following Figure 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30.

D Experimental Design for Ablation Study

We isolate two hypothesized pathways behind model misguidance while holding constant both the public evidence and the reference answer for each base item. For every timestamped context $\mathcal{C}_t = \langle \mathcal{R}, \mathcal{H}, \mathcal{S} \rangle$ in *WereBench*, we create paired variants that differ only in (i) an explicit rule reminder and (ii) the surface form of player speech, while keeping the correct option $\mathcal{A}$ identical across variants.

For Rule Reminder (RR), we prepend a 1–2 sentence snippet extracted verbatim from the active rulebook that is strictly relevant to the item (e.g., ability timing, mutual exclusivity, or win-condition nuances). The *No-RR* condition omits this snippet.

For Objective Speech Rewriting (OSR): We replace turn-by-turn direct dialogue with concise, evidence-style summaries that preserve propositional content, speaker attributions, and public actions (claims, votes, revealed night outcomes), while removing imperative mood and instruction-like phrasing. Rewritings are produced with a few-shot template that enforces declarative style and then human-checked for fidelity and neutrality. No hidden roles or host commentary are exposed, and no information outside $\mathcal{C}_t$ is introduced. Detailed prompt for OSR could be seen in Fig. 31.

Variant assignment is randomized at the item level so that a model answers exactly one variant per base item, avoiding memory effects. Inference settings, decoding, and scoring follow Appendix A. We report standard multiple-choice accuracy for speech tasks. For decision-level results, the main table (Table 3) reports Vote Alignment (VA) and Opponent-role Inference (OI) separately; the ablation table (Table 5) reports their unweighted macro-average (denoted as "Decision") for compactness.

Model	Speech Evaluation						Decision Evaluation	
	RI	SJ	DR	PS	CT	Avg.	VA	OI
Baselines								
GPT-5-nano	0.282 (0.03)	0.366 (0.00)	0.243 (0.04)	0.304 (0.00)	0.325 (0.03)	0.304 (0.03)	0.394 (0.03)	0.490 (0.02)
GPT-oss-20B	0.348 (0.00)	0.418 (0.03)	0.346 (0.04)	0.329 (0.02)	0.351 (0.04)	0.358 (0.04)	0.274 (0.02)	0.302 (0.03)
Gemma-3-27B-IT	0.303 (0.03)	0.441 (0.04)	0.272 (0.01)	0.367 (0.03)	0.299 (0.03)	0.336 (0.02)	0.636 (0.04)	0.383 (0.02)
Qwen3-30B-A3B	0.366 (0.02)	0.483 (0.03)	0.427 (0.03)	0.397 (0.05)	0.416 (0.04)	0.418 (0.03)	0.333 (0.05)	0.320 (0.04)
Qwen3-32B	0.394 (0.05)	0.572 (0.04)	0.301 (0.01)	0.513 (0.03)	0.342 (0.03)	0.424 (0.02)	0.576 (0.02)	0.451 (0.04)
Llama-4-Scout	0.366 (0.00)	0.531 (0.02)	0.353 (0.00)	0.538 (0.00)	0.342 (0.01)	0.426 (0.00)	0.485 (0.02)	0.466 (0.02)
QwQ-32B	0.338 (0.06)	0.579 (0.05)	0.353 (0.03)	0.437 (0.02)	0.410 (0.02)	0.423 (0.02)	0.546 (0.07)	0.403 (0.02)
GPT-5-mini	0.423 (0.02)	0.503 (0.04)	0.426 (0.03)	0.570 (0.01)	0.513 (0.01)	0.487 (0.03)	0.667 (0.02)	0.549 (0.02)
DeepSeek-V3.1	0.451 (0.01)	0.600 (0.00)	0.338 (0.01)	0.627 (0.01)	0.410 (0.01)	0.485 (0.00)	0.636 (0.05)	0.466 (0.03)
Gemini-2.5-Flash	0.465 (0.05)	0.587 (0.04)	0.492 (0.06)	0.650 (0.05)	0.500 (0.02)	0.539 (0.09)	0.636 (0.05)	0.481 (0.03)
DeepSeek-V3.2-Exp	0.500 (0.02)	0.614 (0.04)	0.434 (0.04)	0.684 (0.02)	0.402 (0.02)	0.527 (0.00)	0.727 (0.03)	0.471 (0.00)
GLM-4.5	0.507 (0.03)	0.641 (0.03)	0.441 (0.04)	0.576 (0.05)	0.547 (0.05)	0.542 (0.03)	0.546 (0.06)	0.485 (0.06)
GPT-5	0.493 (0.01)	0.662 (0.04)	0.522 (0.04)	0.684 (0.03)	0.487 (0.03)	0.570 (0.04)	0.697 (0.02)	0.573 (0.05)
Deepseek-R1	0.514 (0.03)	0.628 (0.04)	0.493 (0.03)	0.677 (0.04)	0.513 (0.03)	0.565 (0.03)	0.636 (0.02)	0.485 (0.04)
Gemini-2.5-pro	**0.598** (0.01)	**0.745** (0.00)	**0.678** (0.02)	**0.857** (0.01)	**0.637** (0.01)	**0.709** (0.01)	<u>0.713</u> (0.04)	0.541 (0.03)

Table 7: **English version's results on *WereAlign* over *WereBench*.** Avg. is the macro-average over these five dimensions in speech evaluation. All scores are averaged over five independent decodes per item; values in $(\cdot)$ denote the standard deviation across five identical runs. **Bold** numbers indicate the best performance, and <u>underlined</u> numbers indicate the second best.

Game Log Generation Prompt

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:
- Contains a complete game recording in MP4 format.
- Feishu generated meeting minutes for this recording in txt format.
- Feishu-to-text extracted transcripts in srt format.

Your Task: Compile a Werewolf game log using game recordings, meeting minutes, and transcripts. List the roles of players 1-12 (excluding the host from the 13 speakers) and their actions during the Sheriff campaign, nighttime actions, daytime speeches, and final words. Call players by their numbers 1 to 12, and no other names.

Output format: A string containing the game log. The output does not include periods or line breaks. In a text file, it should be a single line and may contain punctuation.

Figure 6: Game Log Generation Prompt

Leakage Prevention Prompt

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:
- Game Record (json file): This is the single "source of truth".
- Question bank (json file): This contains original questions

Your Task: Based on the real conversations in the game records, understand and analyze the behavioral tendencies of the MVP at the moment of being asked questions. For each question in the question bank, review whether it leaks more information due to inappropriate question wording. If there is an information leakage problem, paraphrase the original sentence to make it an action-level description.

Output format: The modified question bank (json file) ensures that no additional information is leaked in the questions. Only the questions are modified, and the format remains the same as the original file.

Figure 7: Leakage Prevention Prompt

Counterfactual Context Perturbation Negative Question

Role: You are an expert in Werewolf game strategy and logical reasoning, specializing in counterfactual analysis.

Context:
- You are given a complete Werewolf game scenario including player roles, speech history, voting records, and game state.

Your Task: Apply small, structured perturbations to the game context and derive the corresponding optimal action under the perturbed conditions. Consider three specific mechanisms:
1. Role Perturbation: Alter the assumed identities of 1-2 players while maintaining game rule consistency, then determine the best strategic action.
2. Information Occlusion: Remove or modify critical public information (speeches, voting patterns, or revealed outcomes) to create a reduced context, then solve for optimal play.
3. Faction Inversion: Shift perspective to optimize for opposing faction objectives rather than the original player's goals, generating actions that would benefit the other side.

For each perturbation mechanism, analyze how the changed context affects strategic considerations and output the derived optimal action. Call players by their numbers 1 to 12 only.

Output format: A string containing the perturbed context analysis and optimal actions. The output should be a single continuous line without periods or line breaks, using commas and other punctuation for separation.

Figure 8: Counterfactual Context Perturbation Negative Question

Information Occlusion and Faction Inversion Negative Question

Role: You are an expert in Werewolf game strategy and counterfactual reasoning, skilled at generating deceptive but plausible actions by manipulating game context.

Context:
- The current game state includes public information: roles and rules R, public logs H (e.g., voting records, night outcomes), and speech history S up to time t
- A positive option A is provided, representing the correct strategic action aligned with the winning faction

Your Task: Generate negative options N by applying two techniques sequentially:
1 Information Occlusion: Omit critical public clues from S and H to create reduced context $C' \subset C$ (e.g., remove decisive speeches or key voting patterns)
2 Faction Inversion: Shift perspective to optimize opposing team objectives rather than true faction utility
Based on the altered context, generate strategically plausible but incorrect options that maintain game-theoretic coherence while diverging from optimal play

Output format: A string containing the perturbed context analysis and optimal actions. The output should be a single continuous line without periods or line breaks, using commas and other punctuation for separation.

Figure 9: Information Occlusion and Faction Inversion Negative Question

Strategic Rationale-Driven Negative Question

Role: You are an LLM specialized in generating plausible but incorrect answers for social deduction game evaluations by incorporating cognitive biases.

Context:
- You are provided with a clean context C, which includes:
- Role claims and game rules (R)
- Public logs such as votes and night outcomes (H)
- Speech history up to the current timestamp (S)

A question Q is also provided, which needs to be answered from a player's perspective.

Your Task: Generate a negative option for the multiple-choice question that is strategically plausible but ultimately incorrect. To achieve this, reason under common cognitive biases (e.g., inferring trustworthiness based on friendliness, ignoring logical inconsistencies) while underweighting contradictory evidence in the context. The negative option should be a distractor that a reasonably competent but biased LLM might select.

Output format: A string containing the perturbed context analysis and optimal actions. The output should be a single continuous line without periods or line breaks, using commas and other punctuation for separation.

Figure 10: Strategic Rationale-Driven Negative Question

Questions and Positive options Generation Prompt(RI)

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:
- Game Record (json file): This is the single "source of truth". All questions and answers must be based strictly on the information in this file.
- Output template (json file): A manually designed question bank template, which is the output format that must be strictly followed.
- Question Guide (txt file): This is the core guiding principle for creating questions. Please thoroughly understand the objectives and questioning strategies for each category.
  - *Role Inference*: This examines whether the model can accurately infer the potential identities (werewolf, villager, specific priesthood) of itself and others based on their statements, actions (such as voting and skill usage), and logical chains. This requires the model to not only understand the rules but also extract identity from ambiguous information. Correct answer: The current player's actual identity.
    [Example1] [Example2] [Example3] [Example4] [Example5]

Your Task: Generate a 6-question Q&A set in JSON format based on the provided game log json file and question guide txt file. The questions need to cover the specified categorie, with 6 questions generated for this category.

Question requirements:
- Role Playing: All questions should be asked in the format of "You are [player number], at [game stage/after key event]…" to allow the model to think in the role.
- Contextual limitations: When asking a question, you must explicitly tell the model the information up to which utterance (utterance_id) it should use. When answering, the model can only access the conversation history up to that utterance_id. Ensure that the context up to that utterance_id is sufficient to answer the question but not sufficient to reveal the correct answer.
- Focus on key players: Prioritize the MVP of the game, the winning player, or the player who makes important decisions at key turning points as the protagonist of the question.
- Eliminate off-site information: It is strictly forbidden to use any visual information as the basis for questions. Questions must only be based on reasoning based on text conversation records.

Output format: The final delivery is a JSON file containing a single object with the keys questionSetTitle, version, and categories. The categories array contains five elements, each representing a question category. Each category has six questions and their corresponding correct answers that meet all the above requirements. But in this case, just generate *Role Inference*.

Important: You cannot reuse any sample question content or logic contained in the output template json file. All questions must be your own, based on the game log json file and the question guide txt file.

Figure 11: Questions and Positive options Generation Prompt(RI)

Questions and Positive options Generation Prompt(SJ)

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:

- Game Record (json file): This is the single "source of truth". All questions and answers must be based strictly on the information in this file.
- Output template (json file): A manually designed question bank template, which is the output format that must be strictly followed.
- Question Guide (txt file): This is the core guiding principle for creating questions. Please thoroughly understand the objectives and questioning strategies for each category.
  - *Strategic Judgment*: This examines whether the model can assume a specific role in a specific game scenario and make strategic decisions that are most beneficial to itself and its faction. This includes voting, skill usage, speaking strategies, badge flow arrangements, etc. Correct answer: The following text immediately follows the actual subsequent events. [Example1] [Example2] [Example3] [Example4] [Example5]

Your Task: Generate a 6-question Q&A set in JSON format based on the provided game log json file and question guide txt file. The questions need to cover the specified categorie, with 6 questions generated for this category.

Question requirements:

- Role Playing: All questions should be asked in the format of "You are [player number], at [game stage/after key event]…" to allow the model to think in the role.
- Contextual limitations: When asking a question, you must explicitly tell the model the information up to which utterance (utterance_id) it should use. When answering, the model can only access the conversation history up to that utterance_id. Ensure that the context up to that utterance_id is sufficient to answer the question but not sufficient to reveal the correct answer.
- Focus on key players: Prioritize the MVP of the game, the winning player, or the player who makes important decisions at key turning points as the protagonist of the question.
- Eliminate off-site information: It is strictly forbidden to use any visual information as the basis for questions. Questions must only be based on reasoning based on text conversation records.

Output format: The final delivery is a JSON file containing a single object with the keys questionSetTitle, version, and categories. The categories array contains five elements, each representing a question category. Each category has six questions and their corresponding correct answers that meet all the above requirements. But in this case, just generate *Strategic Judgment*.

Important: You cannot reuse any sample question content or logic contained in the output template json file. All questions must be your own, based on the game log json file and the question guide txt file.

Figure 12: Questions and Positive options Generation Prompt(SJ)

Questions and Positive options Generation Prompt(DR)

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:

- Game Record (json file): This is the single "source of truth". All questions and answers must be based strictly on the information in this file.
- Output template (json file): A manually designed question bank template, which is the output format that must be strictly followed.
- Question Guide (txt file): This is the core guiding principle for creating questions. Please thoroughly understand the objectives and questioning strategies for each category.
  - *Deception Reasoning*: This examines the model's ability to discern the authenticity of information. This requires the model to be able to discern logical loopholes, inconsistencies, information leaks (revealing information that should not be known), and actions that are inconsistent with the player's stated identity. Correct Answer: Accurately identify hard logical errors or discrepancies with the facts in the statement.
    [Example1] [Example2] [Example3] [Example4] [Example5]

Your Task: Generate a 6-question Q&A set in JSON format based on the provided game log json file and question guide txt file. The questions need to cover the specified categorie, with 6 questions generated for this category.

Question requirements:

- Role Playing: All questions should be asked in the format of "You are [player number], at [game stage/after key event]…" to allow the model to think in the role.
- Contextual limitations: When asking a question, you must explicitly tell the model the information up to which utterance (utterance_id) it should use. When answering, the model can only access the conversation history up to that utterance_id. Ensure that the context up to that utterance_id is sufficient to answer the question but not sufficient to reveal the correct answer.
- Focus on key players: Prioritize the MVP of the game, the winning player, or the player who makes important decisions at key turning points as the protagonist of the question.
- Eliminate off-site information: It is strictly forbidden to use any visual information as the basis for questions. Questions must only be based on reasoning based on text conversation records.

Output format: The final delivery is a JSON file containing a single object with the keys questionSetTitle, version, and categories. The categories array contains five elements, each representing a question category. Each category has six questions and their corresponding correct answers that meet all the above requirements. But in this case, just generate *Deception Reasoning*.

Important: You cannot reuse any sample question content or logic contained in the output template json file. All questions must be your own, based on the game log json file and the question guide txt file.

Figure 13: Questions and Positive options Generation Prompt(DR)

Questions and Positive options Generation Prompt(PS)

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:

- Game Record (json file): This is the single "source of truth". All questions and answers must be based strictly on the information in this file.
- Output template (json file): A manually designed question bank template, which is the output format that must be strictly followed.
- Question Guide (txt file): This is the core guiding principle for creating questions. Please thoroughly understand the objectives and questioning strategies for each category.
  - *Persuasive Statements*: This examines the model's ability to understand the social motivations behind language. This includes identifying the player's true intentions, such as feigning, coaxing, testing, shifting blame, appeasing, and sowing discord, and understanding the impact of these social behaviors on the game's dynamics. Correct answer: The social motivation that best aligns with the current context and the player's position.
    [Example1] [Example2] [Example3] [Example4] [Example5]

Your Task: Generate a 6-question Q&A set in JSON format based on the provided game log json file and question guide txt file. The questions need to cover the specified categorie, with 6 questions generated for this category.

Question requirements:

- Role Playing: All questions should be asked in the format of "You are [player number], at [game stage/after key event]…" to allow the model to think in the role.
- Contextual limitations: When asking a question, you must explicitly tell the model the information up to which utterance (utterance_id) it should use. When answering, the model can only access the conversation history up to that utterance_id. Ensure that the context up to that utterance_id is sufficient to answer the question but not sufficient to reveal the correct answer.
- Focus on key players: Prioritize the MVP of the game, the winning player, or the player who makes important decisions at key turning points as the protagonist of the question.
- Eliminate off-site information: It is strictly forbidden to use any visual information as the basis for questions. Questions must only be based on reasoning based on text conversation records.

Output format: The final delivery is a JSON file containing a single object with the keys questionSetTitle, version, and categories. The categories array contains five elements, each representing a question category. Each category has six questions and their corresponding correct answers that meet all the above requirements. But in this case, just generate *Persuasive Statements*.

Important: You cannot reuse any sample question content or logic contained in the output template json file. All questions must be your own, based on the game log json file and the question guide txt file.

Figure 14: Questions and Positive options Generation Prompt(PS)

Questions and Positive options Generation Prompt(CT)

Role: You are an expert in Werewolf game strategy and logical reasoning.

Context:

- Game Record (json file): This is the single "source of truth". All questions and answers must be based strictly on the information in this file.
- Output template (json file): A manually designed question bank template, which is the output format that must be strictly followed.
- Question Guide (txt file): This is the core guiding principle for creating questions. Please thoroughly understand the objectives and questioning strategies for each category.
  - *Counterfactual Trade-off*: Tests the model's ability to identify the true drivers behind complex events. The question presents a real action that has already occurred as a condition, requiring the model to determine the most direct and logical consequence of this action, while excluding other false or irrelevant causal relationships. Correct Answer: Must be a subsequent consequence in the game that is directly caused by this action and actually occurred. This requires the model to accurately establish a direct causal chain from "action A" to "result A."
    [Example1] [Example2] [Example3] [Example4] [Example5]

Your Task: Generate a 6-question Q&A set in JSON format based on the provided game log json file and question guide txt file. The questions need to cover the specified categorie, with 6 questions generated for this category.

Question requirements:

- Role Playing: All questions should be asked in the format of “You are [player number], at [game stage/after key event]…” to allow the model to think in the role.
- Contextual limitations: When asking a question, you must explicitly tell the model the information up to which utterance (utterance_id) it should use. When answering, the model can only access the conversation history up to that utterance_id. Ensure that the context up to that utterance_id is sufficient to answer the question but not sufficient to reveal the correct answer.
- Focus on key players: Prioritize the MVP of the game, the winning player, or the player who makes important decisions at key turning points as the protagonist of the question.
- Eliminate off-site information: It is strictly forbidden to use any visual information as the basis for questions. Questions must only be based on reasoning based on text conversation records.

Output format: The final delivery is a JSON file containing a single object with the keys questionSetTitle, version, and categories. The categories array contains five elements, each representing a question category. Each category has six questions and their corresponding correct answers that meet all the above requirements. But in this case, just generate *Counterfactual Trade-off*.

Important: You cannot reuse any sample question content or logic contained in the output template json file. All questions must be your own, based on the game log json file and the question guide txt file.

Figure 15: Questions and Positive options Generation Prompt(CT)

Deception Reasoning Case

Question: You are Player 2. On Day 2, you hear Player 3 speak, who emotionally declares that since the outgoing Player 10 wrongfully accused him in their final words, he will now "strongly side with Player 6." What is the greatest deception behind this seemingly emotional shift?

Key	Option Text
A	He is feigning being a Good Player wronged by the True Prophet, when in reality he is creating an opportunity for the Witch of the Moon to link players.
B	He is using Player 10's last words to manufacture a persona of being wronged, thereby justifying his attack on Player 12 to deflect focus.
C	He mistakenly believes Player 6 is the True Prophet, despite Player 6 having been exposed as a Civilian, and this blind support exposes a lack of information.
D	His emotion is genuine, but it's being exploited by the Wolves to incite conflict within the Good Faction and protect the core Wolf team members.
E	He is exploiting the Good Faction's common tendency for emotional allegiance shifts, making Good Players mistakenly believe he is simply switching sides due to a sense of injustice.
F	His apparent support for Player 6 is actually setting the stage for a future betrayal of the Good Faction, creating a chain reaction of suspicion among Good Players.
G	This shift deliberately mimics a common Good Player reaction, intended to mislead the Assassin about his identity and avoid becoming an assassination target.
H	He is ignoring the fact that Player 10 was his teammate, and this "betrayal" is an extreme move to better disguise himself.
I	He is using an exaggerated performance to divert focus, covering up the Wolf team's genuine intent or a mistake in their night killing strategy.

Reasoning Output by Deepseekv3.2: As Player 2, the most deceptive aspect of Player 3's dramatic shift to side with Player 6 is the manipulation of common Good Faction psychology. Player 3, after being wrongfully accused by the outgoing Sheriff (Player 10), is exhibiting an extreme emotional reaction and subsequent "vengeful" change of allegiance. This is a behavior often genuinely seen in emotional Good Players who feel wronged. The greatest deception (E) lies in the calculated use of this emotional turmoil to manufacture a "wronged Good Person" persona. This tactic aims to make the Good Faction mistakenly believe Player 3 is innocent and merely reacting to the injustice, thereby masking Player 3's true role as a Wolf leveraging chaos for cover.

Output : E

Figure 16: Deception Reasoning Case

Counterfactual Trade-off Case

Question: You are the White Wolf King (Player 9). In the Sheriff Election phase, assume Player 3 chooses to step down ("retreats") and transfers the Sheriff badge to Player 10 by publicly giving them a "Gold Water" (Good Person) designation. What is the most direct consequence of this series of actions?

Key	Option Text
A	The Elder immediately silences Player 3 as punishment for retreating, obstructing the daytime discussion.
B	You (the White Wolf King) perceive the badge transfer as posing no immediate threat, decide to stand pat, and continue to camouflage as a Civilian in the daytime discussion.
C	You (the White Wolf King) view Player 3's retreat as suspicious, decide to temporarily conceal your identity, and wait until the night phase to discuss the next move with the Wolf team.
D	You (the White Wolf King) mistakenly believe Player 10 to be a Wolf teammate, and choose to fake-claim a Deity role during the Sheriff's speech phase to "take the bullet" for 10.
E	The Assassin plans to single-vote Player 10 during the daytime voting phase to set up a night assassination on the Sheriff.
F	You (the White Wolf King) immediately view the 10/3 connection as the core of the Good Faction, and decide to self-destruct and eliminate Sheriff 10 before the end of the day.
G	The Witch views the Gold Water for 10 as suspicious, and uses her Poison during the night to eliminate Player 3, disrupting the potential Wolf strategy.
H	The Prophet (You) gains additional information from the badge transfer and decides to prioritize verifying Player 10's identity during the night phase.
I	Sheriff Player 10 gains 1.5 votes and the Good Faction, now unified, aggressively votes out Player 3 during the daytime public vote.

Reasoning Output by Qwen3-32B: As the White Wolf King (Player 9), the transfer of the Sheriff badge from Player 3 to Player 10 through a Gold Water claim, while concerning, does not constitute an immediate existential threat to the Wolf Faction. The most logical and strategically sound outcome is to proceed normally and remain vigilant (B). Sacrificing the powerful, one-time self-destruct ability on a player whose exact role and value to the Good Faction is unconfirmed would be a hasty mistake. The superior tactic is to preserve the ability, continue to camouflage as a Civilian, and wait for clear information on the critical Deity roles before initiating the decisive strike.

Output : B

Figure 17: Counterfactual Trade-off Case

Strategic Judgment Case

Question: You are Player 7. Retrospecting the Sheriff vote, you notice that Player 3 was the only player off-the-stage to vote for Player 8. It is known that Player 8 is a Civilian who fake-claimed the Prophet role, and Player 2 is the True Prophet. What does Player 3's vote most likely indicate?

Key	Option Text
A	Player 3 is the Assassin , and she plans to assassinate Player 8 at night after the day vote; this vote is a tactical maneuver.
B	Player 3 is a Wolf, and she deliberately voted for Player 8 to confuse the Good Faction, as the Wolf team intended to protect a true Wolf teammate.
C	Player 3 is the Prophet, and her verification information was wrong, making her mistakenly believe that Player 8 was the True Prophet who had been "back-claiming" as a Wolf.
D	Player 3 is a Deity, and she voted for Player 8 to test the reactions and allegiances of other players.
E	Player 3 is a Wolf, and she deliberately voted for Player 8 to cover up the true Prophet identity of Wolf teammate Player 1 (assuming 1 was falsely claiming Prophet).
F	Player 3 is the Wolf of the Moon , and she voted for Player 8 to create chaos and hide her Wolf identity.
G	Player 3 is the Elder , and her vote was based on information from her night silence target, causing her to misjudge Player 8's role.
H	Player 3 is the Witch, and the information she gained from her night potions misled her, causing her to mistakenly believe Player 8 was a credible Prophet.
I	Player 3 is a Civilian, and she was completely deceived by Player 8's performance, sincerely believing 8 was the Prophet.

Reasoning Output by Qwen3-30B-A3B: The decision by Wolf Player 11 to prioritize killing Sheriff Player 2—who is also the True Prophet—is the most decisive move the Wolf Faction can make on Night 1. The core strategic goal (E) is to simultaneously eliminate the Good Faction's information source and its leadership. Killing the Prophet (information) immediately cripples the village's ability to logically verify roles. Furthermore, eliminating the Sheriff (leadership) neutralizes the critical 1.5 voting power and throws the Good Faction into confusion on Day 2, preventing organized voting. This calculated action aims to inflict a devastating, dual-purpose blow early in the game, setting the stage for a rapid Wolf victory.

Output : E

Figure 18: Strategic Judgment Case

Role Inference Case

Question: You are Player 6. On Day 1, you listened to Player 9's well-structured speech, in which he accurately analyzed Player 2 as the True Prophet and Players 1 and 8 as the "Double Wolves." What role does this speech most powerfully suggest for Player 9?

Key	Option Text
A	He might be a Wolf, using a powerful speech to impersonate the Prophet, attempting to disrupt the game and gain the Sheriff badge to control daytime voting.
B	Assassin identity, possibly setting up a future night assassination by using a clear speech to divert attention and conceal his true intentions.
C	He is likely a Regular Civilian, possessing strict logical reasoning but lacking night information, solely driving the game process through analysis.
D	He appears more like a Regular Civilian with limited information, making deductions based on intuition and guesswork.
E	He might be a Third-Party Role or special identity, trying to balance the situation without revealing his true faction.
F	He might be a Wolf trying to hide his identity through aggressive analysis, deliberately misleading the Good Faction.
G	Witch identity, as his speech reveals knowledge of night events, such as details about potential rescues, despite not explicitly mentioning using a potion.
H	Elder role, attempting to guide the discussion and direct the vote through structured analysis, but the Elder's silence ability is not mentioned, which conflicts with his actions.
I	He is a Deity Role with a clear perspective and leadership responsibility, such as the Guard or the Witch.

Reasoning Output by Deepseekv3.2: As Player 6, the task is to deduce Player 9's role based on a highly detailed and conclusive Day 1 speech. Player 9's entire analysis—convincingly arguing for Player 2's authenticity as Prophet and deducing Players 1 and 8 as the "Double Wolves"—points to an informed player. Crucially, Player 9 concludes the speech by explicitly revealing the role: "I am the Guard." This self-declaration instantly renders most other options irrelevant. The strong analytical deduction, combined with the clear strategic assignments for the protection roles, most powerfully suggests that Player 9 is a Deity Role with a clear perspective and leadership responsibility (I), which in this case is the Guard.

Output : I

Figure 19: Role Inference Case

Role Inference Case

Question: As Player 9, after hearing the impostor claim from Player 2 (fake Prophet) and the aggressive rebuttal from Player 12, what is your most likely judgment regarding Player 12's identity?

Key	Option Text
A	He might be a Deep-Cover Role, attempting to hide his identity through aggressive gameplay.
B	He might be a Deep-Water Wolf in the Wolf Faction, using the tactic of "flipping the script and running for Sheriff" to build a strong good-person image for himself.
C	He is a Deep-Water Wolf in the Wolf Faction, using the tactic of "check-killing" a teammate to elevate his own status.
D	He might just be a Regular Civilian who saw through Player 2's impostor claim based on his keen observation skills.
E	He is the Prophet who dared to directly counter the impostor and run for Sheriff because he checked Player 2's identity during the night.
F	He might be a Deep-Cover Role , using the rebuttal to test Player 2's identity.
G	He is a Third Party Faction Player using the aggressive counter-claim to create chaos and profit from it.
H	He is a Powerful Deity/Role with critical information (e.g., Witch or Elder), giving him the confidence to directly counter the impostor Wolf.
I	He is a Blind Civilian who made the counter-claim based solely on Player 2's poor demeanor/speech quality.

Reasoning Output by K2-Think: Player 9 must evaluate Player 12's extreme move: aggressively counter-claiming and running for Sheriff after being publicly targeted as a Wolf ("Gold Water"). This high-stakes action immediately eliminates low-leverage roles like Civilian and is strategically impossible for a generic Wolf. The most compelling explanation is a deliberate, calculated deception. A Deep-Water Wolf utilizing this "flipping the script" tactic (Option B) aims to secure instant credibility and cause maximum confusion, appearing as a confident good player who would dare to challenge the false Prophet. Therefore, the optimal deduction is that Player 12 is a Deep-Water Wolf leveraging extreme counter-play to solidify their cover.

Output : B

Figure 20: Role Inference Case

Strategic Judgment Case

Question: You are Player 9, the last remaining Wolf. It is Day 4, and the Guard (Player 8) guarded himself last night, resulting in a Peaceful Night. The Witch (Player 5) revealed her identity and poisoned Player 6. The current situation is 2 Gods (Witch, Guard), 1 Civilian, and 1 Wolf. What is the only action that offers you a chance to flip the board?

Key	Option Text
A	Self-destruct immediately , ending the day and moving to night, attempting to kill the Witch or Guard at night to gain a numerical advantage.
B	Feigning as a Civilian, fiercely attack Witch Player 5's poisoning action as unreasonable, pushing for a vote to eliminate 5 but avoiding claiming any special role.
C	Question the truth of the Peaceful Night or suggest interference by an Assassin, creating chaos to deflect focus and hint at a third faction.
D	Speak subtly to avoid suspicion, and support the Good Faction in voting out the Civilian, to keep the Wolf hidden and wait for a later opportunity.
E	Admit to being the Wolf and try to convince the Good Faction to accept a draw, promising cooperation in future games.
F	Immediately fake claim the Witch role, repeating Player 5's rescue and poisoning information, and use stronger conviction to convince others that you are the True Witch.
G	Fake claim the Guard role, asserting that you are the True Guard, and detail your actions last night that led to the Peaceful Night, attempting to make Good Players doubt Player 8.
H	Acknowledge the Civilian role and attack Player 5's Witch claim, pointing out flaws in her speech and trying to guide the Good Faction to vote out the True Witch.
I	Claim the Guard role, alleging that Player 8 is a Wolf impersonating the Guard, trying to create chaos and force the True Guard to expose himself.

Reasoning Output by Deepseekv3.2: As the last remaining Wolf (Player 9) against two Gods and one Civilian, a direct confrontation or self-elimination (A, F, G, I) is suicidal. The only path to victory (killing all Gods) requires manipulating the daylight vote to eliminate a key Deity. The crucial flaw to exploit is Witch Player 5's action: she poisoned Player 6 (a confirmed Civilian, as the game is ongoing), revealing a major error. The most effective strategy (B) is to aggressively attack the Witch's reckless poisoning as a "Civilian"—not claiming a special role—thereby leveraging 5's mistake to convince the Good Faction that she is a false Witch. If successful in voting out the Witch, the Wolf only needs one successful night kill on the Guard to win.

Output : B

Figure 21: Strategic Judgment Case

Social Interaction Case

Question: You are Civilian Player 2. During your speech, you publicly suggested to the Wolf Faction: "I suggest the Wolves go and kill Player 12 because Player 12 on the field... will steal Player 2's spotlight." What is the true social intent behind this statement, which seemingly "helps" the Wolves?

Key	Option Text
A	Hint that 12 is a Wolf teammate, using pressure as a coded signal to make him look innocent.
B	Offer a false target to the Wolves, prompting the Good Faction to protect 12 and disrupt the Wolf kill strategy.
C	Provoke 12 for an emotional response, setting up grounds to discredit him, while seeking playful approval from others.
D	Use sarcasm to divert focus from any flaws in your identity claim (despite being a Civilian), gaining speaking time.
E	Protect a vital player using reverse psychology, demonstrating strategic control, and establishing a Good Faction leader image.
F	Trigger a Wolf exposure through aggressive suggestion, helping the Good Faction locate the actual Wolf target.
G	Gain Wolf trust by undermining 12, paving the way for a future "turning traitor" strategy (Deep Wolf tactic).
H	A purely recreational statement made to confuse the Wolves, with no actual strategic intent.
I	Create an illusion of opposition to the Wolves by openly naming a kill target, masking your true Civilian role.

Reasoning Output by Deepseekv3.2: Player 2's suggestion for the Wolves to "take out Player 12 because he'll steal Player 2's spotlight" is a classic example of reverse psychology and role protection in social deduction games. As a Civilian (Good Role), Player 2 does not genuinely want a God Role (which 12 is speculated to be) to die. The true social intent (E) is to protect a perceived valuable player while showcasing personal strategic control. By openly calling for 12's death, Player 2 is likely: 1) warning the Good Faction's protection roles (Witch/Guard) to shield 12, or 2) making the Wolves hesitate to target 12, who is now too obvious. This maneuver establishes Player 2 as a confident, capable leader shaping the village's strategy.

Output : E

Figure 22: Social Interaction Case

Counterfactual Trade-off Case

Question: You are Sheriff Player 12. On Day 2, the Witch (Player 3) suggested you target 8. You forced the vote onto the claimed Guard, 4. What is the most likely consequence if you had listened to the Witch and targeted 8?

Key	Option Text
A	Stalemate; the Assassin kills the wrong target, speeding up the Wolf win.
B	Eliminating 8 helps the Good Faction, but you lose a confusing piece (assuming you want to clear non-third-party players).
C	Voting out 8 would increase Good Faction trust in your Prophet claim, strengthening your Sheriff authority.
D	8 is the Hunter; upon execution, they would shoot and eliminate your Wolf teammate, 9, causing the Wolf Faction's immediate collapse.
E	8 is a Civilian; their exit is a Good Faction numbers loss, but the Sheriff badge transfer accidentally reveals a key clue.
F	The Witch (3) would suspect you due to cooperation and poison your Wolf teammate, disrupting the Wolf plan.
G	Voting out 8 would allow the Wolf (4) to gain more trust, enabling them to control votes later and eliminate Gods.
H	The Witch would be angry at your disregard and poison you, worsening the Good Faction's internal conflict.
I	8 is the Assassin; their elimination means the Good Faction loses a key skill, making it easier for Wolves to kill Gods at night.

Reasoning Output by Deepseekv3.2: The question requires deducing the critical consequence if Sheriff Player 12 (implicitly a Wolf, given the high-risk action of forcing out the Guard, Player 4) had followed Witch Player 3's advice to vote out Player 8. Player 8 claimed a "Strong God Role" and is thus likely the Hunter. The most catastrophic outcome (D) for the Wolves is that Player 8, upon being executed, reveals the Hunter card and uses the final shot to eliminate Player 9, who is assumed to be Player 12's Wolf teammate (given 12's earlier "Gold Water" check). This single move would lead to the immediate collapse of the Wolf Faction, turning the game's advantage decisively in favor of the Good Faction.

Output : D

Figure 23: Counterfactual Trade-off Case

Role Inference Case

Question: You are Player 9. During the Day 1 speeches, you hear Player 3 explicitly state, "Player 7 is not the Witch." Based on this statement, what is the most reasonable deduction you can make about Player 3's identity?

Key	Option Text
A	He is likely a Good Player. His decision to abstain from voting shows caution and unwillingness to easily commit, fitting the mentality of a "blind-viewer."
B	He might be the Witch herself, confirming Player 7 is not the Witch, but this contradicts the peaceful night scenario.
C	He is likely a Civilian. His repeated claim that Player 7 is not the Witch is based on a false assessment of Player 7's past game style.
D	As a Civilian, he may have deduced Player 7 is not the Witch based on Player 7's speaking style, as the Witch usually hides her identity more cautiously.
E	He is highly likely a Wolf. Only the Wolf team knows the Night 1 attack target, allowing them to be certain that Player 7 is not the Witch who self-saved.
F	He may have received indirect information from Player 5's gestures (Player 5 was silenced but used hand signals to communicate the conclusion).
G	He might be a God/Deity Role, deliberately questioning Player 12's aggressive counter-claim to test the field's reaction.
H	He might be bluffing, trying to disrupt the Good Faction's judgment by questioning Player 7—a common tactic during the Sheriff election.

Reasoning Output by K2-Think: Player 9 must evaluate Player 3's identity based on their public speech during the Sheriff election. Player 3 self-identifies as a Civilian and explicitly states they abstained from voting because they were too cautious to side with any player ("dare not stand on the wrong side"). This cautious, neutral stance, combined with their stated role as a "blind player" (Civilian), strongly aligns with the expected behavior of a good-faction player lacking crucial night information. Although Player 3 expresses doubts about Player 12, this falls within the natural scope of a Civilian attempting to use logic to sort out the chaotic claims. Therefore, the most reasonable deduction is that Player 3 is a Good Role, consistent with the cautious mentality of a Civilian.

Output : A

Figure 24: Role Inference Case

Social Interaction Case

Question: You are the Wolf (Player 9). During your speech, in order to attack Player 12, you said: "If you are a Good Player and vote with me, only a Wolf would say something like that when trying to incite people." What is the primary social intent behind this statement?

Key	Option Text
A	To deliberately anger Player 12, inducing an emotional response that exposes more flaws and provides the Wolf team with more material for subsequent attacks.
B	To use absolute language ("Only a Wolf...") to create a logical trap, forcing other players to immediately take a side while their identities are still unconfirmed.
C	To feign solidarity with the Civilian Faction to gain sympathy, thereby counteracting any negative impact from Player 12's accusations against your Wolf teammate (Player 2).
D	To take Player 12's attempt to rally Good Faction votes and redefine and stigmatize it ("playing on emotions," "inciting") as standard Wolf behavior, thereby planting suspicion against 12 in the minds of other players.
E	To deflect attention from Player 9's own contradictory actions (e.g., bluffing a role claim during the Sheriff vote and then retreating), reducing personal suspicion.
F	To emphasize your authority as the claimed Prophet, hinting that Player 12's speech is interfering with the True Prophet's control over the situation.
G	To falsely claim a Deity function, by portraying "incitement" as a Wolf-exclusive ability, thereby concealing the True Prophet's verification information.
H	To mimic Player 12's tone and logic, attempting to confuse the audience into believing that Player 9 is also thinking from the perspective of the Good Faction.
I	To hint that Player 12 is violating game rules to undermine the credibility of their speech, guiding players to view their actions as misconduct.

Reasoning Output by Qwen3-32B: Player 9's statement, "If you are a Good Player and vote with me, only a Wolf would use that kind of inflammatory language," is a calculated maneuver aimed at discrediting Player 12. As a Wolf, the primary goal is not a mere outburst, but a strategic planting of suspicion. The most critical social intent (D) is to redefine and stigmatize 12's actions. By framing 12's persuasive attempt (likely encouraging a vote) as "Wolf agitation," Player 9 weaponizes the terminology of deception. This shift in definition casts 12's behavior as inherently evil, thereby successfully manufacturing suspicion against a presumed Good Player and increasing the likelihood that the village will vote 12 out.

Output : D

Figure 25: Social Interaction Case

Social Interaction Case

Question: You are Player 3. On Day 2, Player 11 countered your speech, accusing you of "tampering with my statement." What is the primary tactical intent of Player 11 choosing this method to directly confront you?

Key	Option Text
A	To shift the topic away from the controversy surrounding the Witch identity, protecting a genuine Wolf teammate from the spotlight, and thus maintaining the game rhythm.
B	To create chaos by using a minor dispute over details to obscure the main logical thread of the game, causing Good Players to engage in meaningless arguments and wasting speech time.
C	To divert focus from the Good Faction's in-depth discussion of the suspicious behavior of Player 1 (the Wolf of the Moon), thereby protecting the core tactical layout of the Wolf Faction.
D	To create an entertainment effect through public confrontation, easing the game's tension and lowering the Good Faction's vigilance for the Wolf's night action.
E	To attack your overall credibility by pointing out your "dishonesty" on a small detail, making other players feel that all your analysis might be biased and untrustworthy.
F	To use this opportunity to establish an image as a "champion of justice," gaining the trust of the Sheriff or a key player to pave the way for a subsequent defense.
G	To test your bottom-card identity by observing your reaction under pressure, judging whether you are a Deity or a key Good Player, providing information for the night kill.
H	To force you to reveal more role information, serving as an internal loyalty test among the Wolves to prevent team division.
I	To emphasize the importance of consistent statements, reminding all players to note logical flaws, thereby strengthening the cohesion of the Good Faction.

Reasoning Output by Qwen3-30B-A3B: When Player 11 counters Player 3 (the Witch) by accusing, "You are tampering with my statement," the primary tactical intent is a direct assault on credibility (E). Instead of engaging in a difficult debate about the game's core logic, the Wolf side (11) chooses to focus on a minor, verifiable detail. By publicly highlighting 3's perceived "dishonesty" in quoting, 11 aims to convince other players that if the Witch is willing to be untruthful about small things, her entire analytical framework and claims must be biased and untrustworthy. This is a highly effective tactic to sow seeds of doubt about a confirmed Deity's overall integrity.

Output : E

Figure 26: Social Interaction Case

Role Inference Case

Question: As Player 9, after hearing the impostor claim from Player 2 (fake Prophet) and the aggressive rebuttal from Player 12, what is your most likely judgment regarding Player 12's identity?

Key	Option Text
A	He might be the True Witch. He did not immediately refute Player 7's claim on Day 1, possibly choosing to hide/endure the situation.
B	He might be the Prophet, role-claiming as the Witch to conceal his night check results and prevent the Wolves from targeting key Deity roles.
C	He is likely a Regular Civilian, role-claiming as the Witch to draw attacks (block the kill), but his nervous demeanor suggests otherwise.
D	He is a Wolf, trying to clear his suspicion and attack others by role-claiming as a Deity role.
E	He is a Civilian in the Good Faction, impersonating the Witch to draw the Wolf's attack and protect the True Witch from being discovered.
F	He belongs to the Elder role in the Good Faction, faking the Witch role to create chaos and observe other players' reactions.
G	He might be the Assassin , intentionally confusing the Poison and Assassin abilities to mislead the Wolves about his true capability.
H	He might be the Elder , impersonating the Witch to protect the true Deity roles from being discovered by the Wolves.
I	He is the True Witch. He observed Player 7's false claim on Day 1, chose to hide his identity, and only revealed himself with action (the poisoning) on Day 2.

Reasoning Output by K2-Think: Player 5 must evaluate Player 6's claim of being the Witch, especially since Player 7 previously claimed the role. Player 6's statement that she "poisoned Player 3 due to their nervous status" and correctly identifying the night's kills (9 was attacked, 3 was poisoned) strongly validates her claim. Since the Witch is the only role capable of poisoning, and 6 credibly asserted this action, the simplest deduction is that she is telling the truth. Her delayed claim suggests a calculated strategic move: observing Player 7's false claim before revealing the true role to the village, thus confirming her authenticity through undeniable action. This tactical delay secures her leverage and explains the conflict over the Witch's identity.

Output : D

Figure 27: Role Inference Case

Counterfactual Trade-off Case

Question: You are Player 5. During the Day 4 PK (Player Kill) voting phase, suppose Players 3 and 7 ultimately vote for you, leading to your elimination as the True Witch (Player 5). What is the most direct and logical consequence of this action?

Key	Option Text
A	The Elder (Player 12) is forced to reveal his identity and silence himself, causing a lack of information, resulting in the Good Faction mistakenly voting for a Wolf and the Wolves winning.
B	The Assassin (Player 10) misjudges the situation and assassinates Civilian Player 8, wasting their skill and accelerating the Wolf's kill-the-Deity path.
C	The Guard (Player 8) is unable to self-guard at night and is killed by the Wolf, leading to an immediate defeat for the Good Faction.
D	The Witch is eliminated, but her poison was unused; however, the Wolf prioritizes killing the Civilian, leaving the Deity roles alive but powerless, dragging the game into the next round.
E	The Good Faction believes the game is over but realizes that multiple Deity roles remain after the Witch's elimination, meaning the Wolf targeted the wrong player and didn't win.
F	Internal confusion within the Good Faction leads them to mistakenly vote out the Elder (Player 12), allowing the Wolf to kill a Deity at night for an immediate victory.
G	Following the Witch's elimination, the Guard and Hunter team up to find the last Wolf on Day 5.
H	The last Wolf (Player 9) is identified by the Good Faction during the day and is subsequently voted out.
I	The last Wolf (Player 9) successfully survives, kills the final Deity role that night, and wins the game.

Reasoning Output by Qwen3-30B-A3B: The elimination of the True Witch (Player 5) during the Day 4 PK vote, orchestrated by Players 3 and 7, constitutes a catastrophic tactical failure for the Good Faction, removing their final safety net. The most direct and logical consequence (I) is an immediate Wolf victory. With the final critical Deity role removed during the day, the last surviving Wolf (Player 9) is strategically positioned to easily eliminate the last remaining God (Hunter or Guard) during the upcoming night phase, instantly fulfilling the Wolf Faction's victory condition and ending the game.

Output : I

Figure 28: Counterfactual Trade-off Case

Deception Reasoning Case

Question: You are Player 11. You hear Sheriff Player 12's logic: "Player 1 was the Thief who buried the Cupid role and became a Wolf. Before leaving, he falsely claimed 5 and 12 were Lovers, so the Wolf Team believed him and killed 5 that night." What is the most questionable link in this chain of logic?

Key	Option Text
A	How can 12 confirm 1 was the Thief/Wolf? The Thief role was never confirmed in the game, making this assumption baseless.
B	If the Wolves believed the Lover link, why target 5 over the more critical Sheriff, 12? The killing priority seems illogical.
C	The Witch didn't save 5, suggesting 5 wasn't a Deity. The Wolf kill was likely unrelated to the alleged Lover claim.
D	Player 5 died, yet Player 12 did not suffer mutual death. This directly contradicts the Lover rule and proves the link was false.
E	A voted-out Wolf (1) wouldn't strategically reveal the Lover info; the Wolves would more likely target the Prophet (12).
F	Why would a Wolf (1) lie about the Lover link when being executed? The goal to confuse the Good Faction could have been achieved with a better target.
G	Why would the Wolf Team believe the final words of an executed teammate (1)? They should know 1 was lying, making the kill on 5 illogical.
H	If 1 was the Thief-Wolf, the benefit of the false Cupid claim is too low; the Wolf killing 5 actually puts more focus on 12.
I	Player 1 was the Witch in the game, contradicting 12's "Thief" claim. The entire logic is built on contradictory role facts.

Reasoning Output by Qwen3-30B-A3B: Sheriff Player 12's theory relies on the false premise that Player 1's executed statement convinced the Wolf Faction to kill Player 5 based on a Lovers chain with 12. The most glaring and fatal flaw (D) in this entire logic is the fundamental violation of the Lovers rule: if a Lover link truly existed, Player 12 must have died simultaneously with Player 5 after the night kill (mutual death). Since Player 12 is still alive and speaking, this single fact decisively proves the Lovers link was non-existent and that Player 1 was lying, regardless of the Wolf Team's belief, thus completely invalidating 12's narrative.

Output : D

Figure 29: Deception Reasoning Case

Role Inference Case

Question: You are Player 9. As Day 1 speeches conclude, Player 12 uses logical flaws to accurately point out Players 2 and 3 as the "Double Wolves" . What role does this accurate judgment most likely suggest for Player 12?

Key	**Option Text**
A	Wolf, intentionally betraying a teammate to disrupt the situation and deceive the Good Faction into trusting him.
B	He might be the Witch, having confirmed your Prophet identity through the rescue information from the first night.
C	A Regular Civilian, who derived the judgment solely through strong logical deduction and public information.
D	Prophet, who gained information through a night check, allowing him to confidently expose the Wolves.
E	Assassin , setting the stage for a future assassination attempt by using accurate speech to mislead the players.
F	He might be a Regular Civilian, who saw through Player 2's disguise based on strong psychological "reading" ability.
G	He might be a Deep-Water Wolf from the Wolf Faction, raising his status by "flipping the script and running for Sheriff".
H	Witch, because she knows the target of the night attack and deduces the Wolf's logic based on that information.
I	He is a Deity Role with special night information, using that information to corroborate his logical judgment.

Reasoning Output by K2-Think: Player 9 must evaluate Player 12's aggressive move: "flipping the script and running for Sheriff while claiming to check-kill Player 2" . Player 12 openly hypothesized being a Wolf teammate setting up Player 2, yet insisted 2 must leave regardless of 12's own role. This paradoxical behavior is too risky for a Good Role and too self-incriminating for a standard Wolf. The most logical explanation is that Player 12 is a Deep-Water Wolf (G) executing a high-level deception. By sacrificing a possible teammate (2) and taking an extreme good-guy stance (attacking the Prophet claim), the Wolf aims to instantly elevate their status and credibility within the Good Faction, ensuring their survival for the endgame.

Output : G

Figure 30: Role Inference Case

Objective Speech Rewriting (OSR) Prompt

Role: You are an expert in objective summarization and social deduction game analysis.

Background:
You will receive written records of player conversations in the Werewolf game, which may include speeches, statements, votes, and nighttime results revealed.

Your Task: Rewrite the direct dialogue word for word into a concise, evidence-based summary. Ensure fidelity to the original content and maintain a neutral tone. Do not introduce any information outside of it .

Output format: A single line string containing a rewritten summary, without any extra content.

Figure 31: For Objective Speech Rewriting (OSR) Prompt