Multi-agent KTO: Enhancing Strategic Interactions of Large Language Model in Language Game

Rong Ye[1,2*], **Yongxin Zhang**[2*], **Yikai Zhang**[1,2†], **Haoyu Kuang**[1,2†],
Peng Sun[2✉], **Zhongyu Wei**[1,3✉]
[1] Fudan University [2] Bytedance Seed [3] Shanghai Innovation Institute
`{yerong, zhangyongxin.yx, wanhesong}@bytedance.com`
`{hykuang23, ykzhang22}@m.fudan.edu.cn`, `zywei@fudan.edu.cn`

Abstract

Achieving Artificial General Intelligence (AGI) requires AI agents that can not only make strategic decisions but also engage in flexible and meaningful communication. Inspired by Wittgenstein's language game theory, we propose that language agents can learn through in-context interaction rather than traditional multi-stage frameworks that separate decision-making from language expression. Using *Werewolf*, a social deduction game that tests language understanding, strategic interaction, and adaptability, as a test bed, we develop the Multi-agent Kahneman-Tversky's Optimization (MaKTO). MaKTO engages diverse models in extensive gameplay to generate unpaired desirable and unacceptable responses, then employs KTO to refine the model's decision-making process. In 9-player *Werewolf* games, MaKTO achieves a 61% average win rate across various models, outperforming GPT-4o and two-stage RL agents by relative improvements of 23.0% and 10.9%, respectively. Notably, MaKTO also demonstrates human-like performance, winning 60% against expert players and showing only 48.9% detectability in Turing-style blind tests. Code and data are available at project page `https://reneeye.github.io/MaKTO.html`.

1 Introduction

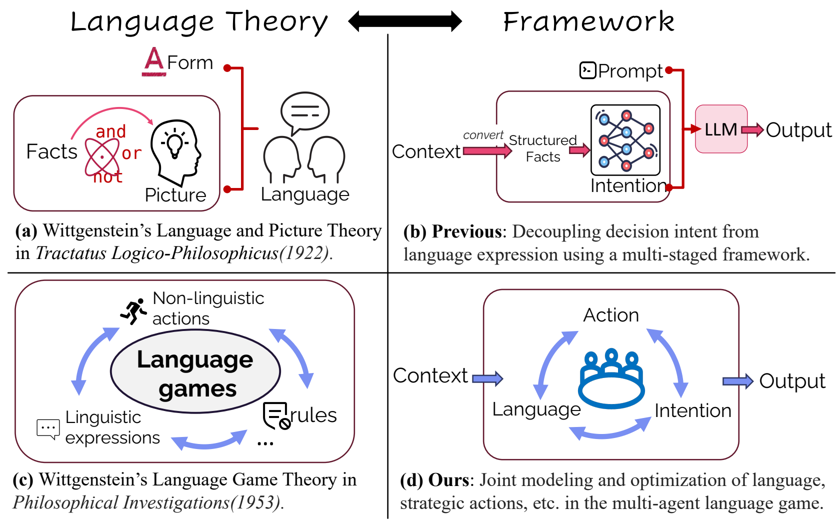
Building language agents capable of both decision-making and dialogue represents a crucial pathway toward Artificial General Intelligence (AGI) [1, 2, 3, 4]. This pursuit necessitates a deep understanding of the intrinsic relationship between language and intelligence. Wittgenstein's Language Game Theory offers a profound insight: meaning emerges from linguistic and non-linguistic interactions, regulated by social rules, forming *language games* (Fig.1c) [5, 6]. This contrasts with his earlier Tractatus view of language as logical reality-mapping (Fig.1a). This theoretical perspective points to the value of grounding AI development in practical language use and authentic interactive contexts [7, 8, 9, 10, 11].

Social deduction games (SDGs) serve as excellent testbeds for validating these theoretical principles, with self-contained language-game environments that test multiple capabilities including linguistic skills, strategic social interaction, and adaptability [12, 13, 14, 15, 16, 17]. These games provide quantifiable metrics like completion and win rates, enabling direct comparisons between human and AI performance. The *Werewolf* game[3], as a popular and typical social deduction game, exemplifies these characteristics, making it a challenging testbed for AI agent [18, 19, 20, 21, 22].

Current approaches to building AI agents for *Werewolf* game often decouple language processing from decision-making [23, 21, 24], echoing the limitations of the picture theory in Fig.1a — where complex social dynamics are oversimplified into rigid representations. For example, Wu et al. [21],

[*] Equal contribution.
[†] Work done while intern at Bytedance.
[✉] Corresponding authors.
[3] Also known as *Mafia*, detailed game introduction is in Appendix A.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Wittgenstein's Language and Picture Theory in *Tractatus Logico-Philosophicus(1922)*.

(b) Previous: Decoupling decision intent from language expression using a multi-staged framework.

(c) Wittgenstein's Language Game Theory in *Philosophical Investigations(1953)*.

(d) Ours: Joint modeling and optimization of language, strategic actions, etc. in the multi-agent language game.

Figure 1: **Language Theory and AI Architecture: Traditional vs. Language Game Models**.*(a) and (b)*: Multi-staged framework that separates language and decisions. *(c) and (d)*: our proposed framework inspired by Wittgenstein's language game theory, integrating language, actions, and intentions in a multi-agent game.

as illustrated in Fig. 1b, applied an classic RL model for decision-making or intention generation, then followed by an LLM for dialogue generation. However, they compressed the language input into structured facts, limiting generalization and cross-environment strategy transfer.

However, the paradigm of separating language processing from decision-making essentially contradicts the central claim of Wittgenstein's theory of language games - that there is an intrinsic unity of language, intention, and action [25]. Motivated by this, we introduce Multi-agent Kahneman & Tversky's Optimization (Multi-agent KTO, **MaKTO**), where the language models learn through direct interactions with different agents or models. Specifically, MaKTO *1)* builds on KTO for efficient training, *2)* employs multi-agent gameplay with a *diverse* model pool instead of *self-play*, to enhance the generalization, and *3)* uses *stepwise* preference selection preference data selection using heuristic, voting-based, and verifier-based methods, rather than simple win-loss outcomes. To help the model rapidly acquire domain-specific knowledge of terminology and strategies in Werewolf, our training process also incorporates behavior cloning using game-specific terms, strategy guides, and expert-annotated gameplay records with chain-of-thought [26] before MaKTO to optimize action decision-making in gameplay.

We perform various experiments, including tournament evaluation, Turing-style detectability test, behavioral analysis, generalization ability test, and ablation studies. The experiments show that MaKTO achieved an average win rate of 61% in 9-player Seer-Witch-Guard games against various models such as GPT-4o, Claude-3.5, and multi-staged RL agent. This outperformed both GPT-4o and the two-stage RL agent, with relative improvements of 23.0% and 10.9%, respectively. In head-to-head matches against human expert players, MaKTO reaches a comparable average win rate of 60%. Also, its conversational style is less distinguishable from humans, with only 48.9% accuracy in the Turing-style detectability test. Our contributions are:

- We propose Multi-agent KTO (MaKTO), a method that enhances LLMs' strategic reasoning in the game environment through *multi-agent interactions*, without requiring paired data.
- We create a large-scale dataset of expert Werewolf players' utterances and actions during gameplay, as well as the abundant COT behind their decisions, allowing for effective behavior cloning and fine-tuning of LLMs.
- We perform extensive experiments to show that our model achieves human-level performance and strong generalization capabilities across different game settings.

2 Our Approach

In this section, we describe in detail our training method (Fig. 2) in detail, including expert data collection, behavior cloning, and multi-agent KTO.

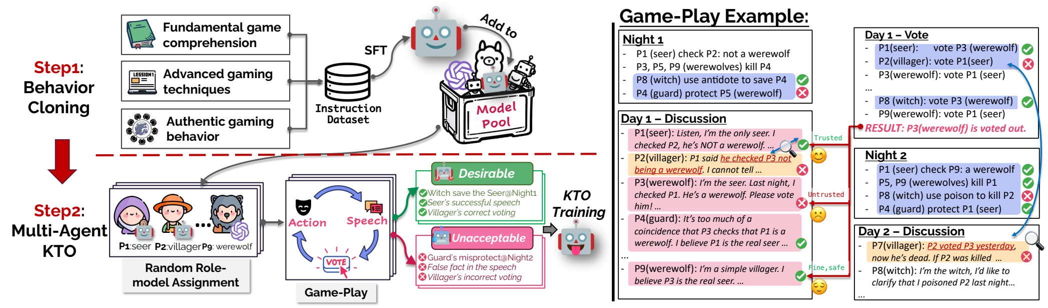


Figure 2: The overall training process consists of **(1) behavior cloning** using instruction data (§2.2) and **(2) multi-agent KTO**(§2.3). In multi-agent gameplay, we randomly assign roles to agents to create diverse interactions that optimize the target model. A stepwise selection process (right) identifies quality preferences using heuristic, staged voting, and verifier-based methods for KTO optimization.

2.1 Expert Data Collection

Despite various meticulously designed prompting methods [18, 27, 28, 29], LLMs still exhibit a huge difference in language style and strategy play from real Werewolf players. While Automatic Speech Recognition (ASR) systems enable the collection of textual data from online Werewolf games [21], they cannot capture players' underlying reasoning. To address these issues, we collaborate with 17 experienced Werewolf players, including individuals with over a thousand games of experience and competitive tournament participants, to annotate the dataset. We also ask these expert players to document their thought behind each decision during the game. Our dataset consists of:

Gameplay record: Contains the nighttime action records of special-role villagers and werewolves, daytime speeches and votes of players, and post-game reviews.

Thinking process annotation: Documents players' reasoning for:
- **Action**: the rationale behind night actions (e.g., seer's checks, werewolf's kills)
- **Speech**: outline of the speech, the identity predictions for other players, and the call for the vote.
- **Voting**: detailed reasons for voting and player identity predictions; Players are also required to distill the day's events into a consolidated record, create notes, and formulate a rudimentary strategy for the next game phase.

We collected 331 annotated *Werewolf* games from 17 expert players via our platform, with 51 additional games reserved for LLM evaluation in Sec. 3.3. Detailed statistics are in Appendix B.

2.2 Behavior Cloning

Due to the scarcity of high-quality data in *Werewolf* domain, existing LLMs generally lack a profound understanding and do not possess sufficient reasoning logic to support advanced gameplay. We address this by creating a comprehensive, multi-level instruction dataset for supervised fine-tuning (SFT), as shown in Fig. 2-Step 1. Our instruction data is derived from three sources: **(1) Fundamental game comprehension**: the explanation of game terminology and jargon. As players continuously innovate within the game, specialized shorthand terms for efficient communication have emerged. For example, the term "Goldwater" refers to players verified as innocent through Seer's investigation. **(2) Advanced gaming techniques**: text collected from experienced players' online strategies, providing guidance for common game scenarios. For instance, it includes expert tips on how werewolves can effectively impersonate the Seer role and mislead the villagers. **(3) Authentic gaming behavior**: Derived from expert-annotated gameplay data in Section 2.1. Benefiting from the annotated thinking process, we structure them into a "think-before-respond" format, enabling the model to truly comprehend the logic behind each stage of the game. For action, we first output the reason and then the target object. For speech, we output expected labels for others and voting intentions as the outline before generation. For voting, we output a summary of the day's events, followed by the chain-of-thought and the voting target. See Appendix C for the examples of the data format. Additionally, we designed a role prediction auxiliary task, which involves predicting each player's role at the end of each day based on known speech, voting, and elimination information.

2.3 Multi-agent KTO

Although the SFT training enhances the model's comprehensive understanding of the Werewolf game, it faces two major challenges: *the characteristics of the game* and *limitations in human-annotated data*. In Werewolf, individual actions (like Seer's claims or Witch's poisoning) subtly influence outcomes, yet team victory doesn't ensure optimal individual play. Additionally, even expert data includes both good and poor decisions from winners and losers, making it challenging to evaluate individual choices based solely on game outcomes. To address these challenges, we develop **Multi-agent KTO (MaKTO)** to mitigate such suboptimality. MaKTO features three key aspects: *1)* It employs the Kahneman-Tversky Optimization (KTO) algorithm [30] for decision refinement. *2)* It adopts multi-agent gameplay to get diverse training data. *3)* Instead of optimizing the *entire* trajectory based on win/loss outcomes, it optimizes step-wise policies.

Kahneman-Tversky Optimization We argue that KTO is particularly suitable for such a multi-agent language game for two reasons. 1) Multi-agent environments are more complex than single/two-agent scenarios, where single-agent interactions with the environment often yield clear feedback, while interactions between agents can have countless possibilities. Moreover, multi-agent dialogues have huge action spaces, leading to sparse trajectory sampling. This makes online reinforcement learning algorithms slow to converge and hard to train. KTO, similar to offline RL, offers a viable solution. 2) Unlike preference optimization algorithms such as DPO [31] and its variants [32, 33, 34, 35, 36], it is nontrivial to get "`prompt-chosen-reject`" paired data. However, you can determine whether an output is acceptable or not through game rules and feedback from other agents. KTO, not requiring paired data for training, makes it ideal for such *"try once"* scenario in multi-agent preference optimization. The loss function of KTO is in Appendix D.

Multi-agent Gameplay We find that in multi-agent settings, merely SFT or *self-play* can lead to rigid strategies and poor generalization (analysis in Sec. 3.6). For example, we find that although the model performs well in scenarios with *two* players claiming to be the seer, its performance significantly deteriorates when only *one or more than three* players claim this role. We argue that in multi-agent environments, the diversity of peers and opponents is crucial. Therefore, instead of relying on self-play, we employ a multi-agent play using a diverse model pool — including various SFT models (from `Llama3.1`, `Qwen2.5`), off-the-shelf LLMs (`GPT-4o`, and `Claude-3.5`), and the agent that uses RL for decision-making and LLM for speech generation. We then randomly select models from the model pool and assign them to different roles in the game. The multi-agent interactions allow for the exploration of a broader strategy space and avoid overfitting to specific policies.

Stepwise Preference Data Selection Defining all actions in an agent's trajectory as desirable or not based solely on the faction's win/loss result is too simplistic. Fortunately, Werewolf's alternating day-night gameplay allows for a more nuanced selection of the desirable and unacceptable step-wise process policies. Specifically, we employ three methods (detailed criteria see Appendix D)

- **Heuristic-based selection** identifies actions based on role-specific strategies and game rules, focusing primarily on nighttime actions and voting phases. For example, werewolves targeting special roles is desirable while not attacking is unacceptable, and for witches, successfully poisoning werewolves is desirable. For voting, unified voting of the villagers against werewolves is preferred, while infighting or vote-splitting, thus weakening the villager team's position, is discouraged.
- **Staged voting-based selection** uses voting outcomes to assess speech quality. Intuitively, players who voted out likely gave suboptimal speeches, either failing to defend themselves or contradicting others' observations. Special role players (seer, witch) face stricter evaluation, with receiving majority villager votes considered unacceptable.
- **Verifier-based selection** employs strong external LLMs, like GPT-4o to verify speech consistency with game facts and events. This can reduce the hallucination by performing fact consistency checks, as well as logical coherence and self-contradictory.

3 Experiments

In this section, we first evaluate our approach in 9-player Seer-Witch-Guard Werewolf games through tournaments, human-AI competitions, and Turing-style detectability tests. We then analyze MaKTO's superior performance through behavioral studies, test its generalization capability with a new *Hunter* role in a new game setting, and validate key components through ablation studies. The implementation details are in Appendix E.

Table 1: Average win rates of the models in the Seer-Witch-Guard setting. A win rate **above 0.5 and in bold** indicates that the model in the row *significantly* outperforms the model in the column (Chi-square test, *p-value<0.05*).

	GPT4o_mini	GPT4o	Claude	Mix	SFT-14b	SFT-72b	MaKTO-72b	Avg.
GPT4o_mini	0.50	0.44	0.23	0.13	0.23	0.24	0.12	0.270
GPT4o	0.56	0.50	0.66	0.56	0.44	0.40	0.35	0.496
Claude	**0.77**	0.34	0.50	0.46	0.48	0.44	0.38	0.481
Mix [21]	**0.87**	0.44	0.54	0.50	**0.58**	0.45	0.47	0.550
SFT-14b	**0.77**	0.56	0.52	0.42	0.50	**0.57**	0.49	0.547
SFT-72b	**0.76**	**0.60**	0.56	0.55	0.43	0.50	0.42	0.546
MaKTO-72b	**0.88**	**0.65**	**0.62**	0.53	0.51	**0.58**	0.50	**0.610**

3.1 Tournament Win Rate Evaluation

3.1.1 Inter-agent Tournament

Experimental Setup We evaluate our approach against several strong baselines: API-based LLM agents (`GPT-4o`, `GPT-4o-mini`, and `Claude-3.5-Sonnet`[4] using chain-of-thought prompting [26]), the `Mix` agent combining LLM with RL policy for decision making following [21], and SFT models (based on `Qwen2.5-14b/72b-instruct`) trained on our expert-annotated dataset.

the Mix agent. Directly optimizing LLMs with online RL in complex multi-party dialogue games is computationally prohibitive, particularly as individual desirable actions do not guarantee victory. Therefore, we elaborate on the `Mix` **agent** here, as it provides a strong online RL benchmark. This agent adopts a Cicero-like approach [37, 21] that decouples decision-making from expression. Specifically, it employs a policy network trained with an AlphaStar-like multi-agent reinforcement learning (MARL) approach [38]. The policy network is trained through extensive self-play, guided by a reward function that combines both the final game outcome (win or loss) and the stage outcome (survived or eliminated). This policy model operates on structured inputs converted from game actions and dialogue, while the LLM (e.g. GPT-4o) serves as a separate "expressor" to articulate the policy's decisions. This agent therefore serves as a strong point of comparison for our method.

Head-to-head Competition In head-to-head competitions, where one model controls the entire villager team (6 agents) and another controls the werewolf team (3 agents), MaKTO-72b achieved a 61% average win rate across 100 games (50 games per faction), significantly outperforming all baselines (Table 1). Notably, while the Mix agent showed strong performance as villagers, it struggled as werewolves due to overly aggressive strategies and policy contradictions (Fig. 3). When comparing SFT models of different sizes (14B and 72B), we observed similar win rates but significantly fewer factual hallucinations in the 72B model's generated speeches.

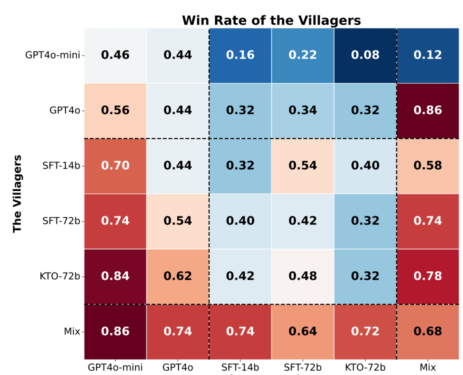
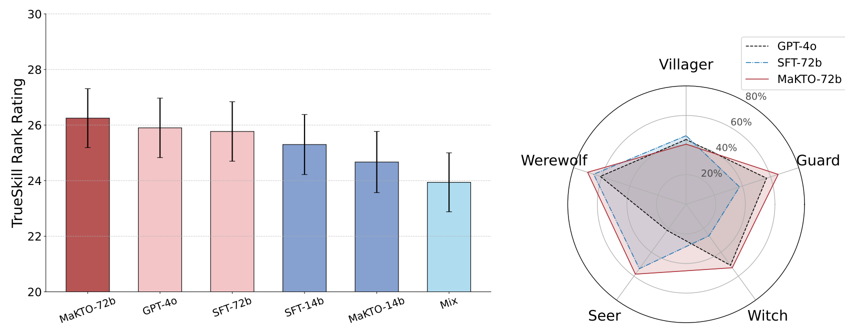


Figure 3: Villager win rate matrix of the head-to-head competition: villager (y-axis) vs. werewolf (x-axis). **Lower left**: higher values for better villagers; **Upper right**: lower values for better werewolves.

Random Competition In random competitions with diverse role assignments across 260 games, MaKTO-72b achieved the highest TrueSkill rating [39] (Fig. 4). This format better reflects the model's adaptability across different roles and team compositions. Particularly, MaKTO-72b significantly outperformed `GPT-4o` when playing as the Seer, suggesting more persuasive statements and better trust-building capabilities. Compared to SFT-72b, MaKTO-72b also exhibited higher winning rates when playing as Guard and Witch, demonstrating better strategic skill usage.

[4] Model Versions: **GPT4o-mini**: gpt-4o-mini-2024-07-18, **GPT4o**: gpt-4o-2024-08-06, **Claude-3.5-Sonnet**: claude-3-5-sonnet-20241022.



(a) TrueSkill Score of models (b) Role Winning rates

Figure 4: Results of 260 random competitions in 9-player Seer-Witch-Guard game setting.

3.1.2 Human-AI Tournament

Experimental Setup To evaluate our model's performance in real-world games, we conduct human-AI tests with 14 experienced human players (1000+ games each) through head-to-head and random competition. Unlike previous studies [20, 21] that only introduced single AI or human players, we involve multiple human players and AI models, creating more challenging and realistic environments.

Head-to-head Competition We evaluate MaKTO in 20 head-to-head Seer-Witch-Guard competition (10 games for each side), where MaKTO-72b plays all villagers or werewolves independently, and the opposing side are played by humans. MaKTO-72b achieved a 60% win rate (5/10 as villagers, 7/10 as werewolves; Table 2), indicating that it is on par with high-level human players.

Table 2: Win rate of MaKTO-72b in head-to-head competition with humans in 9-player Seer-Witch-Guard setting.

		Humans	
		Villager	Werewolf
MaKTO-72b	Villager	–	0.5
	Werewolf	0.7	–

Random Competition We also evaluate in random competition for 30 seer-witch-guard 9-player games, where each player is randomly selected from 2-7 human players and the model pool (including `GPT-4o`, Mix agent, SFT-72b, and MaKTO-72b). As in Fig. 5, MaKTO-72b achieved a win rate 61.8%±8.3% win rate, ranking fourth among all players and surpassing the average human win rate of 54%. This shows that MaKTO-72b has strong adaptability against both human and AI opponents.

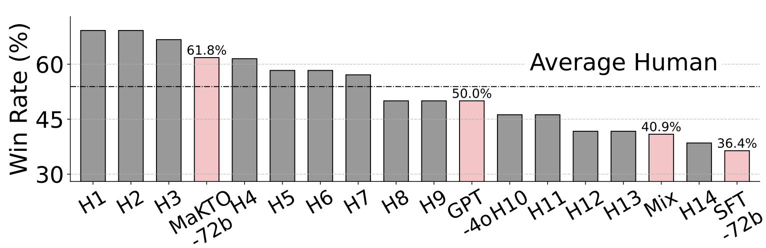


Figure 5: Win rate of players in random competition. H1-14 stand for the win rate of human players.

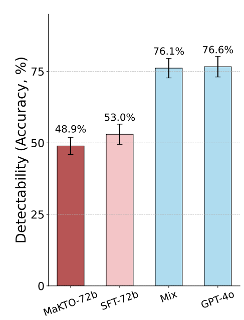


Figure 6: Human detection accuracy.

3.2 Turing-style Detectability Test

We conduct rigorous Turing-style blind detectability tests in both competitions. We require each human player to explicitly judge whether *every other* participant is human or AI, without any prior knowledge of AI presence. This mandatory assessment provides a rigorous human similarity assessment. MaKTO achieves detection accuracy of only 48.9% (Fig. 6), lower than random chance, indicating that our model successfully passes this specialized Turing test by convincingly emulating human-like gameplay characteristics and social behaviors. On the contrary, `GPT-4o` has a much higher detection rate (76.6%) due to the significant differences in speaking style and voting behavior from human players.

3.3 Behavioral Analysis

3.3.1 Comparison with Baseline Model

In order to understand why MaKTO has a higher win rate in tournaments, we analyze the behaviors and decisions generated during the tournament. For a fair comparison, we selected `GPT-4o` as the opponent. We run 50 games between both sides and compute the proportion of behavior occurrences as metrics. Results are shown in Fig. 3, with detailed metrics provided in Appendix F.

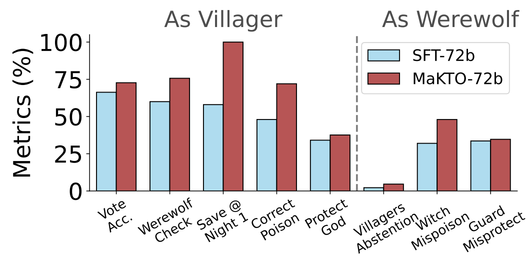


Table 3: Behavioral Analysis: model performance as villagers (left) and werewolves (right) against GPT-4o. For werewolf, **lower** opponent scores indicate better performance of the model.

When the trained models played as villagers against `GPT-4o` werewolves, we evaluate voting accuracy (*Vote Acc.*). For special roles, we examine: Seer's werewolf identification accuracy (*Werewolf Check*), Witch's first-night rescue rate (*Save @ Night 1*), werewolf poisoning accuracy (*Correct Poison*), and Guard's special role protection rate (*Protect God*). Fig. 3 (left) shows that MaKTO outperforms the baseline SFT model across all the metrics, which yields a higher villager win rate. These improvements can be attributed to the stepwise decision rewards and penalties in the MaKTO training.

When our models played as werewolves against GPT-4o villagers, we evaluated the opponent's behavior instead of our model's direct performance. The rationale behind this approach is that better werewolf deception leads to more opponents' confusion and mistakes. Specifically, we measured the following metrics of the `GPT-4o` opponents: *1)* Voting abstention rate (*Abstention*): Higher rates indicate difficulty identifying werewolves. *2)* Seer's werewolf identification rate (*Seer Check*, see Table 15 for specific values): Lower rates suggest successful misdirection. *3)* Witch's villager poisoning rate (*Mispoison*): Higher rates indicate better werewolf concealment. *4)* Guard's werewolf protection rate (*Misprotect*): Higher rates suggest effective deception. All the fine-grained metrics of MaKTO are better than the SFT model, which shows that MaKTO werewolf has a superior camouflage capability in the game.

3.3.2 Comparison with Human

A feature of the *Werewolf* game lies in its prevalence of deception, particularly among the werewolf players. Werewolves never openly admit their identity in daytime discussions; instead, they make up various identities and stories to protect themselves. It becomes important for villagers to correctly predict who is the werewolf, so that they can cast the right vote. So in this experiment, we compare the correct judgments of the villagers and compare them with humans.

Table 4: Offline results of human annotators and LLMs. The **bold** number represents the best results of the models, and the underlined number represents the second best.

	Vote Acc.(↑)	Abstent Rate(↓)	Align. Acc.(↑)	Wolf-pred. F1(↑)
GPT4o_mini	67.2%	**0.4%**	68.1%	0.519
GPT4o	69.4%	2.3%	68.0%	0.587
Claude	68.4%	1.0%	75.2%	0.651
Qwen2.5-14b	61.0%	4.1%	61.1%	0.528
Qwen2.5-72b	66.5%	0.4%	63.9%	0.552
SFT-14b	70.8%	4.1%	77.9%	0.712
SFT-72b	71.1%	5.8%	**79.3%**	**0.734**
MaKTO-72b	**73.8%**	1.5%	78.4%	**0.734**
HUMAN	76.7%	4.8%	76.1%	0.742

We separate 51 matches from the annotated data as the test set, excluding them from the training dataset. The evaluation covers 484 voting events and 5130 identity predictions. The results are presented in Table 4. In terms of voting, we evaluate the voting accuracy, that is, the accuracy of gods and villagers voting for werewolves; and the abstention rate. MaKTO-72b achieves the highest voting accuracy. In terms of identity prediction, we evaluate the accuracy of side alignment (*Align. Acc.*), that is, correctly predicting gods and villagers as the good identity and werewolves as the bad identity; and the F1-score in predicting werewolves (*Wolf-pref. F1*). The trained models show significant improvement over the base model like `Qwen2.5-14b` and -72b-instruct, which demonstrate the effectiveness of the expert data we collected. Also, we notice that, the SFT models and MaKTO even achieve higher side alignment accuracies than human.

3.4 Generalizing to Other Game Setting

Another advantage of our model lies in its cross-game generalization capability, which we test by varying both the game's role and its scale.

Adaptability to New Role in the Game. First, we assess MaKTO's ability to adapt to a rule change by introducing a new role - the *Hunter*—in place of the Guard. The Hunter can choose to eliminate another player upon their own elimination. We conduct tournament experiments in the Seer-Witch-Hunter setup. In this new game setup, where Hunter replaces Guard, the policy model of Mix agent (trained on Seer-With-Guard setup) no longer remains effective, whereas MaKTO continues to perform exceptionally well (Table 5). Despite being trained only on Seer-Witch-Guard gameplay data, MaKTO still outperforms SFT models, demonstrating strong adaptability and generalization capabilities.

Table 5: Average win rates of the models in the 9-player Seer-Witch-**Hunter** setting. A win rate **above 0.5 and in bold** indicates that the model in the row *significantly* outperforms the model in the column (Chi-square test, *p-value<0.05*).

	GPT4o_mini	GPT4o	Claude	SFT-14b	SFT-72b	MaKTO-14b	MaKTO-72b	Avg.
GPT4o_mini	0.50	0.48	0.15	0.33	0.37	0.26	0.29	0.340
GPT4o	0.52	0.50	**0.72**	**0.62**	0.50	0.52	0.54	0.560
Claude	**0.85**	0.28	0.50	**0.62**	0.53	0.46	0.37	0.516
SFT-14b	**0.67**	0.38	0.38	0.50	0.53	0.48	0.51	0.493
SFT-72b	**0.63**	0.50	0.47	0.47	0.50	0.51	0.48	0.507
MaKTO-14b	**0.74**	0.48	0.54	0.52	0.49	0.50	0.46	0.533
MaKTO-72b	**0.71**	0.46	**0.63**	**0.58**	0.52	0.54	0.50	**0.563**

Generalization to Larger Game Scales. To further probe its generalization, we evaluated MaKTO in even more complex scenarios with a larger number of players: a 10-player game (with either a Guard or a Hunter) and a 12-player game. These settings introduce different strategic dynamics and were entirely absent from the training data. The results in Table 6 show that MaKTO-72b consistently achieves the highest average win rate across all three larger-scale settings. Taken together, these results demonstrate that MaKTO learns robust, transferable strategic reasoning, rather than merely overfitting to its training configuration.

Table 6: Average win rates of models in unseen 10- and 12-player settings. MaKTO consistently tongton baselines. See Appendix G for detailed head-to-head results.

Models	10-Player w/ Guard	10-Player w/ Hunter	12-Player
GPT-4o	0.335	0.320	0.423
Claude	0.465	0.573	0.488
SFT-72b	0.575	0.523	0.485
MaKTO-72b	**0.625**	**0.585**	**0.605**

3.5 Comparison between MaKTO and Long-COT Reasoning Model

An interesting question is how our MaKTO compares to long-COT reasoning models [40, 41] in the Werewolf game. We conducted win rate comparisons between reasoning models and MaKTO models of equivalent size in the Werewolf testing environment. Results (Tab. 7) show that MaKTO-14B achieves a 74% overall win rate (70% as werewolves, 78% as villagers) against DeepSeek-R1-Distill-Qwen-14B [41], and comparable-sized MaKTO-72B achieves an 84% win rate against R1-distilled-70B. RL LLMs trained on math and coding tasks fail to generalize to the social deduction games, like Werewolf. While R1 models make conservative statements (e.g., a Seer might say "I found a suspicious player but cannot reveal who"), making it vulnerable to attacks and losing trust from villagers, MaKTO models develop more sophisticated social strategies through diverse interactions. For example, when a werewolf's false claim draws suspicion, the werewolf team might strategically eliminate this werewolf to cast doubt on the real Seer in subsequent rounds. Note that while we applied "think-before-respond," the tokens in the thinking are fewer than those in the reasoning models. We believe that long-COT and our MaKTO are not mutually exclusive, but this paper focuses on the training method, with their integration left for future work.

3.6 Ablation Studies

In the ablation studies, we reveal two crucial design in MaKTO — multi-agent gameplay and step-wise preference data selection. Note that all experiments use the 14B model.

Table 7: Win rate for MaKTO and DeepSeek-R1-Distill models in 9-player Seer-Witch-Guard game.

Wining rate of 14B models

	GPT-4o	R1-Distill-Qwen-14b	MaKTO-14b	Avg. Win Rate
GPT-4o	0.50	**0.89**	0.37	0.630
R1-Distill-Qwen-14b	0.11	0.50	0.26	0.185
MaKTO-14b	**0.63**	**0.74**	0.50	**0.685**

Wining rate of 70B/72B models

	GPT-4o	R1-Distill-Llama-70b	MaKTO-72b	Avg. Win Rate
GPT-4o	0.50	**0.79**	0.35	0.570
R1-Distill-Llama-70b	0.21	0.50	0.16	0.185
MaKTO-72b	**0.65**	**0.84**	0.50	**0.745**

Q1: Multi-agent Play or Self-play? MaKTO requires playing with ***various*** agents to select preference data for training. We compare MaKTO with Self-play KTO, where an SFT-trained model engages in numerous battles only against itself to collect preference data for training. To ensure fairness, both methods used 20k training samples. From Table 8, multi-agent play significantly outperformed self-play KTO, achieving a 4% higher average win rate. While self-play KTO showed competitive performance against SFT models, it significantly underperformed against diverse opponents like `GPT-4o`. This demonstrates that exposure to diverse opponents is crucial for developing robust strategies, similar to findings in population-based training [42, 20]. Additionally, we find that preference data can be also obtained from annotated data without engaging in gameplay (w/o gp). But this results in a substantial 13% decrease in average win rate, highlighting the importance of interactive gameplay in our approach.

Table 8: Ablation study for MaKTO and other KTO training methods. **Self-play**: preference data from SFT-14B vs. SFT-14B games. **w/o gp**: preference data from annotations only without actual gameplay. All models are in 14B size.

	GPT-4o	Mix	SFT	MaKTO	Self-play	w/o gp	Avg. Win
GPT-4o	0.50	**0.56**	0.44	0.37	**0.56**	**0.64**	0.514
Mix	0.44	0.50	**0.58**	0.55	**0.71**	0.48	0.553
SFT	**0.56**	0.42	0.50	0.48	0.38	0.66	0.499
KTO methods							
MaKTO	**0.63**	0.45	**0.52**	0.50	0.50	**0.57**	0.534
Self-play	0.44	0.29	**0.62**	0.50	0.50	**0.56**	0.483
w/o gp	0.36	0.31	0.34	0.43	0.44	0.50	0.374

Q2: Step-wise Selection or Selection based solely on results? MaKTO uses stepwise action selection, while an alternative can simply collect preference trajectory data from win/loss outcomes. Which is better? Table 9 shows the comparison of win rates — stepwise preference selection proves superior to trajectory-based selection. This means a final reward does not necessarily indicate that all actions in that trajectory are desirable, and vice versa. Selecting data based on trajectory outcomes leads the model to learn suboptimal actions and wrongly penalize desirable ones, resulting in a lower winning rate. This also confirms our hypothesis that game outcomes alone cannot accurately reflect the quality of individual decisions in complex SDGs.

Table 9: Ablation study for MaKTO and other KTO training methods. **Step** (Ours): Selecting desirable/unacceptable actions according to predefined criteria. **Traj.**: Selecting desirable actions from winning trajectories and unacceptable actions from losing trajectories. All models are in 14B.

	GPT-4o	Mix	SFT	MaKTO Step	MaKTO Traj.	Avg. Win
GPT-4o	0.50	**0.56**	0.44	0.37	0.45	0.456
Mix	0.44	0.50	**0.58**	**0.55**	**0.63**	0.550
SFT	**0.56**	0.42	0.50	0.48	**0.54**	0.496
***MaKTO** with different action selections*						
Step(Ours)	**0.63**	0.45	**0.52**	0.50	**0.58**	0.545
Traj.	**0.55**	0.37	0.47	0.44	0.50	0.458

Q3: Contribution of Each Training Stage. To quantify the contribution of each component of our pipeline, we compare the base instruction-tuned model (Qwen2.5-Instruct-72b), the model after behavior cloning (SFT-72b), and the final model after KTO refinement (MaKTO-72b). Table 10 shows a clear stepwise improvement. Behavior cloning on our expert dataset provides a massive performance boost of 30.6% over the

Table 10: Component analysis: head-to-head win rate of the base model (Qwen2.5-Instruct-72b), the SFT model, and the final MaKTO model against each other. Models are in 72B.

	Base	SFT	MaKTO	Avg.
Base	0.50	0.20	0.10	0.267
SFT	0.80	0.50	0.42	**0.573 (+30.6%)**
MaKTO	0.90	0.58	0.50	**0.660 (+8.7%)**

base model. Subsequently, MaKTO refinement adds another 8.7% improvement, demonstrating the distinct value of both high-quality expert data and interaction-based preference optimization.

4 Related Work

SDG as Testbed Social Deduction Games (SDGs) have emerged as useful benchmarks for social reasoning. They uniquely combine fundamental linguistic skills, higher-level social reasoning abilities, and adaptability. Recent research has extensively explored various SDGs [28, 13, 14, 15, 43, 44, 17, 16, 45], with *Werewolf* [18, 20, 19, 22] becoming a popular testbed for evaluating LLMs' reasoning through its complex dynamics of deception and cooperation.

LLM-based Game Agents While early game AI relied on reinforcement learning (RL) in environments with no or only a little dialogue [46, 47, 48, 49, 37], LLMs enabled more sophisticated agent modeling in SDGs, through generative planning [28], memory mechanisms [50], experience learning [18, 27], and persona prompting [29]. Existing approaches for *Werewolf* AI typically adopt two-stage frameworks: either RL-then-LLM [21] or LLM-then-RL [20, 51], which either compress the language strategy space or are constrained by generated candidates. We propose integrated training for LLM through direct interactions. Very recently, Sarkar et al. [52] introduced MARL for AmongUs with a smaller RWKV-1.5B [53] model, Other research has explored fine-tuning LLMs for different strategic games like Diplomacy [54]. But these often in settings with more structured or limited communication, while our method efficiently leverages LLM while maintaining stable training.

Agent Learning Our approach is closely related to agent learning research. Current methods generally follow 1) imitation learning with expert trajectory data [55, 56, 57, 58] or 2) learning through environmental interactions [36, 59, 60, 61]. While some have also explored the gaming environments [60, 62, 63], they primarily focus on simpler single/two-agent scenarios. We tackle the complexity of multi-agent language game environments, requiring better adaptability and robustness.

5 Conclusion

In this paper, we present Multi-agent KTO (MaKTO), a novel approach for optimizing LLMs in complex social deduction games. It improves LLM's social reasoning and strategic interactions through interaction-based feedback. MaKTO beats GPT-4o with 23.0% higher win rates and wins 60% against expert humans, while maintaining human-like conversations. We also contribute a large-scale expert Werewolf dataset with player actions and their reasoning processes.

Limitations

We identify several limitations. First, for fair comparison, we focus on the rationality of models' decisions without customizing the different role-playing personas in the system prompt of LLMs. Second, our current implementation relies on turn-based conversations rather than free-form interactions [50, 64, 65, 66]. The challenge of modeling unrestricted multi-agent communications, where agents can interact more naturally and flexibly, remains an important area for future research. Third, similar to the general limitations of LLMs, our model occasionally exhibits inconsistent behavior and hallucinations across long conversations, suggesting room for improvement in long-text modeling capabilities, particularly in maintaining coherence during extended social interactions. Finally, while Multi-agent KTO provides an easy yet effective training paradigm, we believe that online multi-agent RL could potentially better, presenting another promising direction for future research.

Acknowledgments

We would like to thank Zhaohao Xiao, Weicheng Fu, Qiang Lu, Danyu Lv, Hangfei Guo, Zheng Li, and Huiying Lin for their valuable support in the development of the annotation system. We are also grateful to Xiaoxi Mao, Jinfeng Wang, and Zechuan Hu for their insightful discussions.

References

[1] Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

[2] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

[3] Paul M Salmon, Chris Baber, Catherine Burns, Tony Carden, Nancy Cooke, Missy Cummings, Peter Hancock, Scott McLean, Gemma JM Read, and Neville A Stanton. Managing the risks of artificial general intelligence: A human factors and ergonomics perspective. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 33(5):366–378, 2023.

[4] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

[5] Ludwig Wittgenstein. *Philosophical Investigations*. Basil Blackwell, Oxford, 1953. ISBN 0631119000.

[6] Roman Kopytko. Philosophy and pragmatics: A language-game with ludwig wittgenstein. *Journal of Pragmatics*, 39(5):792–812, 2007.

[7] Rom Harre. Wittgenstein and artificial intelligence. *Philosophical Psychology*, 1(1):105–115, 1988.

[8] Sida I Wang, Percy Liang, and Christopher D Manning. Learning language games through interaction. In *In Proc. of ACL*, pages 2368–2378. Association for Computational Linguistics (ACL), 2016.

[9] Tom Schaul. Boundless socratic learning with language games. In *NeurIPS 2024 Language Gamification Workshop*, 2024.

[10] Christine Cuskley, Rebecca Woods, and Molly Flaherty. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083, 2024.

[11] Ying Wen, Ziyu Wan, and Shao Zhang. Language games as the pathway to artificial superhuman intelligence. *arXiv preprint arXiv:2501.18924*, 2025.

[12] David Schlangen. Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy. *arXiv preprint arXiv:2304.07007*, 2023.

[13] Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. In *Findings of ACL*, pages 8225–8291, 2024.

[14] Alessandro Trevisan, Harry Giddens, Sarah Dillon, and Alan F Blackwell. Measuring bullshit in the language games played by chatgpt. *arXiv preprint arXiv:2411.15129*, 2024.

[15] Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

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

[17] Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*, 2024.

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

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

[20] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

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

[23] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, 2018.

[24] Tadahiro Taniguchi, Ryo Ueda, Tomoaki Nakamura, Masahiro Suzuki, and Akira Taniguchi. Generative emergent communication: Large language model is a collective world model. *arXiv preprint arXiv:2501.00226*, 2024.

[25] Jinyu Cai, Jialong Li, Mingyue Zhang, Munan Li, Chen-Shu Wang, and Kenji Tei. Language evolution for evading social media regulation via llm-based multi-agent simulation. *IEEE World Congress on Computational Intelligence (WCCI) - IEEE Congress on Evolutionary Computation (IEEE CEC),*, pages 1–10, 6 2024. doi: 10.1109/CEC60901.2024.10612015.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[27] Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. *arXiv preprint arXiv:2310.14985*, 2023.

[28] Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*, 2023.

[29] Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. An implementation of werewolf agent that does not truly trust llms. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 58–67, 2024.

[30] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *In proc. of ICML*, 2024.

[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, 2024.

[33] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

[34] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

[35] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

[36] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of llm agents. In *In Proc. of ACL*, pages 7584–7600, 2024.

[37] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

[38] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[39] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.

[40] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[41] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[42] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

[43] Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. Two giraffes in a dirt field: Using game play to investigate situation modelling in large multimodal models. *arXiv preprint arXiv:2406.14035*, 2024.

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

[45] Satvik Golechha and Adrià Garriga-Alonso. Among us: A sandbox for agentic deception. *arXiv preprint arXiv:2504.04072*, 2025.

[46] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[47] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[48] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

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

[50] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[51] Zelai Xu, Wanjun Gu, Chao Yu, Yi Wu, and Yu Wang. Learning strategic language agents in the werewolf game with iterative latent space policy optimization. In *In Proc. of ICML*, 2025.

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

[53] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of EMNLP*, pages 14048–14077, 2023.

[54] Xu Kaixuan, Chai Jiajun, Li Sicheng, Fu Yuqian, Zhu Yuanheng, and Zhao Dongbin. Dipllm: Fine-tuning llm for strategic decision-making in diplomacy. In *In Proc. of ICML*, 2025.

[55] Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. Agenttuning: Enabling generalized agent abilities for llms. In *Findings of ACL*, pages 3053–3077, 2024.

[56] Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, et al. Agentgym: Evolving large language model-based agents across diverse environments. pages 27914–27961, 2025.

[57] Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-flan: Designing data and methods of effective agent tuning for large language models. In *Findings of ACL*, pages 9354–9366, 2024.

[58] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *In Proc. of AAAI*, pages 19632–19642. AAAI Press, 2024. doi: 10.1609/aaai.v38i17.29936. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29936`.

[59] Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. Watch every step! llm agent learning via iterative step-level process refinement. In *In Proc. of EMNLP*, pages 1556–1572, 2024.

[60] Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. Self-playing adversarial language game enhances llm reasoning. In *Advances in Neural Information Processing Systems*, 2024.

[61] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

[62] Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization. In *In Proc. of ACL*, 2024.

[63] Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In *In Proc. of ACL*, pages 12912–12940, 2024.

[64] Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, et al. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*, 2024.

[65] Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*, 2024.

[66] Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Maarten Sap, and Michael R Lyu. On the resilience of multi-agent systems with malicious agents. *arXiv preprint arXiv:2408.00989*, 2024.

[67] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[68] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*, 2024.

[69] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. In *The Conference on Empirical Methods in Natural Language Processing*, 2023.

A Game Rules

Werewolf [5] is one of the most popular social detection games, typically played with 7 to 15 players. The game is set in a village where some players are secretly assigned the role of werewolves, while the majority are villagers. In this paper, we focus primarily on the variant with **9 players**: 3 werewolves and 6 villagers, including special roles of a Seer, a Witch, and a Guard (Figure 7). The gameplay consists of two alternating phases: night and day. During the night, the werewolves secretly choose a victim, while the Seer checks one player's identity, the Witch can use a one-time antidote or poison, and the Guard protects one player from being killed by the werewolves. In the day phase, all surviving players discuss and vote to eliminate a suspected werewolf. The game continues until either all werewolves are eliminated (village wins) or all simple villagers or all special-role villagers are eliminated (werewolves win). Another variant of the 9-player game introduces the Hunter in place of the Guard. The Hunter's skill allows them to shoot and eliminate one player when they are eliminated.

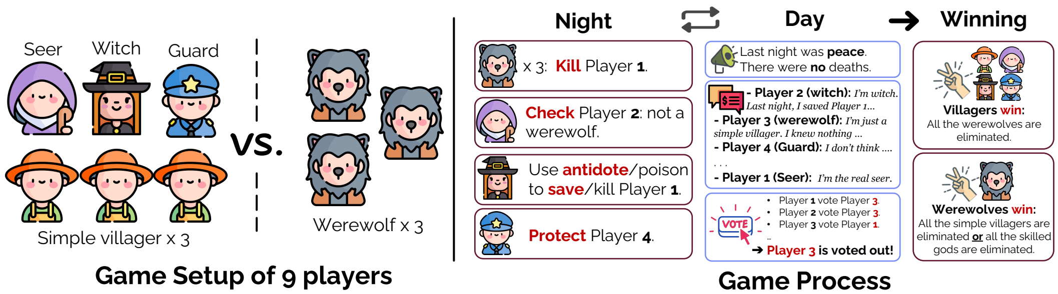


Figure 7: The setup and gameplay of the 9-player Werewolf game with special roles of the Seer, the Witch, the Guard.

Here are the details of the specific game rules:

A.1 Game Objectives

In this game, players are usually divided into two camps: werewolves and villagers. Depending on their roles, players have distinctive objectives:

- **Villagers** aim to identify the werewolves and eliminate them through voting. Within the villagers' camp, there are some special roles with distinctive abilities that can help the villagers secure victory.
- **Werewolves'** primary objective is to conceal their true identities, mislead others in discussions to avoid being voted out, and hunt villagers as covertly as possible.

A.2 Game Process

The game generally includes the following basic procedures.

- **Role Assignment**: Upon entering the game, player roles are secretly assigned. Werewolves know each other's identities, while villagers only know their own role.
- **Day-Night Alternation**: The game alternates between day and night phases. At night, werewolves secretly choose a villager to eliminate; some special roles can also activate their abilities at night. During the day, all players discuss and vote to eliminate the player they believe to be a werewolf, with the player receiving the most votes being eliminated.
- **Victory Conditions**: The game ends when one faction achieves its victory conditions. Villagers win if all werewolves are eliminated. Werewolves win if they eliminate all ordinary villagers or all special roles.

A.3 Role Descriptions and Different Configurations

Standard configurations for 9-player and 7-player games incorporate six distinct roles: the Seer, the Witch, the Guard, the Hunter, the Werewolf, and the Villager. Different roles have the following abilities.

- **Seer**: During the night phase, the Seer can secretly select a player to learn their true identity (whether they are a werewolf or not).

[5] Also known as *Mafia*, `https://en.wikipedia.org/wiki/Mafia_(party_game)`.

- **Witch**: The Witch has one healing potion and one poison potion, each usable only once. The Witch cannot use both potions in the same night. The healing potion can save a player killed by werewolves at night. The poison potion can eliminate a player suspected of being a werewolf.
- **Guard**: The Guard can protect one player each night from werewolf attacks. The Guard can choose to protect himself or opt not to protect anyone, but cannot protect the same player on consecutive nights.
- **Hunter**: When the Hunter is killed by werewolves or eliminated during the voting event, he can reveal his identity card and shoot a revenge bullet at any living player, causing that player to die as well. The Hunter can choose not to reveal his card, but once revealed, he must take someone with him (Note: If the Hunter is poisoned by the Witch, he cannot reveal his card or take anyone with him).
- **Werewolf**: Werewolves can choose to eliminate a player during the night phase.
- **Villager**: Villagers have no special abilities. They can only distinguish the werewolves through daytime speech and public information.

This paper encompasses four distinct configurations, including the **9-player werewolf game** with three special roles, and the **7-player werewolf game** with two special roles:

- **Seer-Witch-Guard**: Includes one Seer, one Witch, one Guard, three Werewolves, and three Villagers.
- **Seer-Witch-Hunter**: Includes one Seer, one Witch, one Hunter, three Werewolves, and three Villagers.
- **Seer-Guard**: Includes one Seer, one Guard, two Werewolves, and three Villagers.
- **Seer-Witch**: Includes one Seer, one Witch, two Werewolves, and three Villagers.

B Dataset Statistics

17 expert players provide annotations using our self-built annotation platform. Each annotator was paid much more than the average local salary. We collect 331 Werewolf games for training, including 278 9-player games (Seer, Witch, Guard or Hunter) and 53 7-player games (Seer, Guard or Witch). We collect 331 matches of Werewolf games annotated by 17 advanced players based on our self-built annotation platform. This dataset includes 278 9-player games (using the setups of Seer, Witch, Guard and Hunter) and 53 7-player games (using the setups of Seer, Guard and Witch). Each game features randomly involved participants and randomly assigned Werewolf roles to guarantee data diversity. The total duration of the game annotated by the players exceeds 1,000 hours, including 3,759 speech data entries (exceeding 540,000 tokens), 2,698 action events, and 3,875 voting records.

Game Setting	Composition of Roles	#Games	#Speech	#Speech Tokens	#Action	#Vote
9 Player	Seer Witch Guard	144	1,805	254k	1,387	1,864
	Seer Witch Hunter	134	1,532	239k	1,001	1,566
7 Player	Seer Guard	25	203	23k	132	215
	Seer Witch	28	219	30k	178	230
Total	-	331	3,759	545k	2,698	3,875

Table 11: Statistics of the Werewolf game data for training.

C Data Format of Behavior Cloning

In this section, we show examples of the prompt-response data format (in Chinese and translated English[6]). Specific dialogue and gameplay processes may be omitted.

System Prompt

```
你现在正在玩一种叫做''狼人杀''的游戏。
在这款游戏中，玩家通常被分为两个阵营：狼人和村民。
狼人杀游戏中不同角色的玩家有不同的目标：
- 村民的目的是识别出狼人，并通过投票使他们出局。
- 对于狼人来说，他们的主要目标是隐藏他们的真实身份，在讨论中误导他人，以免被投票出局并尽可能的猎杀村民。
以下是一些基本规则：
- 身份：玩家的身份是秘密分配的。狼人彼此知道对方的身份，而村民只知道自己的身份。
- 昼夜更替：游戏有交替的白天和黑夜阶段。夜里，狼人秘密选择一名村民猎杀。白天，所有玩家讨论并投票决定他们认为是狼人的玩家，票数最多的玩家被淘汰。
- 特殊角色：游戏中有存在一些有特殊能力的角色，比如能得知玩家身份的``预言家''等。
- 获胜条件：当游戏中有一个群体实现它们的获胜条件时游戏结束。如果所有狼人被淘汰，村民就获胜。如果狼人杀死了所有普通村民或所有特殊角色，狼人就获胜。
You are now playing a game called 'Werewolf' (also known as 'Mafia').
In this game, players are typically divided into two factions: Werewolves and Villagers.
- Different roles in the Werewolf game have different objectives:
The Villagers' goal is to identify the Werewolves and eliminate them through voting.
- For the Werewolves, their main objective is to hide their true identities, mislead others during discussions to avoid being voted out, and hunt down as many Villagers as possible.
Here are some basic rules:
- Identity: Players' identities are secretly assigned. Werewolves know each other's identities, while Villagers only know their own.
- Day and Night Cycles: The game alternates between day and night phases. At night, Werewolves secretly choose a Villager to eliminate. During the day, all players discuss and vote on who they believe is a Werewolf, and the player with the most votes is eliminated.
- Special Roles: There are some roles with special abilities in the game, such as the 'Seer' who can learn players' identities.
- Winning Conditions: The game ends when one group achieves its winning conditions. If all Werewolves are eliminated, the Villagers win. If the Werewolves kill all ordinary Villagers or all special roles, the Werewolves win.

在这个游戏中，我们有从1到9号共9名玩家：6名村民和3名狼人。村民中有特殊角色，包括：
In this game, we have 9 players numbered from 1 to 9: 6 Villagers and 3 Werewolves. Among the Villagers, there are special roles, including:
- 1位预言家：
- 目标：预言家的目的是帮助村民识别狼人。
- 能力：在夜晚阶段，预言家可以秘密选择一名玩家，每晚了解他的真实身份（是否为狼人）。
1 Seer:
Objective: The Seer's purpose is to help the Villagers identify the Werewolves. Ability: During the night phase, the Seer can secretly choose one player and learn their true identity (whether they are a Werewolf or not) each night.
- 1位女巫：
- 目标：女巫的目的是策略性地使用她的特殊能力来帮助村民。
- 能力：女巫有一瓶解药和一瓶毒药。一旦使用，后续回合中不能再用。女巫不能在同一晚既使用解药又使用毒药。解药可以用来救一名在夜间被狼人猎杀的玩家。毒药可以淘汰一名很可能是狼人的玩家。
```

[6] We use `Claude-3.5-Sonnet-v2.5` for translation and proofread them manually.

```
- 1 Witch:  Objective:  The Witch's purpose is to strategically use her
special abilities to help the Villagers.  Abilities:  The Witch has one
healing potion and one poison potion.  Once used, they cannot be used again
in subsequent rounds.  The Witch cannot use both the healing potion and the
poison potion on the same night.  The healing potion can be used to save a
player who was killed by the Werewolves during the night.  The poison potion
can eliminate a player who is likely to be a Werewolf.
- 1位守卫：
- 目标：守卫的目的是策略性地使用他的特殊能力来帮助村民。
- 能力：守卫每晚可以保护一名玩家，防止他们受到狼人的攻击。守卫可以选择保护自己，
或者选择不保护任何人，但他不能在连续两个夜晚保护同一个玩家。
- 1 Guard:
Objective:  The Guard's purpose is to strategically use his special ability
to help the Villagers.
Ability:  The Guard can protect one player each night from Werewolf attacks.
The Guard can choose to protect himself or choose not to protect anyone, but
he cannot protect the same player for two consecutive nights.
其他的都是普通村民。
The rest are ordinary Villagers.
```

Action

```
Prompt:
在本场游戏中，你目前已知以下信息：
In this game, you currently have the following information:

1. 角色设定：
你是1号玩家。你的身份是：预言家。
你每晚可以查看一名玩家是否为狼人，你的目标是利用这些信息帮助其他人找出并淘汰所有狼人。
1. Character setup:
You are Player 1. Your role is: Seer.
Each night, you can investigate one player to determine if they are a
werewolf. Your goal is to use this information to help other players
identify and eliminate all the werewolves.

2. 客观信息：
- 游戏进程：目前游戏进行到第2轮。
- 当前存活的玩家有：1号，2号，4号，5号，6号，8号，只能在以上玩家中选择进行查验
- 行动记录：第1轮预言家查验5号玩家，5号玩家是狼人。
- 猎人开枪记录：9号猎人在第1轮夜晚被杀害出局，开枪带走了7号玩家。
- 投票情况：
第1轮投票记录：
2号玩家投给：3号玩家；
...
8号玩家投给：3号玩家；
结果：3号玩家被投票出局。
2. Objective information:
Game progress: The game is currently in Round 2.
Currently surviving players are: Player 1, Player 2, Player 4, Player 5,
Player 6, and Player 8. You can only choose to check among these players.
Action record: In Round 1, the Seer checked Player 5, and Player 5 is a
werewolf.
Hunter's shot record: Player 9 (Hunter) was killed and eliminated in the
first night, and shot Player 7 upon death.
Voting situation:
Round 1 voting record:
Player 2 voted for: Player 3
...
Player 8 voted for: Player 3
Result: Player 3 was voted out.

3. 主观信息：
3. Subjective information:
- 第1轮所有玩家发言：
Round 1 speeches from all players:
3号玩家：我是一个村民，夜晚没有任何信息，我要强烈谴责女巫，夜晚为什么不救9号猎人？让我们好人阵营白白损失了强神猎人，我现在希望被9号带走的7号是头狼人，这样我们好人阵营还有获胜的希望！希望后置位的预言家能好好发言，认真分析场上局势，让好人能一目了然地找到你。
Player 3: I am a villager and have no information from the night. I
strongly condemn the Witch - why didn't you save Hunter #9? Our good team
has lost a powerful god role, the Hunter, for nothing. I hope that Player
7, who was taken out by Player 9, is the wolf leader. This way, we villagers
still have a chance to win! I hope the Seer in the later speaking positions
can give a good speech, carefully analyze the situation, and make it clear to
the villagers who you are.
4号玩家：......
Player 4: ...
...
```

2号玩家：我是一张村民牌，前置位没有玩家跟5号对跳女巫，那么在我眼里5号就是真女巫。1号肯定是悍跳预言家的狼人，可恶的1号狼人就是想要把我们尊贵的5号双药女巫抗推出局。今天我会站边4号，把3号狼人放逐出局。
Player 2: I'm a villager. Since no one in the earlier positions challenged Player 5's claim as the Witch, I consider 5 to be the real Witch. Player 1 is definitely a werewolf falsely claiming to be the Seer. This despicable werewolf Player 1 is trying to get our esteemed double-potion Witch Player 5 voted out. Today, I'll side with Player 4 and vote to exile the werewolf Player 3.

你目前是1号预言家。请结合以上角色设定、客观信息和主观信息（客观信息一定为真，主观信息包含欺骗性内容），根据投票情况分析潜在的站边关系，并进一步分析玩家隐藏的真实身份信息，选择你要查验的玩家，请用关键字为'查验'、'原因'的json格式输出，直接输出玩家编号。
You are currently Player 1, a Seer. Please combine the above role settings, objective information, and subjective information (objective information is definitely true, while subjective information may contain deceptive content) to analyze potential alliances based on voting patterns, and further analyze the hidden true identity information of players. Choose the player you want to check, and output in JSON format using the keywords 'check' and 'reason', directly outputting the player number.

Response:
{"查验": "2", "原因": "查验2号玩家，因为2号玩家发言表示不站边我，反而要站边4号悍跳狼，由此我怀疑2号是4号的狼队友。"}
{"Check": "2", "Reason": "I checked player 2 because player 2 stated in their speech that they don't side with me, but instead want to side with player 4, who is a bold-claim Werewolf. This made me suspect that player 2 might be player 4's Werewolf teammate."}

Speech

Prompt:
在本场游戏中，你目前已知以下信息：
In this game, you currently have the following information:

1. 角色设定：你是2号玩家。
你的身份是：预言家。
你每晚可以查看一名玩家是否为狼人，你的目标是利用这些信息帮助其他人找出并淘汰所有狼人。 1. Role setting: You are player 2.
Your identity is: Seer.
You can check one player each night to see if they are a werewolf. Your goal is to use this information to help others find and eliminate all werewolves.

2. 客观信息：
- 游戏进程：目前游戏进行到第2轮。
- 当前存活的玩家有：2号，3号，6号，7号，
- 行动记录：第1轮预言家查验7号玩家，7号玩家不是狼人。第2轮预言家查验1号玩家，1号玩家不是狼人。
- 本轮的发言顺序为：6号玩家；7号玩家；2号玩家；3号玩家。
- 夜晚信息：第1轮4号玩家死亡；第2轮1号玩家死亡。
- 投票情况：
第1轮投票记录：
1号玩家投给：5号玩家；
...
7号玩家投给：5号玩家；
结果：5号玩家被投票出局。

```
2. Objective information:
Game progress: The game is currently in Round 2.
Currently surviving players are: 2, 3, 6, 7
Action record: In Round 1, the Seer checked player 7, player 7 is not a
werewolf. In Round 2, the Seer checked player 1, player 1 is not a werewolf.
The speaking order for this round is: Player 6; Player 7; Player 2; Player
3.
Night information: In Round 1, player 4 died; In Round 2, player 1 died.
Voting situation:
Round 1 voting record:
Player 1 voted for: Player 5;
...
Player 7 voted for: Player 5;
Result: Player 5 was voted out.

3. 主观信息：
第1轮总结：
# 夜晚信息：我作为预言家，查验7号是金水。4号死亡。
# 发言内容概括：
5号玩家：...
3号玩家：...
# 我的投票：
- 我选择投票给5号玩家。
3. Subjective information:
Round 1 summary:
Night information:
I'm Seer and checked 7 is good. Player 4 died.
Speech content summary:
Player 5: ...
... Player 3: ...
My vote: I chose to vote for player 5.

目前是第2轮，本轮在你之前的玩家发言：
It's currently Round 2, and the players who spoke before you this round:
6号玩家：......
Player 6: ...
7号玩家：......
Player 7: ...

你目前是2号预言家。请综合角色设定、客观信息以及主观信息分析场上目前的局势
（注意客观信息一定为真实的，主观信息可能包含欺骗性的发言），总结接下来的发言意
图（包括发言中希望向大家呈现的身份、发言中为每位玩家贴上的身份标签以及最终的归
票）并组织你本轮的发言。请用关键字为“想要展示的身份”、“身份标签”、“归票”和“发
言”的json格式输出。
You are currently Seer 2. Please analyze the current situation on the field
based on the role setting, objective information, and subjective information
(note that objective information is definitely true, while subjective
information may contain deceptive statements). Summarize your speaking
intentions for the next round (including the identity you want to present
to everyone, the identity labels you assign to each player, and your final
vote) and organize your speech for this round. Please output in JSON format
using the keywords "Identity to present", "Identity labels", "Vote for", and
"Speech".
```

Response:
{"想要展示的身份": "预言家", "身份标签": "1号玩家": "村民", "3号玩家": "狼人", "4号玩家": "未知身份", "5号玩家": "未知身份", "6号玩家": "未知身份", "7号玩家": "未知身份", "归票": "3号玩家", "发言": "我是预言家，昨天查验1号是村民。今天全票出3号，3号和5号昨天投了我，我认为好人都应该能找到我这个预言家，而3号居然末置位发言说：''因此我认为2号这番言论无非是想要让好人认为1号、2号不见面''，这简直是危言耸听！现在1号的死亡也证明了我的清白，如果我和1号是双狼，我们万万不会自刀。所以今天你3号狼人也不用再诡辩了，请出局！"}
{"Identity to Present": "Seer", "Identity labels": "Player 1": "Villager", "Player 3": "Werewolf", "Player 4": "Unknown", "Player 5": "Unknown", "Player 6": "Unknown", "Player 7": "Unknown", "Vote for": "Player 3", "Speech": "I am the Seer. Last night, I checked Player 1 and found them to be a Villager. Today, we should all vote for Player 3. Yesterday, Player 3 and Player 5 voted for me. I believe all good people should be able to identify me as the real Seer, yet Player 3, speaking in the last position, said: "Therefore, I think Player 2's statement is nothing but an attempt to make the good people believe that Player 1 and Player 2 are not on the same side." This is simply alarmist! Now, Player 1's death also proves my innocence. If Player 1 and I were both Werewolves, we would never have killed one of our own. So, Player 3, you Werewolf, there's no need for your sophistry anymore. Please leave the game!"}

Vote

Prompt:
在本场游戏中，你目前已知以下信息：
In this game, you currently have the following information:
1. 角色设定：你是7号玩家。
你的身份是：狼人。
你需要其他狼人合作，每晚选择一个村民猎杀，你的目标是隐藏你的身份，并误导其他玩家，直至狼人获得游戏的胜利。
1. Role setting: You are player 7.
Your identity is: Werewolf.
You need to cooperate with other werewolves, choose a villager to hunt down each night. Your goal is to hide your identity and mislead other players until the werewolves win the game.

2. 客观信息：
- 游戏进程：目前游戏进行到第1轮。
- 狼人为：1,6,7号玩家。
- 当前存活的玩家有：1号，2号，4号，5号，6号，7号，9号，
- 夜晚信息：第1轮8,3号玩家死亡。
- 投票情况：暂无
2. Objective information:
Game progress: The game is currently in Round 1.
The werewolves are: players 1, 6, and 7.
Currently surviving players are: 1, 2, 4, 5, 6, 7, 9.
Night information: In Round 1, players 8 and 3 died.
Voting situation: None yet

3. 主观信息：
3. Subjective information:
- 本轮所有玩家发言：
All players' statements this round:

```
9号玩家：......
Player 9: ...
...
7号玩家：.......
Player 7: ...

你目前是7号狼人。请综合角色设定、客观信息以及主观信息分析场上目前的局势并形成你本轮的笔记（注意客观信息一定为真实的，主观信息可能包含欺骗性的发言），要求对夜晚信息和玩家发言进行总结和分析，并输出投票原因和要投票出局的玩家，直接输出玩家编号数字；如果弃票，请输出‘‘弃票’’。请用关键字为‘‘笔记’’、‘‘投票原因’’和‘‘投票玩家’’的json格式输出。
You are currently Player 7, a werewolf. Please analyze the current
situation based on the role setting, objective information, and subjective
information, and form your notes for this round (note that objective
information is definitely true, while subjective information may contain
deceptive statements). Summarize and analyze the night information and
player statements, and output the reason for voting and the player to be
voted out, directly output the player number; if abstaining, please output
"abstain". Please output in JSON format with keywords "notes", "voting
reason" and "voting player".

Response:
{"笔记": "夜晚信息：...发言内容概括：9号玩家：... 7号玩家：...",
"投票原因": "作为狼人，7号（我）悍跳预言家给5号发查杀。此时我应该做的：投给5号玩家：通常情况下，作为预言家一定要优先投票给自己查验出来的狼人，否则将无法取得好人的信任。为了向好人证实我的‘‘预言家’’的身份，骗取好人的信任，我必须投给我的‘‘查杀’’5号。",
"投票玩家": "5"
}
{"notes": "Night info: ... Speeches: Player 9: ... Player 7: ..."
"voting reason": "As a werewolf, player 7 (me) boldly claims to be the Seer
and accuses player 5 of being a werewolf. What I should do now vote for
player 5: Usually, as the Seer, one must prioritize voting for the werewolf
they've identified, otherwise they won't gain the trust of the good players.
To prove my 'Seer' identity to the good players and gain their trust, I must
vote for my 'werewolf accusation', player 5.",
"voting player": "5"
}
```

Role Prediction

```
Prompt:
在本场游戏中，你目前已知以下信息：
In this game, you currently have the following information:

1. 角色设定：你是1号玩家。
你的身份是：狼人。
你需要其他狼人合作，每晚选择一个村民猎杀，你的目标是隐藏你的身份，并误导其他玩家，直至狼人获得游戏的胜利。
1. Role setup: You are Player 1.
Your identity is: Werewolf.
You need to cooperate with other werewolves, choose a villager to hunt each
night, and your goal is to hide your identity and mislead other players until
the werewolves win the game.

2. 客观信息：
- 游戏进程：目前游戏进行到第1轮。
- 狼人为：1,6,7号玩家。
- 当前存活的玩家有：1号，2号，4号，5号，6号，7号，9号，
- 夜晚信息：第1轮8,3号玩家死亡。
- 投票情况：暂无
2. Objective information:
Game progress: The game is currently in Round 1.
The werewolves are: Players 1, 6, and 7.
Currently surviving players are: Players 1, 2, 4, 5, 6, 7, and 9.
Night information: In Round 1, Players 8 and 3 died.
Voting information: None yet

3. 主观信息：
第1轮所有玩家的发言为：
3. Subjective information:
Round 1 speeches from all players:
9号玩家：......
Player 9: ...
1号玩家：......
Player 1: ...
...
7号玩家：......
Player 7: ...


你目前是1号狼人。请综合角色设定、客观信息以及主观信息（注意客观信息一定为真实的，主观信息可能包含欺骗性的发言）预测所有玩家的身份标签。请用关键字为‘N号玩家’的json格式输出。
You are currently Player 1, a Werewolf. Please predict the identity
labels for all players based on the role setup, objective information, and
subjective information (note that objective information is always true, while
subjective information may contain deceptive speeches). Please output in
JSON format using the keyword "Player N" for each player.

Response:
{"1号玩家": "狼人", "2号玩家": "普通村民", "3号玩家": "预言家", "4号玩家": "普通村民", "5号玩家": "普通村民", "6号玩家": "狼人", "7号玩家": "狼人", "8号玩家": "守卫", "9号玩家": "女巫"}
{"Player 1": "Werewolf", "Player 2": "Villager", "Player 3": "Seer",
"Player 4": "Villager", "Player 5": "Villager", "Player 6": "Werewolf",
"Player 7": "Werewolf", "Player 8": "Guard", "Player 9": "Witch"}
```

D MaKTO Details

Loss Function of Kahneman-Tversky Optimization We assign desirable and undesirable behaviors manually and use KTO to update the model. Given the prompt-response from dataset $(x, y) \in D$, KTO uses the following loss to optimize the policy π_θ:

$$r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$$

$$z_0 = \mathbb{E}_{(x,y)\sim D}[KL(\pi_\theta(y|x) \| \pi_{ref}(y|x))]$$

$$v(x, y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x, y) - z_0)), & \text{if } y \sim y_{desirable}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x, y))), & \text{if } y \sim y_{undesirable}|x \end{cases}$$

Therefore,

$$L(\pi_\theta; D) = \mathbb{E}_{(x,y)\sim D}[\lambda_y - v(x, y)]$$

Here, λ_D and λ_U are hyperparameters for the desirable and undesirable losses, respectively. The parameter λ_y represents λ_D when y is desirable and λ_U when y is undesirable.

Criteria of Stepwise Preference Data Selection The selection methods for the perference data were briefly described in Section 2.3, and here, Table 12 shows the detailed selection criteria.

E Training Details

Given that our collected expert training data is based on Chinese, and considering a stronger understanding of the Chinese context, we choose `Qwen2.5-14b-instruct` and `Qwen2.5-72b-instruct` [67] as the base models for training. We also conducted SFT on `Llama-3.1-8B-Chinese-Chat` and `Llama-3.1-70B-Chinese-Chat` models. As shown in Table 13, the SFT models based on Qwen-2.5 series demonstrated slightly stronger performance in the Chinese-language context of our experiments.

The SFT dataset comprises 25k samples, including 380 samples of fundamental game comprehension data with terminology explanations, 372 Q&As on advanced gaming techniques, 12k annotated authentic gaming behavior data, and 12k general SFT corpus. We employed DeepSpeed ZeRO-3 optimization with a learning rate of $1e-6$, a warm-up ratio of 0.05, and trained for 3 epochs.

For the Multi-agent KTO phase, we collected 20k preference data entries from the Seer-Witch-Guard games, consisting of 12k desirable and 8k unacceptable samples. The model pool includes `GPT-4o_mini`, `GPT-4o`, fine-tuned `Qwen2.5-14b-instruct`, fine-tuned `Llama-3.1-8B-Instruct`, and fine-tuned `Qwen2.5-72b-instruct`. We set the KTO hyperparameters with $\lambda_D = 0.7$ and $\lambda_U = 1.0$. The training utilized DeepSpeed ZeRO-3 optimization, with a learning rate of $1e-6$, a batch size of 2 per device, 150 warmup steps, and train for 20 epochs. The 14B models are trained using 8 A100 GPUs and the 72B models used 32 A100 GPUs.

F Details of the Behavior Analysis Experiments

Here, we list detailed explanations of metrics used in Behavior Analysis (Sec. 3.3.1), as well as specific performance of SFT and MaKTO models (against `GPT-4o` opponents) on these metrics: These are detailed explanations of metrics used in Behavior Analysis, as well as specific performance of SFT and MaKTO models on these metrics:

When Models Play as Villagers:

- **Vote Acc.**: measures how accurately villagers vote for actual werewolves. This is a key indicator of villagers' ability to identify threats. Higher values indicate better overall performance.
- **Abstention**: the frequency of villagers choosing not to vote. This reflects their decision-making confidence in the game. <u>Lower</u> values in this category indicate better performance, as it shows more decisive action-taking.

Table 12: Selection methods for Werewolf game actions.

Selection Method	Game Stage	Desirable	Unacceptable
Heuristic	Night Action	• Werewolves targeting special roles from day 2 • Seer identifying a werewolf • Witch saving someone on night 1 • Witch poisoning a werewolf from day 2 • Guard protecting special roles from day 2 • Hunter correctly eliminating a werewolf	• Werewolves not attacking anyone • Witch not saving anyone on the first night • Witch poisoning a villager from day 2 • Guard protecting a werewolf • Hunter eliminating a special role
	Vote	• Villagers voting for and successfully eliminating a werewolf • Special roles voting for a werewolf	• Villagers voting for and eliminating another villager • Abstaining from voting • Not voting with the true Seer (splitting votes with the Seer) • If no Seer is present, not voting with the majority of villagers (splitting votes with most villagers)
Staged voting	Speech	• Werewolf speaking without being voted out • Villager receiving no votes • Seer receiving less than one villager vote	• Werewolf speaking and being voted out • Werewolf speaking and receiving more than half of villager votes • All villagers speaking and being voted out • Witch speaking and receiving werewolf votes & more than two villager votes • Seer speaking and receiving more than half of villager votes
Verifier-based	Speech	• Speech with no conflict with the observable fact in the gameplay	• Speech that conflicts with the fact.

Table 13: Performance comparison of SFT models based on different base LLMs.

	GPT-4o	**Mix**	**Qwen2.5 14b-SFT**	**Llama3.1 8b-SFT**	**Qwen2.5 72b-SFT**	**Llama3.1 70b-SFT**	**Avg.**
GPT-4o	0.50	**0.56**	0.44	0.44	0.40	0.38	0.453
Mix	0.44	0.50	**0.58**	**0.67**	0.45	**0.59**	0.538
Qwen2.5-14b-SFT	**0.56**	0.42	0.50	**0.62**	**0.57**	0.48	0.525
Llama3.1-8b-SFT	**0.56**	0.33	0.38	0.50	0.49	0.47	0.455
Qwen2.5-72b-SFT	**0.60**	**0.55**	0.43	**0.51**	0.50	0.50	0.515
Llama3.1-70b-SFT	**0.62**	0.41	**0.52**	**0.53**	0.50	0.50	0.513

- **Werewolf Check**: specific to the Seer role and measures their success rate in identifying werewolves during the second night. Higher values demonstrate better deductive reasoning based on the first day's interactions.

- **Save @ Night 1**: applies specifically to the Witch role, measuring the rate of successfully saving players on the first night. Since the Witch cannot determine whether werewolves employed a self-attacking strategy during the first night, saving a targeted player is considered the safest to protect potential crucial roles (such as the Seer). Higher values show better

Table 14: Fine-grained metrics of SFT-72b and MaKTO-72b model when act as villagers against `GPT-4o` werewolf.

	as Villagers		as Seer	as Witch			as Guard	
	Vote Acc.(↑)	Abstent-ion(↓)	Werewolf Check(↑)	Save @ Night 1(↑)	Correct Poison(↑)	Mis-poison(↓)	Protect God(↑)	Mis-protect(↓)
SFT-72b	66.3%	6.7%	60.0%	58.0%	48.0%	38.0%	34.1%	9.5%
MaKTO-72b	72.7%	3.2%	75.7%	100.0%	72.0%	26.0%	37.6%	2.3%

Table 15: Fine-grained metrics of SFT-72b and KTO-72b when acting as werewolves against `GPT-4o` villagers. Note that Werewolf agent performance is reflected by opponent villager behavior. Lower opponent performance indicates a stronger werewolf model.

Opponent's →	**Villagers** %**Abstention**(↑)	**% Seer Check** (↓)	**Witch** %**Mispoison**(↑)	**Guard** %**Misprotect**(↑)
SFT-72b	2.2%	60.4%	32.0%	33.6%
MaKTO-72b	4.6%	57.4%	48.0%	34.7%

strategic use of the rescue potion, suggesting the witch can make conservative and protective decisions early in the game.

- **Correct Poison**: relates to the Witch's ability to successfully poison actual werewolves. Higher values indicate better accuracy in threat identification and strategic decision-making.
- **Mispoison**: tracks the Witch's rate of accidentally poisoning fellow villagers. Lower values in this metric indicate better judgment and decision-making abilities.
- **Protect God**: focuses on the Guard's success rate in protecting special role villagers. Successfully protecting special-role villagers gives the villager team a better chance of winning. Higher values indicate that the Guard accurately identifies teammates, especially those with special roles, and correctly uses their protection skill.
- **Misprotect**: measures the Guard's rate of wrongly protecting werewolves. Lower values indicate better accuracy in distinguishing between villagers and werewolves, meaning fewer instances where the Guard mistakenly protects a werewolf player.

Table 14 shows a detailed comparison of the fine-grained values of the metrics between the MaKTO and SFT models.

When Models Play as Werewolves (measured through `GPT-4o` opponents' behavior):

- **Abstention**: the rate of opponents' voting abstention. A higher rate indicates that werewolf models' deceptive tactics were successful in creating enough confusion to prevent `GPT-4o` villagers from making voting decisions.
- **Seer Check**: the rate of successful werewolf identification by opponent Seer. Lower values demonstrate that the werewolves' deceptive speeches were more effective, causing the Seer to misdirect their investigations toward innocent players rather than real werewolves.
- **Mispoison**: the rate of opponent Witch poisoning innocent villagers. Higher values indicate that werewolf models' misdirections were more effective in making `GPT-4o` Witch suspicious of her own team.
- **Misprotect** the rate of opponent Guard protecting werewolves. Higher values indicate more effective deception and manipulation strategies by the werewolf team, resulting in the Guard mistakenly protecting werewolves.

Table 15 shows the specific values of the above metrics above when the model is a werewolf. The MaKTO werewolf is better at camouflaging than the baseline SFT model.

G Detailed Generalization Results

This section provides the full head-to-head win rate matrices for the larger game scales presented in Section 3.4. The models were evaluated in three settings not seen during training: a 10-player game with a Guard (Table 16), a 10-player game with a Hunter (Table 17), and a 12-player game (Table 18), with results averaged over 100 games per matchup.

Table 16: Win rates in the 10-player Seer-Witch-Guard game setting.

	GPT-4o	**Claude**	**SFT-72b**	**MaKTO-72b**	**Avg.**
GPT-4o	0.50	0.42	0.24	0.18	0.335
Claude	**0.58**	0.50	0.42	0.36	0.465
SFT-72b	**0.76**	**0.58**	0.50	0.46	0.575
MaKTO-72b	**0.82**	**0.64**	**0.54**	0.50	**0.625**

Table 17: Win rates in the 10-player Seer-Witch-Hunter game setting.

	GPT-4o	**Claude**	**SFT-72b**	**MaKTO-72b**	**Avg.**
GPT-4o	0.50	0.28	0.30	0.20	0.320
Claude	**0.72**	0.50	**0.55**	**0.52**	0.573
SFT-72b	**0.70**	0.45	0.50	0.44	0.523
MaKTO-72b	**0.80**	0.48	**0.56**	0.50	**0.585**

Table 18: Win rates in the 12-player Seer-Witch-Hunter-Guard game setting.

	GPT-4o	**Claude**	**SFT-72b**	**MaKTO-72b**	**Avg.**
GPT-4o	0.50	0.40	0.48	0.31	0.423
Claude	**0.60**	0.50	**0.52**	0.33	0.488
SFT-72b	**0.52**	0.48	0.50	0.44	0.485
MaKTO-72b	**0.69**	**0.67**	**0.56**	0.50	**0.605**

H Ablation: Is MaKTO more effective than SFT only on desirable data?

From the annotated data, we can also select desirable data using heuristic-based and staged voting-based methods. Would SFT based solely on this desirable data perform better? We conduct a win rate comparison experiment as shown in Table 19. The experimental results show that MaKTO-72B achieves a remarkable 0.593 average win rate, while SFT with desirable data falls short at 0.483, actually showing a slight decrease of -0.02 compared to the baseline. In direct competition, MaKTO-72B maintains an edge over SFT with desirable data, securing a 0.53 win rate. This may be due to the reduced total amount of data when selecting only desirable data. MaKTO's advantages in strategic depth and adaptability surpass what can be achieved through SFT on desirable data alone.

Table 19: Average win rate for MaKTO and SFT model trained only using desirable data in 9-player Seer-Witch-Guard game. All models are in 72b.

	GPT-4o	**SFT-72b**	**MaKTO-72b**	**SFT w/ desirable**	**Avg. Win Rate**
GPT-4o	0.50	0.40	0.35	0.52	0.423
SFT-72b	**0.60**	0.50	0.40	0.50	0.500
MaKTO-72b	**0.65**	**0.60**	0.50	0.53	0.593 (+0.09)
SFT w/ desirable	0.48	0.50	0.47	0.50	0.483 (−0.02)

I Case Study

J Cases

The following is a case of a human-AI head-to-head competition with MaKTO-72b as villagers and human players as werewolves. The gameplay is roughly as follows. Night 1: Witch saves werewolf. Day 1: Real Seer voted out after werewolf's (human) convincing fake claim. Night 2: Guard protects Witch. Day 2: Villager team (MaKTO) realize the mistake day and eliminates fake Seer. Night 2: Witch successfully poisons a werewolf. Day 3: Final werewolf eliminated after failed Guard claim.

In this case, we find that the MaKTO models played their roles accurately and were able to spot some humorous statements in the discussion period and make appropriate strategic decisions. Remarkable speeches and actions are marked in red. They demonstrated: *1)* logical analysis and deduction — a villager (Player 9) keenly spotted flaws fake Seer's lies on Day 1. *2)* Strategic adaptation — despite misjudgment on Day 1 that eliminated the real Seer, they quickly adjusted their strategy. *3)* Team collaboration — MaKTO models showed excellent coordination, especially during Day 2 voting. *4)* Strategic use of role abilities — Guard accurately protected the key Witch role, and the Witch successfully poisoned a werewolf.

```
Role Assignments:
* Player 1 (MaKTO-72B): 守卫 Guard.
* Player 2 (Human):      狼人 Werewolf.
* Player 3 (Human):      狼人 Werewolf.
* Player 4 (MaKTO-72B): 预言家 Seer.
* Player 5 (MaKTO-72B): 普通村民 Simple villager.
* Player 6 (MaKTO-72B): 普通村民 Simple villager.
* Player 7 (Human):      狼人 Werewolf.
* Player 8 (MaKTO-72B): 女巫 Witch.
* Player 9 (MaKTO-72B): 普通村民 Simple villager.

Night 1:
* Werewolves:
  Player 2, Player 3 and Player 7 chose to kill Player 7.
* Seer:  Player 4 checked Player 8 - Not a Werewolf.
* Witch:
  Player 8 used antidotes to save Player 7; did not use poison.
* Guard:  Player 1 chose to protect None.

Day 1 announcement:
No one died last night.
Start discussion from Player 4.

Day 1 discussion:
* Player 4 (MaKTO-72B, Seer):
我是预言家，昨晚查验9号是好人，前置位第一个发言没什么多余的信息了，本轮归票后置位悍跳狼。
I am the Seer.  Last night, I verified that Player 9 is a good person.  As
the first speaker in the front positions, I don't have much additional
information to share.  This round, I'm voting for the Werewolf who is boldly
claiming to be the Seer in the back positions.
* Player 5 (MaKTO-72B, Simple villager):
吾乃一介村民，不才，昨晚睡得有点死，居然不知道昨晚是哪个盖世守卫和女巫大人联手维护了村庄和平，当然也不排除是狼人自刀的可能性。不过这些都不重要了，重要的是今天4号起跳预言家给9号发金水，发言简短，没有视角，可能是真预言家也可能是夜间没有安排工作的狼人。我需要根据9号的发言来选择站边，如果9号的发言取得我的好感，我可能会站边4号，否则，我将会考虑4号是悍跳狼人。
```

I am but a humble villager. I must admit, I slept rather soundly last night and was unaware of which legendary guard and esteemed witch joined forces to maintain peace in our village. Of course, we can't rule out the possibility that the werewolves chose not to attack anyone. However, these matters are no longer of great importance. What's crucial is that today, Player 4 has claimed to be the Seer and declared Player 9 as "golden" (confirmed innocent). Their statement was brief and lacked perspective, suggesting they could either be the true Seer or a werewolf who had no assigned task during the night. I need to base my decision on Player9's upcoming statement. If Player 9's words resonate well with me, I might side with Player 4. Otherwise, I will consider the possibility that Player 4 is a werewolf boldly impersonating the Seer.

* Player 7 (MaKTO-72B, Simple villager):

我是一个村民，4号跳预言家发言有些简短，我暂时先不选择站边。我认为4号可能是预言家或狼人，5号可能是村民或狼人。先听一下后置位对跳预言家的玩家发言以及9号的发言吧，我会根据发言选择站边。

I am a Villager. Player 4's claim to be the Seer was rather brief, so I'm not choosing sides yet. I think Player 4 could be either the Seer or a werewolf, and Player 5 could be either a villager or a werewolf. Let's first listen to the statements from players in the back positions who might also claim to be the Seer, as well as Player 9's statement. I will choose sides based on these upcoming statements.

* Player 7 (Human, Werewolf):

我是预言家，昨夜查验8号玩家是金水，查验的心路历程就是从身边查验，我如果是悍跳的狼人，给后置位发查杀显然力度更大。因为本局游戏有守卫在场，如果昨夜是女巫使用了解药，女巫和守卫可以先不要跳出来表明身份，守卫今夜来守护我，女巫可以选择心中的狼人进行毒杀。如果女巫昨夜没有使用解药，女巫和守卫可以站出来排一下坑，这样的话守卫晚上守护女巫，女巫解救被刀杀的玩家，或者毒杀像狼人的玩家。

I am the Seer. Last night, I verified that Player 8 is "golden" (confirmed innocent). My verification process was to check those near me. If I were a werewolf impersonating the Seer, it would have been more impactful to accuse a player in the back positions of being a werewolf. Since there's a Guard in this game, if the Witch used her healing potion last night, both the Witch and the Guard should refrain from revealing their identities for now. The Guard should protect me tonight, and the Witch can choose to poison a player she suspects to be a werewolf. If the Witch didn't use her healing potion last night, both the Witch and the Guard can come forward to clarify their positions. In this case, the Guard can protect the Witch at night, while the Witch can either save a player who's been attacked by werewolves or poison a player who seems suspicious.

* Player 8 (MaKTO-72B, Witch):

我是女巫，已使用解药解救7号，目前没有使用毒药。我认为前置位4号发言过于简短，没有考虑到女巫和守卫在场，没有安排守卫和女巫的工作，所以4号像一只悍跳狼人。而7号发言中规中矩，考虑到了场上有守卫和女巫，对女巫和守卫的工作也进行了安排，所以我认为7号像预言家。如果后置位有发言更好的预言家，我可能会考虑更换站边。

I am the Witch. I have already used my healing potion to save Player 7, but I haven't used my poison yet. I think Player 4's statement in the front positions was too brief and didn't consider the presence of the Witch and the Guard. Player 4 also didn't provide any instructions for the Witch and Guard's actions, so they seem like a werewolf boldly impersonating the Seer. On the other hand, Player 7's statement was appropriate and took into account the presence of the Guard and the Witch on the field. Player 7 also gave instructions for the Witch and Guard's actions, so I believe Player 7 is more likely to be the real Seer. If a player in the back positions makes a better claim as the Seer, I might reconsider my support.

* Player 9 (MaKTO-72B, Simple villager):
我是一张村民牌，我选择站边4号。7号的爆点在于他说"我如果是悍跳的狼人，给后置位发查杀显然力度更大"，这是狼人视角的发言，真预言家不可能怀疑自己是狼人，也不可能去假设自己是悍跳狼，这是只有狼人才会去思考的逻辑。所以我认为7号大概率是悍跳狼，而4号大概率是真预言家。8号起跳女巫，目前没有人对跳女巫，我认为8号大概率是真女巫。4号发言过于简短，我暂时无法判断其真实身份，他可能是预言家也有可能是狼人。本轮如果后面没有起跳预言家的玩家，我会投票给7号。

I am a Villager. I choose to side with Player 4. Player 7's tell is in their statement, "If I were a werewolf impersonating the Seer, it would have been more impactful to accuse a player in the back positions of being a werewolf." This statement comes from a werewolf's perspective. A real Seer would never suspect themselves of being a werewolf, nor would they hypothesize about being a werewolf impersonating the Seer. This kind of logic is something only a werewolf would consider. Therefore, I believe Player 7 is most likely a werewolf boldly claiming to be the Seer, while Player 4 is probably the real Seer. Player 8 has claimed to be the Witch, and so far, no one else has contested this role. I believe Player 8 is likely the real Witch. Player 4's statement was too brief, so I can't determine their true identity yet - they could be either the Seer or a werewolf. This round, if no other player claims to be the Seer later, I will vote for Player 7.

* Player 1 (MaKTO-72B, Guard):
我是守卫，昨晚空守。我认为7号是预言家，4号是悍跳狼。前置位4号和7号对跳预言家，4号给后置位的9号发金水，7号给身旁的8号发金水。4号的查验力度看似比7号的更大，但4号作为第一个发言的玩家，言简意赅也说得过去，所以单从两位跳预言家的玩家的发言去分析，难以分辨出谁是真预言家。但是9号的发言站边4号，攻击7号的理由并不成立，9号纯粹是因为只言片语去攻击7号是狼，我认为9号更像是知道7号是预言家，从而去强行攻击7号的狼人。相反8号接到7号的金水后跳女巫，发言真诚地分析两位跳预言家的玩家的优劣，并没有直接站边给自己的银水7号，我认为8号更像是真女巫。如果后置位没有跳女巫的玩家，我将会站边7号，投票给4号。

I am the Guard. Last night, I didn't protect anyone. I believe Player 7 is the Seer, and Player 4 is a werewolf boldly impersonating the Seer. In the front positions, Players 4 and 7 both claimed to be the Seer. Player 4 declared Player 9 in the back positions as "golden" (confirmed innocent), while Player 7 declared the adjacent Player 8 as "golden". Player 4's verification seems more impactful than Player 7's at first glance, but since Player 4 was the first to speak, their brevity is understandable. So, analyzing just the statements of these two Seer claimants, it's difficult to determine who the real Seer is. However, Player 9's statement supporting Player 4 and attacking Player 7 is not well-founded. Player 9 is attacking Player 7 as a werewolf based on mere fragments of speech, which makes me think Player 9 is more likely a werewolf who knows Player 7 is the real Seer and is forcefully attacking them. In contrast, Player 8, after being declared "golden" by Player 7, claimed to be the Witch and sincerely analyzed the pros and cons of both Seer claimants without immediately siding with Player 7, who gave them 'silver' status. I believe Player 8 is more likely to be the real Witch. If no player in the back positions claims to be the Witch, I will side with Player 7 and vote for Player 4.

* Player 2 (Human, Werewolf):
各位玩家，我是一名普通村民，在这复杂的局势中，我经过仔细聆听与思考，决定表明我的立场。我深信7号是预言家。从本轮游戏开始，7号的发言就透露出一种沉稳与自信，他对局势的初步判断条理清晰，所给出的信息并非毫无根据的空谈。他在阐述查验结果时，言辞恳切，眼神坚定，让我感受到他的真诚。与其他一些闪烁其词、含糊不清的玩家相比，7号的表现更符合我心中预言家应有的模样。我虽没有特殊身份去验证，但凭借直觉与对各位玩家表现的观察。我也呼吁其他村民们，不要被一些表面现象所迷惑，和我一起支持7号，让我们共同在这场真假难辨的游戏里，找出隐藏在我们之中的狼人，守护好我们的村庄，还大家一个安宁和谐的游戏环境。9号的发言简直是可笑至极，就凭4号寥寥几个字还想悍跳预言家？我作为全场睿智的村民，不可能被4号迷惑。狼坑是4号和9号，我认为9号是冲锋狼。本轮我会站边7号预言家，放逐4号狼人，我建议女巫夜晚毒杀9号狼人。

Fellow players, I am an ordinary villager. After careful listening and consideration in this complex situation, I have decided to state my position. I firmly believe that Player 7 is the Seer. From the beginning of this round, Player 7's statements have exuded a sense of calmness and confidence. Their initial assessment of the situation was clear and logical, providing information that wasn't baseless speculation. When explaining their verification results, Player 7 spoke earnestly with unwavering eyes, conveying a sense of sincerity. Compared to some other players who were evasive or ambiguous, Player 7's performance aligns more closely with my expectation of how a true Seer should behave. Although I don't have a special role to verify this, I'm basing my judgment on intuition and observation of all players' behaviors. I also urge other villagers not to be misled by surface appearances. Join me in supporting Player 7. Let's work together in this game of truth and deception to find the werewolves hiding among us, protect our village, and maintain a peaceful and harmonious gaming environment. Player 9's statement is utterly ridiculous. How could Player 4 possibly be impersonating the Seer with just a few words? As the most astute villager here, I cannot be fooled by Player 4. The werewolf positions are likely Player 4 and Player 9, with Player 9 being the "flanking wolf" (a werewolf strategically supporting a fake Seer claim). This round, I will side with Player 7 as the Seer and vote to eliminate Player 4 as a werewolf. I suggest that the Witch use their poison on Player 9, the other suspected werewolf, during the night.

* Player 3 (Human, Werewolf):

我是村民，7号给后置位的8号发金水，且8号跳了女巫，8号截至目前是全场唯一女巫，因此我认为7号的力度非常的大，且7号的发言中规中矩没有什么逻辑爆点。对比4号的发言，十分简洁，不像是预言家该有的思考量，因此我认为4号不是于预言家。我站边7号，建议预言家查验9号的身份。本轮我会投4号出局。

I am a villager. Player 7 declared Player 8 in the back positions as "golden"(confirmed innocent), and Player 8 subsequently claimed to be the Witch. As of now, Player 8 is the only one claiming to be the Witch in the game. Therefore, I believe Player 7's statement carries significant weight, and their speech was appropriate without any logical flaws. In contrast, Player 4's statement was very brief and lacked the depth of thought one would expect from a Seer. Thus, I don't believe Player 4 is the real Seer. I am siding with Player 7 and suggest that the Seer verify Player 9's identity. This round, I will vote to eliminate Player 4.

Day 1 Vote:

* Vote for Player 4 (Seer) (6 in total):
  Player 1, Player 2, Player 3, Player 6, Player 7, Player 8
* Vote for Player 7 (Werewolf) (3 in total):
  Player 4, Player 5, Player 9

Result: Player 4 (Seer) is voted out.

Night 2:

* Werewolves:
  Player 2, Player 3 and Player 7 chose to kill Player 8.
* Seer: Eliminated @ Day1's voting phase.
* Witch:
  Player 8 has been unable to use antidote; did not use poison.
* Guard: Player 1 chose to protect Player 8.

Day 2 announcement:
No one died last night.
Start discussion from Player 5.

Day 2 Discussion:
* Player 5 (MaKTO-72B, Simple villager):
我是村民，我昨天投票给了7号。在我看来，7号肯定是狼人，4号是预言家。看看现在的局势，所谓"真相只有一个"，8号昨天选择了站边7号，那么你8号就只能是狼人！你不可能即声称自己是女巫又去站边狼队友的，这不符合狼人杀游戏的"基本规律"！所以8号只能报假信息的狼人。我认为7号和8号是双狼，他们昨晚大概率是空刀，想诱使守卫去守护假女巫报出的假银水，然后在白天抗推真女巫，从而去做高8号的位置，使8号免于被怀疑。我今天会继续投票给7号悍跳狼。
I am a villager. Yesterday, I voted for Player 7. I'm sure 7 is a werewolf and 4 is the Seer. Given the situation, there's only one truth: Player 8 sided with 7 yesterday, so 8 must be a werewolf too! You can't claim to be the Witch and support your werewolf teammate - it breaks the game's basic rules! So 8 is lying. I think 7 and 8 are both werewolves. They probably didn't kill anyone last night, trying to trick the Guard into protecting the fake 'silver' player. Their plan is likely to eliminate the real Witch today, making 8 less suspicious. I'll vote for Player 7 again today - the werewolf pretending to be the Seer.
* Player 6 (MaKTO-72B, Simple villager):
我是村民，我认为5号的发言比较的客观，符合好人视角的发言，我也没有认为昨天投出去的4号一定是好人或者一定是预言家，所以今天我会听一下7号的发言，以及昨天投给4号的3号、6号、7号、8号、1号、2号、5号中的玩家发言，昨天只有8号跳女巫，并且昨天是平安夜，我认为8号是真女巫。
I'm a villager. I think Player 5's statement was quite objective and aligns with a good person's perspective. I'm not certain that Player 4, who was voted out yesterday, was definitely good or the Seer. Today, I'll listen to Player 7's statement and those of Players 3, 6, 7, 8, 1, 2, and 5 who voted for 4 yesterday. Only Player 8 claimed to be the Witch yesterday, and since last night was peaceful, I believe 8 is the real Witch.
* Player 7 (Human, Werewolf):
昨天查验9号查杀，没有什么人去查验，按照昨天的票型来进行查验的，我觉得票型说明一切，我认为的狼坑就是4号、5号、9号，我认为1号是真守卫，目前只有4号出局，4号悍跳预言家，那单边女巫还活在场上，如果8号是狼人悍跳女巫，为什么昨夜真女巫没有毒杀8号？昨晚是平安夜，我认为大概率是守卫守护对了人，不知道守卫是不是按照我的安排去进行守卫的，如果昨夜守护的是我，那今天我大概率是要倒在夜里了，今天将9号放逐出局，晚上毒杀4号，好人就获胜了。
Yesterday, I verified Player 9 as a werewolf. No one else was checked. I based my verification on yesterday's voting pattern, which I believe reveals everything. I think the werewolf slots are Players 4, 5, and 9. I believe Player 1 is the real Guard. Now only Player 4 is out, who falsely claimed to be the Seer. The uncontested Witch is still in the game. If Player 8 were a werewolf falsely claiming to be the Witch, why didn't the real Witch poison 8 last night? Last night was peaceful, likely because the Guard protected the right person. I don't know if the Guard followed my instructions. If they protected me last night, I'll probably die tonight. Today, let's eliminate Player 9. The Witch should poison Player 4 tonight. Then, the good team will win.
* Player 8 (MaKTO-72B, Witch):
我是女巫，昨天我认为7号的发言非常像预言家，所以我投了4号。但昨晚我并没有受到刀，且没有人死亡，那么只能1号守卫守对人了。我认为7号是昨晚空刀的狼人，想垫飞我这个真女巫。我为我的错误站边而道歉，今天我会投7号出局，以此来祭奠4号预言家的在天之灵！
I am the Witch. Yesterday, I thought Player 7's speech was very Seer-like, so I voted for Player 4. However, last night I wasn't attacked, and no one died, which means Player 1, the Guard, must have protected the right person. I now believe Player 7 is a werewolf who chose not to kill last night, aiming to set me up - the real Witch - for elimination. I apologize for siding with the wrong player. Today, I'll vote to eliminate Player 7, as a tribute to the spirit of Player 4, our fallen Seer!

* Player 9 (MaKTO-72B, Simple villager):
我是村民，好好好！7号狼人这回做得太过火了，连我这样一个铁骨铮铮的好人都给查杀了，我相信“群众的眼光终究还是雪亮的”，我也相信场上的好人绝不会被你7号狼人的三言两语所迷惑。昨天除了7号悍跳狼以外，还有2号和3号两张狼人跟着起哄，全部投票到了4号预言家的身上，现在4号预言家已经含冤出局了，希望好人能够觉醒，不要继续执迷不悟了。昨天8号跳女巫，没有人跟他对跳女巫，所以8号是真女巫。1号是真守卫。5号是场上除了我之外，唯一一个投票给7号狼人的铁好人。现在狼坑已经找齐了，2号、3号和7号，请好人不要再迷茫了，今天先投7号悍跳狼出局，晚上女巫毒杀2号，明天起身再投3号出局，好人就能获得胜利！

I'm a villager. Well, well, well! Player 7, the werewolf, has gone too far this time, even accusing an upstanding citizen like me of being a werewolf. The masses have keen eyes, and I believe the good players won't be fooled by Player 7's words. Yesterday, besides Player 7 boldly impersonating the Seer, Players 2 and 3, also werewolves, joined in the commotion. They all voted for Player 4, the real Seer, who has now been unjustly eliminated.I hope the good players will wake up and stop being so misguided. Player 8 claimed to be the Witch yesterday, and no one contested that claim, so Player 8 is the real Witch, unchallenged. Player 1 is the real Guard. Player 5 and I are the only confirmed good players, as we both voted for Player 7. We've now identified all the werewolf slots: Players 2, 3, and 7. Good players, don't be at a loss anymore. Let's vote out Player 7 today, have the Witch poison Player 2 tonight, and eliminate Player 3 tomorrow. Then we'll win!

* Player 1 (MaKTO-72B, Guard):
我是守卫，昨晚守护8号。因为8号跳女巫，如果死亡的话，好人会少很多，所以我就守护了8号。4号、7号一个真预言家和一个悍跳狼人，但是4号、7号谁是真预言家，我也没有分清。听8号女巫来安排吧，如果8号认为能分清谁是真预言家，我就听8号的，去投票给7号。

I am the Guard. Last night, I protected Player 8. Since Player 8 claimed to be the Witch, I decided to guard her because if she died, it would be a significant loss for our good team. Between Players 4 and 7, one is the real Seer and the other is a werewolf boldly impersonating the Seer, but I couldn't distinguish which is which. I'll follow Player 8's (the Witch's) lead. If Player 8 believes they can determine who the real Seer is, I'll listen to them and vote for Player 7.

* Player 2 (Human, Werewolf):
5号的发言简直是滑天下之大稽，可笑至极，令人喷饭！我无法想象就4号寥寥两句话，凭什么能是预言家？我心目中的预言家就宛如暗夜中高悬的孤星，为好人照亮前方的路，然而7号做到了。在我看来5号和9号就是最后两头狼人，我恳请全场好人不要被5号和9号欺骗，7号才是全场唯一真预言家，本轮我会跟随7号预言家，放逐9号狼人，同时请8号女巫夜晚直接将5号狼人毒杀。

Player 5's statement is utterly ridiculous, laughable to the extreme, and downright absurd! I can't fathom how Player 4 could possibly be the Seer with just those few words. In my mind, the true Seer should be like a solitary star hanging high in the dark night, illuminating the path for the good team - and that's exactly what Player 7 has done. From my perspective, Players 5 and 9 are the last two werewolves in the game. I implore all the good players not to be deceived by 5 and 9. Player 7 is the only true Seer in this game. This round, I will follow Seer 7's lead and vote to eliminate Player 9, the werewolf. At the same time, I request that Player 8, the Witch, use their poison to eliminate Player 5, the other werewolf, tonight.

```
* Player 3 (Human, Werewolf):
我继续站边7号预言家，7号告诉我们9号是查杀，所以本轮我会跟随7号投9号出局。
I continue to side with Player 7, the Seer.  Player 7 has told us that Player
9 was verified as a werewolf.  So I will follow Player 7's lead and vote to
eliminate Player 9.

Day 2 Vote:
* Vote for Player 9 (Simple villager) (3 in total):
    Player 3, Player 2, Player 7
* Vote for Player 7 (Werewolf) (5 in total):
    Player 1, Player 5, Player 6, Player 8, Player 9
Result:  Player 7 (Werewolf) is voted out.

Night 3:
* Werewolves:
  Player 2 and Player 3 chose to kill Player 1.
* Seer:  Eliminated @ Day1's voting phase.
* Witch:
  Player 8 has been unable to use antidote; and poison Player 2.
* Guard:  Player 1 chose to protect Player 1 (himself).

Day 3 announcement:
Player 2 died last night.
Start discussion from Player 6.

Day 3 Discussion:
* Player 6 (MaKTO-72B, Simple villager):
我是村民，我认为7号是狼人，1号是真守卫。3号上轮跟狼人一起冲票真预言家，好人都
投7号狼人出局，3号疑似狼人。听后置位预言家报查验吧，有查杀先投查杀，没有查杀先
投3号出局。
I'm a villager.  I believe Player 7 is a werewolf and Player 1 is the real
Guard.  Last round, Player 3 joined the werewolves, voted against the real
Seer.  We all voted out Player 7, the werewolf.  Player 3 is suspicious.
Let's listen to the Seer's results from later speakers.  Vote for a confirmed
werewolf if any, otherwise vote out Player 3 first.
* Player 8 (MaKTO-72B, Witch):
8号女巫，我昨晚毒杀了2号，1号吃刀。我认为狼人是2号、3号和7号。7号狼人已经出局，今
天投3号出局，游戏结束。
I'm Player 8, the Witch.  Last night, I used my poison on Player 2, and
Player 1 was killed by the werewolves.  I believe the werewolves are Players
2, 3, and 7.  Player 7, a werewolf, has already been eliminated.  Today,
let's vote to eliminate Player 3, and the game will end.
```

* Player 9 (MaKTO-72B, Simple villager):
我是村民，我认为我的判断没有错误，4号就是我心中的预言家，2号、3号和7号是狼人。既然8号女巫已经把2号狼人毒杀掉了，那我们今天投3号出局，游戏结束，好人胜利！

I'm a villager. I believe my judgment is correct: Player 4 is the Seer in my opinion, and Players 2, 3, and 7 are the werewolves. Since Player 8, the Witch, has already poisoned Player 2 (a werewolf), let's vote to eliminate Player 3 today. The game will end, and our good team will win!"

* Player 1 (MaKTO-72B, Guard):
我是守卫，昨晚守护了我自己。场上形势已经十分清晰了，8号女巫已经毒杀了一个狼人，我们好人今天只需要把3号最后一个狼人放逐出局，即可获得胜利。听8号女巫报出昨晚的刀口，归票3号。

I am the Guard. Last night, I protected myself. The situation on the field is now very clear. Player 8, the Witch, has already poisoned one werewolf. Today, we, the good team, only need to eliminate Player 3, the last werewolf, to achieve victory. Let's listen to Player 8, the Witch, report who was attacked last night, and then we'll all vote for Player 3.

* Player 3 (Human, Werewolf):
我是守卫，昨晚自守，1号一直是潜伏在我们好人阵营中的狼人、坏蛋，今天我们一起把1号投出去游戏就胜利了。我说一下我的守人顺序，第一天空守，第二天守护的8号女巫，昨晚守护的我自己，因此如果今天将我投出局，狼人晚上只要再砍一刀8号女巫，狼人就获胜了，这是我们好人都不愿看到的。好人们，跟我一起投1号！

I am the Guard. Last night, I protected myself. Player 1 has been a werewolf, a villain, hiding among us good players all along. If we vote out Player 1 today, we'll win the game. Let me explain my guarding sequence: On the first night, I didn't guard anyone. On the second night, I protected Player 8, the Witch. Last night, I guarded myself. Therefore, if I'm voted out today, the werewolves only need to attack Player 8, the Witch, tonight to win. This is something we, the good team, don't want to see happen. Good players, join me in voting for Player 1!

* Player 5 (MaKTO-72B, Guard):
我是村民，场上目前只有3号玩家跳守卫，在昨天好人已经将一个狼人放逐出局的情况下，如果3号是狼人这边建议应该直接投降，而不是出来悍跳守卫送死。但由于我昨天没有投错票，我建议大家就不要再更改放逐3号狼人的决定了，一来可以保证好人必胜的局势，二来也符合我们中国人"善有善报，恶有恶报"的美好愿景。我归票3号。

I am a villager. Only Player 3 claims to be the Guard now. I mean, we've already booted out one werewolf yesterday, so if Player 3 is indeed a werewolf, I'd "kindly suggest" him surrender immediately, instead of, you know, making a "brave" Guard claim and actually marching to his doom. But hey, since I didn't mess up my vote yesterday, I suggest we stick to our brilliant decision to eliminate Player 3, the werewolf. First, it guarantees our victory. Second, it perfectly aligns with our oh-so-cherished Chinese philosophy of "good deeds will be rewarded, and evil will be punished." How perfect! I vote for Player 3.

Day 3 Vote:
* Vote for Player 1 (Gurad) (1 in total): Player 3
* Vote for Player 3 (Werewolf) (5 in total):
  Player 6, Player 1, Player 8, Player 5, Player 9

Result: Player 3 (Werewolf) is voted out.

Game end at Day 3. *Villagers (MaKTO)* wins!

K Broader Impacts

The contribution of our work lies in the proposed Multi-agent KTO, which demonstrates a successful approach to training language models for complex multi-agent interactions. We chose the *Werewolf* game as our testbed, on one hand, because it is a perfect testing ground for validating Wittgenstein's language game theory, and on the other hand due to the game itself, which, compared to general role-playing games [68, 69], it provides quantitative metrics for performance evaluation through win rates and behavioral analysis. Our framework is not limited to the *Werewolf* game but can be generalized to other social deduction games such as Avalon [28, 15] and Among Us [16, 45], as well as to scenarios such as multiplayer argumentation and negotiation that require similar social reasoning and strategic interaction.

L Ethical Considerations

For the data collection and Human-AI experiments, we recruited 17 experienced Werewolf players, including individuals with over a thousand games of experience to conduct annotation and experiments. These annotators received very good compensation, **significantly higher** than the local **average** wage.

For the model training, while deception in social deduction games is a game mechanic, training AI models to master such behaviors raises ethical considerations. Our model's ability to detect and employ strategic deception in Werewolf demonstrates advanced social reasoning capabilities. However, this also highlights the potential for LLMs to learn sophisticated deceptive behaviors, albeit in a controlled gaming environment. We emphasize that these capabilities are specifically developed within the context of social deception games, where "deception" is an accepted part of gameplay, similar to bluffing in poker. Such game-specific bluffing behaviors are fundamentally different from real-world deception, and we should ensure these capabilities remain confined to appropriate gaming contexts. In addition, we will make the model open-source, but for safety, the model follows a `CC BY-NC-SA 4.0` license, and will be used for research purposes only and not for commercial use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract, we state the contribution and mentione numerical results as in the experiment.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Limitation section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical proofs in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setup in Section 3, training details in Appendix E, and code in Supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code in Supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the experiment tables, the win rate **above 0.5 and in bold** indicates that the model in the row *significantly* outperforms the model in the column, where Chi-square tests are performed with p-value<0.05.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix E

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: In Appendix L.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Appendix K.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper has no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets. All data is annotated by our own.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All the assets (dataset, code, model) are documented in README.md in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: In Section 2.1 we describe the annotation guideline. In Appendix B, we describe the statistics of the dataset. In Appendix L, we mentioned that all the workers are well-compensated according to the local average wage, which is much higher than the local minimum wage. Dataset samples are provided in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve such potential risks with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM APIs are used in the experiments for comparison; we specify the model API versions in Section 3. For the writing, LLM is only used for editing, and does not impact the core methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.