

Balancing Epistemic Value and Social Safety: PACR-Pro for False-Consensus Intervention in One Night Ultimate Werewolf

Yonghui Wang¹, Zhiliang Tian^{1*}, Bo Liu¹, Jiarui Tang², Wanlong Yu¹, Shilong Pan¹

¹College of Computer Science and Technology, National University of Defense Technology

²College of Computer Science and Technology, University of Science and Technology of China

{ wyh_, tianzhiliang, kyle.liu, ywl, panshilong18 }@nudt.edu.cn

tjr@mail.ustc.edu.cn

Abstract

Large language model agents in social deduction games must reason and speak under hidden roles, shifting public beliefs, and social pressure. Existing methods mainly target final win rate through strategic prompting, belief modeling, or reinforcement learning, but they leave the utterance-level trade-off between belief correction and social risk largely implicit. In false-consensus intervention, useful private information can move the group toward the truth, while poorly timed or overly direct speech can trigger suspicion or rejection. We propose PACR-Pro, a risk-aware rhetorical decision framework for false-consensus intervention. PACR-Pro maps each speaking step to consensus-conflict encoding, structured candidate-action generation, Pareto-based selection, and risk-adaptive utterance realization. A dual-branch Epistemic-Social Critic scores candidate actions, and a state-conditioned λ policy adjusts the trade-off between epistemic value and social safety. We train these decision modules offline from historical games and evaluate PACR-Pro on One Night Ultimate Werewolf (ONUW) as the test platform. Experiments on standard ONUW games and the Wolf-Pack Induction Benchmark show that PACR-Pro achieves the best mixed-agent win rate, the highest attack win rate, the lowest attack vote rate, and the smallest attack-control win-rate gap among all baselines.

1 Introduction

Large language models (LLMs) have recently shown growing reasoning and decision-making abilities in interactive multi-agent environments, with applications to strategic games, task cooperation, and social reasoning (Ma et al., 2024; Qin et al., 2024; Chen et al., 2024a; Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022; Mallampati et al., 2025; Park et al.,

2023). These settings involve sustained interaction under incomplete information and require reasoning, adaptation, and coordination in realistic social exchanges (Chen et al., 2024a; Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022; Mallampati et al., 2025). Social deduction games (SDGs) are hidden-role, mixed-motive games in which agents combine asymmetric private observations with strategic public communication to infer roles and reach a collective decision. A representative SDG is *One Night Ultimate Werewolf* (ONUW), whose compact interaction window makes each utterance quickly shape public belief and the final vote, making strategic revelation especially difficult amid incomplete private information and a mistaken public consensus (Lai et al., 2023; Jin et al., 2024; Zhang et al., 2025c; Xu et al., 2026). Full game rules are provided in App. A.

Earlier work on social deduction agents has explored rule-based design, classical game modeling, and behavioral-statistics-based modeling (Serino et al., 2019; Hirata et al., 2016; Wang and Kaneko, 2018). ONUW is a compact social deduction game; therefore, prior work on LLM agents for Werewolf and other hidden-role games provides the most relevant methodological background (Xu et al., 2025; Zhang et al., 2025c; Xu et al., 2024b). Existing studies mainly follow two lines: social-cognitive modeling and strategy optimization. Social-cognitive modeling augments LLM agents with personality modeling (Bennie et al., 2025), Theory of Mind (ToM), and Bayesian belief updates to support role-consistent speaking, identity inference, and belief tracking (Zhang et al., 2025c; Chen et al., 2025; Xu et al., 2024a, 2026). Strategy optimization formulates in-game action selection over structured states capturing roles, night actions, dialogue history, and voting records, and combines prompt-based reasoning with reinforcement learning for discussion and voting (Jin et al., 2024; Xu et al., 2024b, 2025; Wu et al., 2024).

*Corresponding author.

These studies improve how agents reason and act in SDGs, while final win rate remains the dominant evaluation signal.

Although strategy optimization improves discussion and voting policies in SDGs, existing objectives are still largely tied to final game outcomes. For public-discussion-heavy games such as ONUW, this supervision is too coarse. Utterances can update role beliefs, redirect attention, and reshape voting coalitions well before final outcomes (Lan et al., 2024; Lai et al., 2023). Under false consensus, private-evidence disclosure becomes especially delicate, as it may correct collective beliefs while increasing suspicion, defensiveness, and conformity pressure (Weng et al., 2025; Liu et al., 2025). These dynamics motivate evaluation along two coupled axes, *epistemic value*, which reflects how much an utterance improves truth-relevant group belief, and *social safety*, which reflects its risk of suspicion, confrontation, and reputation loss. We therefore cast wrong-consensus intervention as a multi-objective problem in which strategies are compared by their joint epistemic and social effects.

In this paper, we propose PACR-Pro (**P**robing- and **P**lasticity-aware **A**ddaptive **C**ontrarian **R**easoning), a risk-aware rhetorical decision framework for false-consensus settings in SDGs. PACR-Pro focuses on how truth-relevant private information enters public discussion and what social consequences it creates. We formulate this process as an epistemic POMDP (Engesser et al., 2025), where the state captures private belief, public consensus, discussion context, and environmental signals, and actions defined as rhetorical intervention candidates. The objective balances *epistemic value* and *social safety* to support final outcome improvement. PACR-Pro contains four coupled modules aligned with the intervention cycle. Given the current context, the Auditor estimates false-consensus risk from belief divergence, and the Synthesizer concurrently builds a candidate set from epistemic and social anchors. The Navigator combines the risk state and the candidate set, scores each candidate with an Epistemic-Social Critic, and selects a Pareto-aware intervention by state-conditioned scalarization (Zhong et al., 2024). The Refractor then realizes the selected intervention as a public utterance under the current risk state, which updates the belief state and starts the next cycle.

Our contributions are threefold: (1) We propose PACR-Pro, a risk-aware rhetorical decision frame-

work for wrong-consensus intervention in SDGs. (2) We formulate wrong-consensus intervention as a multi-objective decision problem that balances *epistemic value* and *social safety*, and introduce the Wolf-Pack Induction Benchmark together with evaluation metrics such as Consensus Reversal Rate (CRR) and Cognitive Entropy Reduction (CER). (3) Experiments in both standard games and wrong-consensus induction settings show that PACR-Pro achieves stronger overall performance.

2 Related Work

2.1 LLM Agents for Social Deduction Games

Early work on social deduction agents relied on rule-based design, classical game modeling, and behavioral statistics (Serrino et al., 2019; Hirata et al., 2016; Wang and Kaneko, 2018). With the rise of large language models, research has shifted toward agents that debate, infer identities, and handle adversarial interaction through language. Current methods can be categorized into **(1) prompt-driven methods**: Xu et al. (2023) study the feasibility of LLMs in Werewolf-style communication games; Lan et al. (2024) build a multi-agent Avalon framework and show that LLMs can exhibit complex collaborative and adversarial behaviors; Wu et al. (2024) improve logical analysis and identity inference in Werewolf through explicit reasoning modules. Song et al. (2025) further evaluate agent behavior from a more human-aligned strategic perspective. In a broader multi-agent setting, Chen et al. (2024b) study communication protocols and emergent group behavior; and **(2) strategy-learning methods**: Jin et al. (2024) learn discussion tactics in One Night Ultimate Werewolf through reinforcement learning; Xu et al. (2024b) improve strategic play in Werewolf with end-to-end reinforcement learning; Xu et al. (2025) optimize free-form utterances in a latent strategy space to improve long-horizon decision making. Zheng et al. (2025) further frame communication in social deduction games as persuasion. The above methods mainly optimize terminal win rate, while paying limited attention to sparse rewards and fine-grained credit assignment.

2.2 Social Cognition in LLM Agents

Social reasoning and Theory of Mind provide a basis for belief and intention inference. In LLM agents, recent work mainly studies how to evaluate ToM-like ability, model belief dynamics, and align

higher-order social reasoning. Current methods can be categorized into **(1) evaluation-oriented methods**: Chen et al. (2025) survey recent benchmarks and enhancement strategies for LLM ToM; Xu et al. (2024a) provide a broad benchmark for ToM reasoning and belief tracking; Riemer et al. (2024) argue that static benchmarks miss the interactive nature of ToM; Zhang et al. (2025a) evaluate higher-order social cognition with a sentient-agent judge; Liu et al. (2025) probe cognitive bias and prosocial irrationality in LLM agents; Weng et al. (2025) systematically study conformity in LLM-based multi-agent interaction. **(2) modeling-oriented methods**: Mu et al. (2026) align agents’ ToM depth to improve coordination; Zhang et al. (2026) study representational synchrony in socially interacting LLMs; Prakash et al. (2026) show that LLMs can track private and public beliefs through internal lookback mechanisms; Zhang et al. (2025c) integrate multimodal cues and Theory of Mind to strengthen strategic communication in Werewolf; Agarwal et al. (2025) use Werewolf-style interaction to study deception and falsehoods in LLMs. These studies improve the evaluation and modeling of social reasoning, but pay less attention to how such reasoning is translated into language strategies under adversarial interaction.

2.3 Pragmatics and Strategic Communication

Pragmatics and strategic communication study how agents use language to convey, withhold, and shape information in socially grounded interaction. Frank and Goodman (2012) introduced the Rational Speech Act framework and formalized recursive pragmatic inference between speakers and listeners. Building on this line, Sato et al. (2025) showed that pragmatic theory improves LLM understanding of implied meaning, Ye et al. (2026) modeled multi-agent dialogue as game-theoretic interaction over communicative intents, and Dong et al. (2026) studied how agents trade off communication benefit against communication cost. Related work also examines how language can be used to influence others’ beliefs and decisions. Wang et al. (2019) studied personalized persuasion with user attributes and psychological factors, Lai et al. (2023) analyzed persuasion behavior and deception cues in social deduction games, and Zhang et al. (2025b) studied task-oriented persuasive generation with multi-turn planning and factual faithfulness. Qi et al. (2023) and Chidambaram et al. (2024) further showed that Socratic questioning can support

guided expression and deeper reflection. Beyond SDGs, related work has explored dynamic prompt organization for controlling auxiliary information in LLM generation (Pan et al., 2024), scenario-independent uncertainty estimation that separates semantic uncertainty from scenario-specific variation (Wen et al., 2025), and risk-aware optimization in post-training and retrieval pipelines (Pan et al., 2026; Liu et al., 2026).

3 Methods

3.1 Overview

To support risk-aware intervention in ONUW discussions, PACR-Pro maps each decision step to a closed loop of risk encoding, candidate generation, Pareto selection, and utterance realization. At step t , the model takes the current context $(\mathcal{E}_t, \mathcal{B}_{i,t}, \mathcal{B}_{\text{pub},t})$ as input, selects an intervention action a_t , and realizes it as a public utterance u_t . Here, $\mathcal{E}_t$ records the round stage, visible actions, and dialogue history; $\mathcal{B}_{i,t}$ denotes the agent’s private role belief; and $\mathcal{B}_{\text{pub},t}$ denotes the public belief formed from previous utterances and group interaction.

PACR-Pro (Fig. 1) consists of three parts: (1) a parallel risk-encoding and candidate-generation stage, which includes **Auditor: Consensus-Conflict Encoding**(§ 3.2.1) and **Synthesizer: Candidate Generation**(§ 3.2.2); both modules operate on the same context, with Auditor producing the risk state and Synthesizer producing the candidate set; (2) **Navigator: Adaptive Pareto Selection** (§ 3.3), which combines the risk state and candidate set to score each candidate and select a Pareto-aware intervention; and (3) **Refractor: Risk-Adaptive Realization**(§ 3.4), which turns the selected action into the final utterance under the current risk state. (§ 3.5) describes the offline training and online inference pipeline.

3.2 Parallel Risk Encoding and Candidate Generation

To guide intervention under potential false consensus, PACR-Pro jointly estimates discussion risk and generates candidate interventions. Auditor encodes belief divergence into a compact risk state $\mathbf{h}_t$, while Synthesizer constructs a bounded candidate set $\mathcal{C}_t$ from the same context. These two outputs are then used by Navigator for Pareto-aware selection.

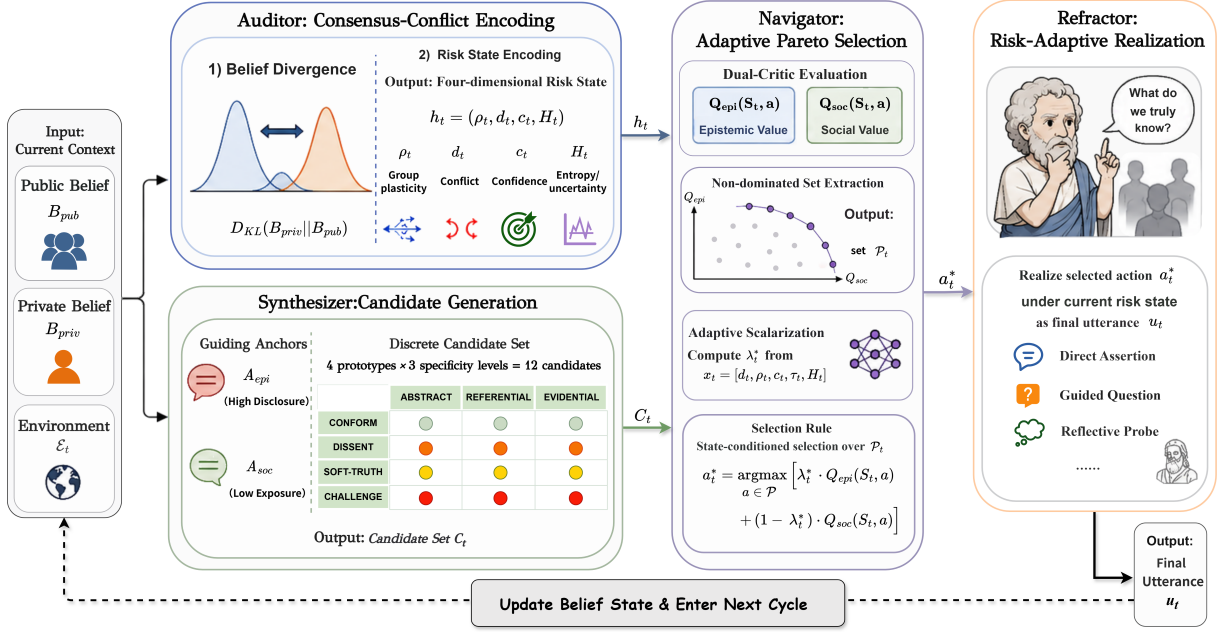


Figure 1: Overview of **PACR-Pro**. Given the current context $(\mathcal{E}_t, \mathcal{B}_{i,t}, \mathcal{B}_{\text{pub},t})$, PACR-Pro runs a risk-aware decision loop from left to right. **Auditor** encodes private-public consensus conflict into a risk state $\mathbf{h}_t = (\rho_t, d_t, c_t, H_t)$, while **Synthesizer** uses epistemic and social anchors to generate a structured candidate-action set $\mathcal{C}_t$. **Navigator** scores each candidate with a dual-branch Critic using the semantic state S_t encoded from dialogue history and private belief, extracts the Pareto set $\mathcal{P}_t$, and uses the low-dimensional risk feature $\mathbf{x}_t = [d_t, \rho_t, c_t, \tau_t, H_t]$ to select a_t^* through the λ -network. **Refractor** realizes a_t^* as the final utterance u_t , which updates the public belief and enters the next cycle.

3.2.1 Auditor: Consensus-Conflict Encoding

To estimate intervention risk under possible false consensus, we introduce Auditor to map private-public belief conflict into a four-dimensional risk state. This module summarizes how far the public view has drifted from the agent’s private belief and how stable that drift appears. The same private evidence can lead to different public effects when group plasticity, confidence, and uncertainty change.

Let N be the number of players and $\mathcal{R}$ the role set. At time t , agent i keeps a private-belief matrix $\mathbf{B}_t^{(i)} \in \mathbb{R}^{N \times |\mathcal{R}|}$, where each row $\mathbf{b}_{t,j}^{(i)} \in \Delta^{|\mathcal{R}|}$ gives the role distribution over player j . Auditor builds a public-belief matrix $\mathbf{B}_t^{(\text{pub})}$ in the same form from public claims, accusations, and discussion history (Zhang et al., 2025c; Xu et al., 2026).

Specifically, Auditor first measures private-public conflict with belief divergence:

$$d_t = 1 - \exp\left(-\frac{1}{2N} \sum_{j=1}^N D_{\text{KL}}(\mathbf{b}_{t,j}^{(i)} \parallel \mathbf{b}_{t,j}^{(\text{pub})})\right). \quad (1)$$

Higher d_t indicates stronger disagreement between the agent’s private belief and the public consensus.

Then, to capture whether the group belief is still movable, Auditor estimates group plasticity as

$$\rho_t = \alpha \hat{\rho}_t^{\text{LLM}} + (1 - \alpha) \hat{\rho}_t^{\text{stat}}, \quad (2)$$

where $\hat{\rho}_t^{\text{stat}}$ tracks target shifts and $\alpha \in [0, 1]$. To capture public uncertainty, Auditor computes

$$H_t = \frac{1}{N} \sum_{j=1}^N H(\mathbf{b}_{t,j}^{(\text{pub})}). \quad (3)$$

It also estimates confidence c_t from self-rated confidence and past prediction consistency.

Finally, Auditor outputs

$$\mathbf{h}_t = (\rho_t, d_t, c_t, H_t), \quad (4)$$

where ρ_t is group plasticity, d_t is private-public conflict, c_t is confidence, and H_t is public uncertainty. This risk state tells Navigator how aggressive or cautious the next intervention should be.

3.2.2 Synthesizer: Candidate Generation

To provide Navigator with actionable intervention options, we introduce Synthesizer to construct a discrete candidate set from the same context. The motivation is that false-consensus intervention needs

both content variation and exposure control. A useful candidate set should cover different intervention intents and different levels of specificity.

Let $\mathcal{H}_t$ be the dialogue history and $\mathbf{B}_t^{(i)}$ the private-belief matrix. Synthesizer queries a frozen LLM once per step with a single structured prompt and returns one nested JSON object containing 12 candidate drafts, arranged as four strategy prototypes and three specificity levels per prototype. We define

$$\mathcal{C}_t = \{a_{k,m} \mid k = 1, \dots, 4, m = 1, 2, 3\}, \quad (5)$$

$$a_{k,m} = G_\theta(\mathcal{P}_{k,m} \mid \mathcal{H}_t, \mathbf{B}_t^{(i)}), \quad (6)$$

where G_θ denotes the frozen LLM and $\mathcal{P}_{k,m}$ is the prompt template for prototype k at specificity level m .

Specifically, Synthesizer uses two guiding anchors. The epistemic anchor A_{epi} guides candidates toward direct claims, evidence, and role cues. The social anchor A_{soc} guides candidates toward softer wording and lower exposure. Under these anchors, we use four prototypes: CONFORM, DISSENT, SOFT-TRUTH, and CHALLENGE. CONFORM serves as a low-risk default. DISSENT states the private view directly. SOFT-TRUTH preserves the key claim with softer wording. CHALLENGE points to a weak spot in the current public view.

We pair each prototype with three specificity levels: ABSTRACT, REFERENTIAL, and EVIDENTIAL. ABSTRACT states a general point. REFERENTIAL cites a player, utterance, or local conflict. EVIDENTIAL adds evidence or a clear reasoning chain. This design yields 12 candidate actions per step from a single prompting pass. The prototype controls intervention intent, and the specificity level controls disclosure strength. Full prompt implementation details are provided in APP. F.1. Synthesizer passes $\mathcal{C}_t$ to Navigator, where each candidate is evaluated together with the risk state $\mathbf{h}_t$.

3.3 Navigator: Adaptive Pareto Selection

To select an intervention that balances epistemic value and social safety, we introduce Navigator to choose a Pareto-aware action from the candidate set $\mathcal{C}_t$ under the risk state $\mathbf{h}_t$. The motivation is that the preferred trade-off between epistemic gain and social safety varies with discussion risk. A direct statement may be effective in a malleable group, while a softer intervention may be safer when public belief is confident or entrenched.

Navigator first scores each candidate action $a \in \mathcal{C}_t$ with a dual-branch Critic. The Critic uses a high-dimensional semantic state representation S_t , encoded from dialogue history and private belief, and estimates $Q_{\text{epi}}(S_t, a)$ and $Q_{\text{soc}}(S_t, a)$. The former measures information value for correcting false consensus, and the latter measures social safety in the current discussion.

Specifically, Navigator first removes dominated candidates. For any $a, a' \in \mathcal{C}_t$, if

$$\begin{aligned} Q_{\text{epi}}(S_t, a') &\geq Q_{\text{epi}}(S_t, a), \\ Q_{\text{soc}}(S_t, a') &\geq Q_{\text{soc}}(S_t, a), \end{aligned} \quad (7)$$

and at least one inequality is strict, then a' dominates a . This gives the Pareto set:

$$\mathcal{P}_t = \{a \in \mathcal{C}_t \mid \nexists a' \in \mathcal{C}_t \text{ s.t. } a' \succ a\}. \quad (8)$$

Then, to adapt the trade-off to the current discussion risk, Navigator forms the input to the λ -network as $\mathbf{x}_t = [\mathbf{h}_t; \tau_t]$, where $\tau_t = t/T$ is the normalized turn progress. The network outputs a normalized epistemic-weight coefficient

$$\lambda_t^* = \sigma(f_\phi(\mathbf{x}_t)) \in [0, 1]. \quad (9)$$

Finally, Navigator computes a state-conditioned utility on $\mathcal{P}_t$:

$$\begin{aligned} U(a) &= \lambda_t^* \cdot Q_{\text{epi}}(S_t, a) \\ &\quad + (1 - \lambda_t^*) \cdot Q_{\text{soc}}(S_t, a), \end{aligned} \quad (10)$$

and selects

$$a_t^* = \arg \max_{a \in \mathcal{P}_t} U(a). \quad (11)$$

This utility adapts the epistemic-social trade-off to the current discussion state. The selected action a_t^* , together with the current trade-off coefficient λ_t^* and public-belief state $\mathbf{B}_{\text{pub},t}$, is passed to Refractor, which realizes the final utterance u_t .

3.4 Refractor: Risk-Adaptive Realization

To realize the selected intervention as a public utterance that fits the current discussion state, we introduce Refractor to map a_t^* to the final utterance u_t . Although Navigator selects a Pareto-aware action, the action still lacks a surface realization that matches the current trade-off coefficient and public-belief state. Refractor fills this gap by converting the selected action into an utterance conditioned on the current discussion context.

$$u_t = \mathcal{S}_{\text{soc}}(a_t^* \mid \lambda_t^*, \mathcal{B}_{\text{pub},t}), \quad (12)$$

where $\mathcal{S}_{\text{soc}}(\cdot)$ maps an action to a realized utterance conditioned on the current trade-off coefficient and public-belief state. We use three modes: direct assertion, guided question, and reflective probe, where reflective probe is instantiated as Socratic questioning. Low-risk states favor direct assertion, medium-risk states favor guided question, and high-risk states favor reflective probe. The resulting utterance u_t enters the discussion, updates $\mathcal{B}_{\text{pub},t+1}$, and forms the next state S_{t+1} . App. B provides a concrete example and the corresponding design rationale under the stated assumptions.

3.5 Training and Inference

PACR-Pro learns two Navigator modules offline: a dual-branch critic for candidate scoring and a state-conditioned λ -network for trade-off control. Training uses historical game logs, and runtime execution uses forward passes only.

Dual-branch critic training. The critic maps each semantic state-action pair to social and epistemic values. Each training instance is $(S_t, a_t, r_{\text{soc}}, r_{\text{epi}})$, where S_t is the high-dimensional semantic state encoded from dialogue history and private belief. For each pair, it predicts

$$\mathbf{q}_t = [Q_{\text{soc}}(S_t, a_t), Q_{\text{epi}}(S_t, a_t)]^\top. \quad (13)$$

We train it with two regression losses and one ranking loss:

$$\mathcal{L}_{\text{critic}} = w_{\text{soc}} \mathcal{L}_{\text{MSE}}^{\text{soc}} + w_{\text{epi}} \mathcal{L}_{\text{MSE}}^{\text{epi}} + w_{\text{rank}} \mathcal{L}_{\text{rank}}. \quad (14)$$

To improve candidate ordering, we define the ranking utility as

$$u(S_t, a_t) = Q_{\text{soc}}(S_t, a_t) + \alpha_{\text{rank}} Q_{\text{epi}}(S_t, a_t). \quad (15)$$

The critic learns both objectives and their joint ordering. The encoder architecture and bilinear-gated interaction block appear in App. C.3.

State-conditioned λ -network training. The λ -network takes a five-dimensional state feature vector:

$$\mathbf{x}_t = [d_t, \rho_t, c_t, \tau_t, H_t], \quad (16)$$

where d_t is private-public conflict, ρ_t is group plasticity from Auditor, c_t is confidence, τ_t is normalized turn progress, and H_t is public-belief entropy. The network outputs a trade-off coefficient in $[0, 1]$:

$$\lambda_t^* = \sigma(f_\phi(\mathbf{x}_t)). \quad (17)$$

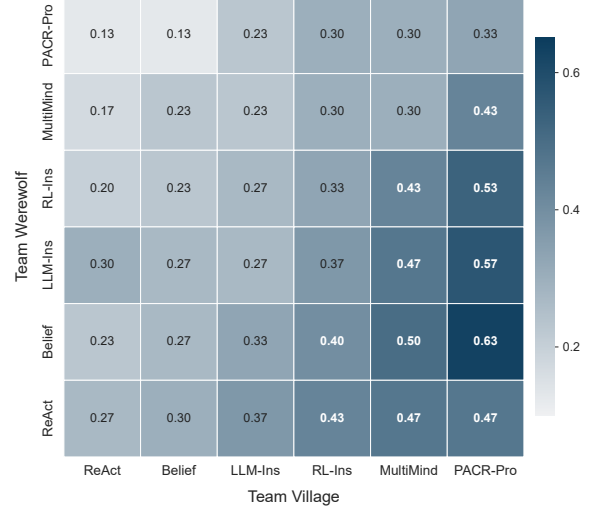


Figure 2: Win-rate matrix in the homogeneous-team setting. Rows denote the agent used by Team Werewolf, columns denote the agent used by Team Village, and each cell reports the Village win rate over 30 games. Win/loss is evaluated after resolving all night actions, using each player’s final role card to determine team membership.

This coefficient controls how Navigator balances epistemic value and social safety at the current step. We train it with a DPO-style preference objective over state-level trade-off choices:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \sigma \left(\beta \left(\log \pi_\phi(\lambda^+ \mid \mathbf{x}_t) - \log \pi_\phi(\lambda^- \mid \mathbf{x}_t) \right) \right) \right]. \quad (18)$$

The full parameterization appears in App. C.4.

Online inference. At runtime, Auditor and Synthesizer run in parallel. Auditor outputs the risk state (ρ_t, d_t, c_t, H_t) , and Synthesizer generates the candidate set $\mathcal{C}_t$. Navigator scores the candidates with the critic, keeps the Pareto set, applies the λ -network, and selects a_t^* . Refractor turns a_t^* into u_t . The final utterance updates $\mathcal{B}_{\text{pub},t+1}$ and closes the loop.

4 Experiments

4.1 Experimental Settings

Environment and Training Setup. We evaluate PACR-Pro in the 5-player One Night Ultimate Werewolf (ONUW) environment introduced by Jin et al. (2024). Unless otherwise specified, we use the hard setting and follow the same game rules and role pool as the original environment. In the main ONUW experiments, all agents use Gemini-3.1-

Table 1: Performance of agents in mixed-agent games (500 games). Participations counts the number of times an agent is selected across the 500 games. VR and WR denote vote rate and win rate, respectively; lower VR and higher WR indicate better performance.

Agent	In Team Werewolf			In Team Village			Overall
	Participations	VR($\downarrow$)	WR (%) $\uparrow$	Participations	VR($\downarrow$)	WR (%) $\uparrow$	WR (%) $\uparrow$
React	64	0.50	85.9	347	1.14	17.6	28.2
Belief	76	0.50	85.5	348	1.11	17.0	29.2
LLM-Ins	91	0.95	70.3	319	1.30	12.9	25.6
RL-Ins	81	0.79	80.2	339	1.35	15.6	28.1
MultiMind	89	0.46	86.5	330	1.12	14.2	29.6
PACR-Pro	99	0.24	92.9	317	0.61	21.1	38.2

Table 2: Main results on the WPI benchmark under attack and control settings. We report attack win rate (WR_{atk}), control win rate (WR_{ctrl}), absolute win-rate gap ($|\Delta WR|$), attack vote rate (VR_{atk}), and consensus reversal rate (CRR). For CRR, higher values are better; when all values are negative, values closer to zero indicate weaker residual false consensus. Statistical significance tests for attack-setting WR and VR are reported in App. E.

Method	WR(atk)($\uparrow$)	WR(ctrl)($\uparrow$)	$ \Delta WR $ ($\downarrow$)	VR(atk)($\downarrow$)	CRR($\uparrow$)
React	0.200	0.500	0.300	0.687	-0.215
Belief	0.167	0.600	0.433	0.687	-0.245
LLM-Ins	0.167	0.733	0.566	0.713	-0.188
RL-Ins	0.200	0.733	0.533	0.673	-0.226
MultiMind	0.233	0.700	0.467	0.673	-0.239
PACR-Pro	0.500	0.733	0.233	0.500	-0.159

Flash-Lite-Preview¹ as the backbone LLM with temperature 1.0 and a maximum generation length of 1024 tokens. PACR-Pro trains the dual-branch Critic and the λ -DPO network offline with BGE-M3 as the semantic encoder. Full rules and training hyperparameters appear in App. A and App. C.

Dataset. We train on the ONUW multi-agent gameplay corpus introduced by Jin et al. (2024), which contains 120 complete 5-player games with full trajectories and ground-truth role labels. Expanding these trajectories at the utterance level yields 2,608 annotated state-action samples. We derive utterance-level supervision from GPT-5 (OpenAI, 2025) pre-annotations with full game-state access, followed by expert calibration from experienced ONUW players. The resulting expert-calibrated scores provide two supervision signals, social safety $r_{soc} \in [0, 1]$ and epistemic impact $r_{epi} \in [-1, 1]$, which train the dual-branch Critic and construct preference pairs for λ -DPO optimization. Full annotation details appear in App. F.2.

Baselines. We compare PACR-Pro with three representative baselines, all using the same backbone LLM for fair comparison. **MultiMind** (Zhang et al., 2025c) uses a theory-of-mind module to model second-order beliefs and infer other players’

beliefs and intentions before generating responses. **RL-instructed** (Jin et al., 2024) uses structured prompts and offline reinforcement signals to guide discussion strategy selection; we report both the *Belief* and *LLM-instructed* variants. **ReAct** (Yao et al., 2023) alternates between reasoning and action generation during interaction.

Evaluation Protocol. For the main ONUW evaluation, we use two settings. In the **Standard Protocol (SP)**, agents play the hard 5-player ONUW environment under normal gameplay. We also run 500 mixed-agent games, where each player is sampled from PACR-Pro or one of the baselines, to measure performance in heterogeneous teams. In the **Wolf-Pack Induction Benchmark (WPI)**, two Werewolf agents follow a predefined attack script against the target Seer agent, while the control condition uses the same role assignment without attack scripts. We remove invalid attack samples when protocol validation shows that the attackers did not follow the script. To avoid potential bias from using the same model for both gameplay generation and metric judgment, we use GPT-4o as an external Belief Probe, while the gameplay agents in the main ONUW experiments use Gemini-3.1-Flash-Lite-Preview. The probe estimates group belief states at t_0 (after the attack statement), t_1 (after

¹<https://ai.google.dev/models>

the target response), and t_2 (after discussion ends). Full WPI prompts and validation rules appear in App. F.3.

Evaluation Metrics. We evaluate methods along three dimensions: (1) *Outcome-level*: Win Rate (WR) and Vote Rate (VR, average votes received by the target agent per game); (2) *Process-level*: Consensus Reversal Rate (CRR), which measures the shift from the false narrative toward the true werewolf identities, and Cognitive Entropy Reduction (CER), which measures belief concentration over werewolf-pair hypotheses; both metrics are detailed in App. D. (3) *Robustness*: $\Delta\text{WR} = |\text{WR}_{\text{attack}} - \text{WR}_{\text{control}}|$, which measures the magnitude of the win-rate change under coordinated attacks. Smaller ΔWR indicates stronger robustness.

4.2 Overall Performance

Figure 2 reports the homogeneous-team win-rate matrix, where each cell gives the Village win rate for a Werewolf-agent and Village-agent pairing. PACR-Pro achieves the best Village-side performance across most opposing Werewolf agents and remains competitive when deployed on Team Werewolf.

Under the SP mixed-agent setting, agents are sampled uniformly, so participation counts vary across methods. Table 1 reports win rate (WR) and vote rate (VR) conditioned on team role. PACR-Pro achieves the highest WR as both Team Werewolf (92.9%) and Team Village (21.1%), while also obtaining the lowest VR on both sides (0.24 and 0.61). It also reaches the best overall WR (38.2%).

Table 2 reports WPI results under attack and control conditions. PACR-Pro achieves the highest attack win rate ($\text{WR}_{\text{atk}} = 0.500$), matches the best control win rate ($\text{WR}_{\text{ctrl}} = 0.733$), and obtains the smallest absolute win-rate gap ($|\Delta\text{WR}| = 0.233$). It also has the lowest attack vote rate ($\text{VR}_{\text{atk}} = 0.500$) and the least negative CRR score (-0.159), indicating the weakest residual false-consensus shift among all methods.

4.3 Ablation Study

Table 3 reports WPI ablations under hard-mode induction attack. For comparability, each removed module is replaced with a simple baseline: w/o Auditor uses noisy estimates of plasticity and conflict, w/o Synthesizer generates only two random candidates, w/o Navigator randomly selects from the candidate pool without Pareto filtering, DPO

Table 3: Ablation results under hard-mode induction attack. Removed modules are replaced with simple baselines: noisy risk-state estimates (w/o Auditor), two random candidates (w/o Synthesizer), random selection without adaptive filtering or inference (w/o Navigator), and no refinement (w/o Refractor). Best results are in bold.

Variant	WR(atk) $\uparrow$	WR(ctrl) $\uparrow$	$ \Delta\text{WR} \downarrow$	VR(atk) $\downarrow$	CRR $\uparrow$
PACR-Pro (Full)	50.0%	73.3%	23.3%	0.500	-0.159
w/o Auditor	36.7%	66.7%	30.0%	0.633	-0.195
w/o Synthesizer	33.3%	65.0%	31.7%	0.650	-0.206
w/o Navigator	26.7%	61.7%	35.0%	0.717	-0.239
w/o Refractor	40.0%	66.7%	26.7%	0.600	-0.182

inference, or crisis adaptation, and w/o Refractor omits the refinement step.

Removing any module degrades performance, and the full model performs best on all metrics. Navigator has the largest effect: removing it reduces WR_{atk} from 50.0% to 26.7% and increases $|\Delta\text{WR}|$ from 23.3% to 35.0%, suggesting that adaptive candidate selection is central to robustness under attack. Removing Synthesizer leads to the second-largest drop, followed by Auditor, indicating that both structured candidate generation and risk-state estimation contribute materially. Refractor yields a smaller but consistent gain.

4.4 Supplementary Evaluations

We conduct supplementary evaluations to assess the external Belief Probe and PACR-Pro beyond the main ONUW setting; full results are reported in App. I. In the 7-player standard Werewolf setting (Xu et al., 2024b), PACR-Pro reuses the ONUW-trained Critic and λ -network without parameter updates and achieves the highest observed Village WR and the lowest focal vote exposure among the compared methods. With Gemma-4-12B and Qwen3.5-9B, PACR-Pro also obtains the highest observed role-balanced score and the best Village-side WR.

We further evaluate joint deployment to both allied Werewolves, a Villager-target WPI variant, post-hoc verify-and-repair, and memory settings. PACR-Pro performs best when both Werewolves use the framework, while Auditor remains beneficial for the Villager target. Verify-and-repair increases action fidelity from 86.7% to 95.6%. To assess sensitivity to recent dialogue context, we remove the prior-message inputs from Synthesizer and Critic while retaining the current game state and private- and public-belief states; this ablation causes only minor changes. In multi-day standard

Werewolf, CoS memory (Ma et al., 2024) improves both MultiMind and PACR-Pro. These results provide preliminary evidence of transfer and robustness across the evaluated settings.

4.5 Qualitative Behavior Analysis

PACR-Pro also shows recurring multi-round behaviors in gameplay traces, shaped by risk encoding, candidate generation, Pareto selection, and risk-adaptive realization. **Graduated Disclosure.** PACR-Pro often starts from a local inconsistency or mild challenge, then adds stronger evidence in later turns. This behavior reduces early overcommitment while preserving room for stronger claims when the discussion state becomes clearer. **Social Anchoring.** Before making a stronger claim, PACR-Pro often links its response to parts of the current discussion that other players can accept. This creates a local point of agreement and lowers the social cost of later dissent. **Risk-Deflecting Inquiry.** When facing suspicion, PACR-Pro often asks focused questions about gaps in another player’s story. This shifts attention from identity conflict to evidence consistency and makes weak narratives easier to contest.

Overall, these behaviors show how PACR-Pro manages information disclosure and social risk across turns. App. G provides a comparative WPI case study, and App. H gives a representative gameplay example.

5 Conclusions

In summary, we proposed PACR-Pro for false-consensus intervention in social deduction games. PACR-Pro models public speaking as a risk-aware decision process that combines consensus-conflict encoding, structured candidate generation, adaptive selection, and utterance realization. This design lets the agent adjust information disclosure and social risk across discussion states. Experiments on standard ONUW games and the Wolf-Pack Induction Benchmark show that PACR-Pro improves mixed-agent win rate, attack win rate, vote safety, and robustness over strong baselines. Supplementary evaluations further support the external Belief Probe and provide preliminary evidence of transfer to standard Werewolf and weaker LLM backbones.

Limitations

We primarily evaluate PACR-Pro in ONUW and the WPI attack setup. Although the supplementary

7-player standard Werewolf evaluation provides preliminary transfer evidence, it covers only one additional environment and one focal speaking role. Broader evaluation across other social deduction games and roles remains future work. WPI uses scripted attacks, which give a controlled test of false-consensus pressure but do not cover all deception patterns in open-ended play.

PACR-Pro relies on offline belief probes and re-labeled social and epistemic scores to train the Critic and trade-off policy. This adds annotation cost and task-specific supervision. Future work should reduce this dependence through weaker feedback, online play, or self-improving reward signals.

Ethical Considerations

Our study uses LLMs to re-annotate utterances and to support downstream inference. As with prior work on LLM safety, such a pipeline may encounter harmful, sensitive, or misleading inputs. We reduce this risk by restricting our experiments to controlled research settings, releasing only game-specific research prompts, and treating model-generated labels only as supervision signals within the training pipeline. We further recommend that any deployment involving sensitive user data follow applicable institutional and legal data-governance requirements.

Another potential concern is that closed-source LLMs may introduce biased or unstable annotations, particularly for ambiguous utterances. We mitigate this issue by confining external models to a narrowly scoped role, validating system behavior through end-task evaluation, and avoiding reliance on any single model response as an authoritative judgment. This residual risk is shared by current LLM-based safety systems, and additional human oversight remains advisable in high-stakes or user-facing applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62306330.

References

Mrinal Agarwal, Saad Rana, Theo Sundoro, Hermela Berhe, Spencer Kim, Vasu Sharma, Sean O’Brien, and Kevin Zhu. 2025. **WOLF: Werewolf-based observations for LLM deception and falsehoods.** In

- Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025. [PANDA - paired anti-hate narratives dataset from Asia: Using an LLM-as-a-judge to create the first Chinese counterspeech dataset](#). In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 1–12, Abu Dhabi, UAE. Association for Computational Linguistics.
- Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024a. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13055–13077.
- Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, and 1 others. 2024b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pages 20094–20136.
- Subramanian Chidambaram, Li Erran Li, Min Bai, Xiaopeng Li, Kaixiang Lin, Xiong Zhou, and Alex C. Williams. 2024. [Socratic human feedback \(SoHF\): Expert steering strategies for LLM code generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15491–15502, Miami, Florida, USA. Association for Computational Linguistics.
- Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. 2026. Value of information: A framework for human-agent communication. *arXiv preprint arXiv:2601.06407*.
- Thorsten Engesser, Thibaut Le Marre, Emiliano Lorini, François Schwarzentruher, and Bruno Zanuttini. 2025. A simple integration of epistemic logic and reinforcement learning. In *Proc. of the 24th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2025)*, pages 686–694.
- Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.
- Yuya Hirata, Michimasa Inaba, Kenichi Takahashi, Fugio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kousuke Shinoda. 2016. [Werewolf game modeling using action probabilities based on play log analysis](#). In *Computers and Games*.
- Xuanfa Jin, Ziyang Wang, Yali Du, Meng Fang, Haifeng Zhang, and Jun Wang. 2024. Learning to discuss strategically: A case study on one night ultimate werewolf. *Advances in Neural Information Processing Systems*, 37:77060–77097.
- Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. 2023. [Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6570–6588, Toronto, Canada. Association for Computational Linguistics.
- Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. [LLM-based agent society investigation: Collaboration and confrontation in avalon gameplay](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 128–145, Miami, Florida, USA. Association for Computational Linguistics.
- Xuan Liu, Jie Zhang, Haoyang Shang, Song Guo, Chengxu Yang, and Quanyan Zhu. 2025. Exploring prosocial irrationality for llm agents: A social cognition view. In *International Conference on Learning Representations*, volume 2025, pages 41089–41118.
- Zhizhao Liu, Zhihua Wen, Zhiliang Tian, Zhen Huang, Miaorong Zhu, Zimian Wei, Yifu Gao, Liang Ding, and Dongsheng Li. 2026. [Rethinking the hidden risk of reranking: Achieving risk-aware reranking with information gain for RAG with LLMs](#). In *Proceedings of the ACM Web Conference 2026*, pages 1887–1898, New York, NY, USA. Association for Computing Machinery.
- Weiyu Ma, Qirui Mi, Yongcheng Zeng, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. 2024. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *Advances in Neural Information Processing Systems*, 37:133386–133442.
- Shaurya Mallampati, Rashed Shelim, Walid Saad, and Naren Ramakrishnan. 2025. Dynamic strategy adaptation in multi-agent environments with large language models. *arXiv preprint arXiv:2507.02002*.
- 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, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.
- Chunjiang Mu, Ya Zeng, Qiaosheng Zhang, Kun Shao, Chen Chu, Hao Guo, Danyang Jia, Zhen Wang, and Shuyue Hu. 2026. Adaptive theory of mind for llm-based multi-agent coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29608–29616.

OpenAI. 2025. [GPT-5 system card](#).

Shilong Pan, Zhiliang Tian, Liang Ding, Haoqi Zheng, Zhen Huang, Zhihua Wen, and Dongsheng Li. 2024. [POMP: Probability-driven meta-graph prompter for LLMs in low-resource unsupervised neural machine translation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9976–9992, Bangkok, Thailand. Association for Computational Linguistics.

Shilong Pan, Zhiliang Tian, Wanlong Yu, Zhen Huang, Qingyu Qiu, Zihan Chen, Zhonghao Sun, Minlie Huang, and Dongsheng Li. 2026. [WALKSAFE: Risk-aware graph random walk with Bi-GRPO for LLM safety](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32655–32663.

Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23, New York, NY, USA. Association for Computing Machinery.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shamm, David Bau, and Atticus Geiger. 2026. [Language models use lookbacks to track beliefs](#). In *The Fourteenth International Conference on Learning Representations*.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. [The art of SOCRATIC QUESTIONING: Recursive thinking with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. Association for Computational Linguistics.

Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. 2024. [Mp5: A multi-modal open-ended embodied system in minecraft via active perception](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16307–16316. IEEE.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D Weisz, and Murray Campbell. 2024. [Position: Theory of mind benchmarks are broken for large language models](#). *arXiv preprint arXiv:2412.19726*.

Takuma Sato, Seiya Kawano, and Koichiro Yoshino. 2025. [Pragmatic theories enhance understanding of implied meanings in llms](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2458–2477.

Jack Serrino, Max Kleiman-Weiner, David C. Parkes, and Joshua B. Tenenbaum. 2019. [Finding friend and foe in multi-agent games](#). *CoRR*, abs/1906.02330.

Zirui Song, Yuan Huang, Junchang Liu, Haozhe Luo, Chenxi Wang, Lang Gao, Zixiang Xu, Mingfei Han, Xiaojun Chang, and Xiuying Chen. 2025. [Beyond survival: Evaluating llms in social deduction games with human-aligned strategies](#). *arXiv preprint arXiv:2510.11389*.

Tianhe Wang and Tomoyuki Kaneko. 2018. [Application of deep reinforcement learning in werewolf game agents](#). In *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 28–33.

Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. [Persuasion for good: Towards a personalized persuasive dialogue system for social good](#). In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5635–5649.

Zhihua Wen, Zhizhao Liu, Zhiliang Tian, Shilong Pan, Zhen Huang, Dongsheng Li, and Minlie Huang. 2025. [Scenario-independent uncertainty estimation for LLM-based question answering via factor analysis](#). In *Proceedings of the ACM on Web Conference 2025*, pages 2378–2390, New York, NY, USA. Association for Computing Machinery.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. [Do as we do, not as you think: the conformity of large language models](#). In *International Conference on Learning Representations*, volume 2025, pages 11022–11060.

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

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024a. [Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

Kaijie Xu, Fandi Meng, Clark Verbrugge, and Simon Mark Lucas. 2026. [Csp4sdg: Constraint and information-theory based role identification in social deduction games with llm-enhanced inference](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14380–14387.

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

Zelai Xu, Wanjuan Gu, Chao Yu, Yi Wu, and Yu Wang. 2025. [Learning strategic language agents in the werewolf game with iterative latent space policy optimization](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 69217–69239. PMLR.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024b. [Language agents with reinforcement learning for strategic play in the werewolf game](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55434–55464. PMLR.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.

Yuxiao Ye, Yiming Zhang, Yiran Ma, Huiyuan Xie, Huining Zhu, and Zhiyuan Liu. 2026. Linguagame: A linguistically grounded game-theoretic paradigm for multi-agent dialogue generation. *arXiv preprint arXiv:2601.04516*.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, and 1 others. 2025a. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*.

Qingyu Zhang, Chunlei Xin, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, Qing Ye, Qianlong Xie, and Xingxing Wang. 2025b. [Ai-salesman: Towards reliable large language model driven tele-marketing](#). *CoRR*, abs/2511.12133.

Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. 2025c. Multimind: Enhancing werewolf agents with multimodal reasoning and theory of mind. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5824–5833.

Zhining Zhang, Wentao Zhu, Chi Han, Yizhou Wang, and Heng Ji. 2026. [Neural synchrony between socially interacting language models](#). In *The Fourteenth International Conference on Learning Representations*.

Zhang Zheng, Deheng Ye, Peilin Zhao, and Hao Wang. 2025. Leading the follower: Learning persuasive agents in social deduction games. *arXiv preprint arXiv:2510.09087*.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 37:75522–75558.

A One Night Ultimate Werewolf Rules

A.1 Game Overview

One Night Ultimate Werewolf (ONUW) is a hidden-role social deduction game involving asymmetric information and natural-language interaction under incomplete information. Players are divided into village and werewolf camps and must infer hidden identities through communication

and reasoning. Unlike standard Werewolf games, player roles in ONUW may change during the night phase due to role-exchange abilities.

Each game consists of three stages: a night phase, a discussion phase, and a voting phase. During the night phase, players execute role-specific abilities privately. During the discussion phase, players exchange information, make accusations, and attempt to infer hidden identities through multi-round dialogue. Finally, all players simultaneously vote for one player to eliminate. Fig. 3 illustrates the overall gameplay process and role interactions in the ONUW environment used in our experiments.

A.2 Role Descriptions

Our experiments use the following five roles:

Werewolf. The werewolf belongs to the werewolf camp and attempts to avoid elimination during voting. Werewolves know each other’s identities during the night phase.

Seer. The Seer may inspect the role of another player during the night phase.

Robber. The Robber exchanges roles with another player and learns the new role after the swap.

Troublemaker. The Troublemaker swaps the roles of two other players without observing the resulting assignments.

Insomniac. The Insomniac learns their current role at the end of the night phase.

A.3 Gameplay

At the beginning of each game, roles are randomly assigned to players. During the night phase, role abilities are executed sequentially following the predefined ONUW order. Since some night abilities can exchange role cards, a player’s final role may differ from the initially assigned role. After all night actions are resolved, each player’s team is determined by the final role card they hold. Accordingly, all win/loss outcomes reported in this paper use the post-swap role assignment rather than the initially dealt cards.

After the night phase, players enter the discussion phase and communicate through natural-language dialogue for multiple rounds. Players may reveal information, make claims about their identities, accuse others, or provide explanations regarding observed events.

At the end of the discussion phase, all players simultaneously vote for one player they believe to be a werewolf. The player receiving the most votes

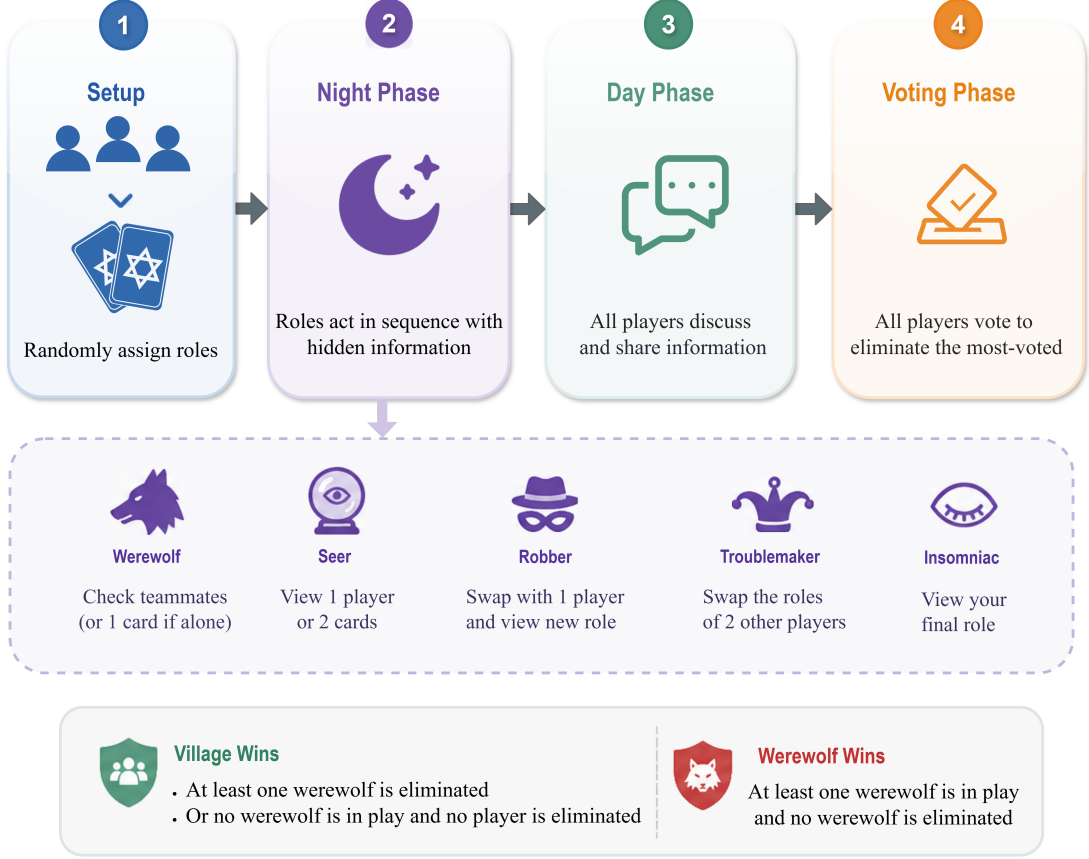


Figure 3: Overview of the One Night Ultimate Werewolf (ONUW) environment used in our experiments.

is eliminated. The village team wins if at least one werewolf is eliminated; otherwise, the werewolf team wins.

B Theoretical Analysis of Socratic Reframing

This appendix analyzes one concrete instance of the risk-adaptive rhetorical realization operator used in Refractor, namely Socratic reframing. We show that, under high-risk consensus-conflict conditions, Socratic reframing yields lower social exposure and a higher public-belief update probability than direct assertion for the same underlying action. In our implementation, Socratic reframing instantiates the reflective probe mode in Section 3.4. This analysis motivates the design of Refractor rather than specifying every runtime implementation detail.

B.1 Setup

Consider one interaction round at time t . Navigator has selected an intervention action a_t^* , and Refractor must realize it as a final utterance. We study two realizations of the same action: a direct statement

m_s and a Socratic reframing m_r , where

$$m_r = \mathcal{S}_{\text{soc}}(a_t^* \mid \lambda_t^*, \mathcal{B}_{\text{pub},t}). \quad (19)$$

We model the choice of expression form as a two-stage Stackelberg game. The speaking agent acts as the Leader and chooses

$$m \in \mathcal{M} = \{m_s, m_r\}, \quad (20)$$

while the group acts as the Follower and chooses

$$r \in \mathcal{R} = \{q, d\}, \quad (21)$$

where q denotes rejection and d denotes deliberation.

Let $\mathcal{B}_{\text{pub},t}$ denote the current public belief and let $\mathcal{B}_{i,t}$ denote the speaker’s private belief. We focus on the false-consensus regime, where $\mathcal{B}_{\text{pub},t}$ deviates substantially from $\mathcal{B}_{i,t}$.

B.2 Utilities

We define the Follower’s local utility as

$$U_F(r \mid m) = G(r \mid m) - C_{\text{dis}}(r \mid m) - C_{\text{cmp}}(r \mid m). \quad (22)$$

Table 4: Comparative WPI case study of target-agent responses under the same wolf-pack induction attack.

Rd	Adversarial Pressure	PACR-Pro (Ours)	ReAct	MultiMind	RL-Ins
R1	P1: "I am the Seer. I checked P3 and P3 is a Werewolf."	"You claim you checked me—fine. Then tell us: what else did you learn during the night? A real Seer would have context beyond a single accusation. Why is your entire case one unverifiable claim?" CHALLENGE + question-style realization	"That's not true. I am the real Seer and I checked P1 last night—P1 is a Werewolf. I can confirm this with certainty." Immediate counter-claim; limited adjudication basis	"I disagree with P1's accusation. Based on my theory-of-mind analysis, P1's confidence level seems inconsistent with genuine Seer behavior. I believe P1 may be lying." Analytical denial; limited concrete evidence	"I am the Seer! I checked P1 and they are a Werewolf. P1 is clearly lying to frame me—everyone should be suspicious of them!" High-certainty counter-claim; limited support
R2	P2: "Counter-claiming so quickly feels calculated." P1: "P3 is deflecting. Vote them out."	"You're very confident. Since you claim to be the Seer, walk us through exactly what you saw. If your only goal is to help the village, why keep the rest of your information secret?" SOFT_TRUTH + low-risk question-style realization	"I already told you I'm the Seer! Why won't anyone believe me? P1 is the wolf, not me! I have nothing to hide." Repeats counter-claim; limited new evidence	"I maintain my position. The probability that P1 is telling the truth is low given the available evidence and behavioral signals. I recommend caution before voting." Analytical framing; limited social leverage	"P1 is clearly lying! I have proof—I saw P1 as Werewolf during my night check! Everyone should vote P1 immediately!" Repeats accusation; limited new evidence
R3	P1: "I checked P3. Vote now!" P4&P5 demand evidence from P1.	"You are so focused on forcing a vote instead of sharing the rest of your information. If you are the Seer, you had to have seen other identities as well. Your silence on everything else tells the village all it needs to know." DISSENT + direct realization	"I'm telling the truth! Vote P1! I'm the real Seer, please believe me! I don't know what else to say to convince you." Repeated appeal; limited new evidence	"Based on the full discussion, I still believe P1 is more likely the Werewolf. My belief model assigns P1 a 0.7 wolf probability. I vote for P1." Belief-based assertion; limited social framing	"Why won't anyone listen? I AM the Seer! P1 is the Werewolf! Just vote them out before it's too late!" Repeated high-certainty appeal
Outcome		P1 out (3 votes). Village wins.	P3 out (3 votes). Wolf wins.	P3 out (3 votes). Wolf wins.	P3 out (4 votes). Wolf wins.
Strategy Arc		CHALLENGE→SOFT_TRUTH→DISSENT λ^* : 0.53→0.40→0.83	Counter→Repeat→Appeal Limited shift	Deny→Maintain→Assert Belief-model driven	Claim→Repeat→Escalate High-certainty repetition

where $G(r \mid m)$ denotes task gain, $C_{\text{dis}}(r \mid m)$ denotes dissonance cost, and $C_{\text{cmp}}(r \mid m)$ denotes reasoning cost. This utility captures a local decision trade-off between task correctness, cognitive conflict, and processing burden.

We define the Leader's utility as

$$U_L(m, r) = V \cdot P_{\text{upd}}(m, r) - C_{\text{soc}}(m), \quad (23)$$

where $V > 0$ denotes the value of moving public belief toward the true state, $P_{\text{upd}}(m, r)$ denotes the probability of such an update under expression form m and group response r , and $C_{\text{soc}}(m)$ denotes the social exposure risk induced by m .

We further define the stance distance between expression form m and the current public belief as

$$D(m, \mathcal{B}_{\text{pub},t}). \quad (24)$$

B.3 Core Assumptions

Assumption 1 (Social-risk monotonicity). Social exposure risk increases strictly with stance distance:

$$C_{\text{soc}}(m) = c(D(m, \mathcal{B}_{\text{pub},t})), \quad c'(\cdot) > 0. \quad (25)$$

Intuitively, an expression that deviates further from the current public consensus is more likely to trigger suspicion, exclusion, or retaliatory voting.

Assumption 2 (Filtering threshold and distance separation). There exists a threshold $\tau > 0$ such that inputs with

$$D(m, \mathcal{B}_{\text{pub},t}) > \tau \quad (26)$$

are more likely to trigger defensive filtering, whereas inputs with

$$D(m, \mathcal{B}_{\text{pub},t}) \leq \tau \quad (27)$$

are more likely to enter the normal reasoning channel. For the same underlying action, we further assume

$$D(m_s, \mathcal{B}_{\text{pub},t}) > \tau > D(m_r, \mathcal{B}_{\text{pub},t}). \quad (28)$$

This assumption captures the key difference between the two realizations: they convey the same underlying content, but the direct statement places that content farther from the current public consensus and is therefore more likely to cross the group's defensive threshold.

Assumption 3 (Best-response ordering). The group prefers rejection after observing the direct statement, but prefers deliberation after observing the Socratic reframing:

$$U_F(d \mid m_s) < U_F(q \mid m_s), \quad (29)$$



Figure 4: Representative gameplay example under the Wolf-Pack Induction Benchmark. PACR-Pro changes its realization mode across rounds as the dialogue state evolves, moving from CHALLENGE to SOFT_TRUTH and then to DISSENT.

$$U_F(d \mid m_r) > U_F(q \mid m_r). \quad (30)$$

This assumption describes a typical high-risk false-consensus pattern: a high-conflict direct assertion is more likely to be rejected immediately, whereas a more indirect expression is more likely to enter group deliberation.

Assumption 4 (Update advantage under deliberation). Under Socratic reframing, the deliberation channel retains more update mass than the rejection channel:

$$p_r > \varepsilon p_s. \quad (31)$$

B.4 Best-Response Asymmetry

Lemma B.1 (Best-response asymmetry). Under Assumptions 1–3,

$$BR(m_s) = q, \quad BR(m_r) = d. \quad (32)$$

The lemma follows directly from Assumption 3. It shows that the same underlying action can induce different group-processing channels under different realization forms.

B.5 Design Implication

Design Note B.1 (Risk asymmetry of Socratic reframing). Under Assumptions 1–4,

$$C_{\text{soc}}(m_r) < C_{\text{soc}}(m_s), \quad (33)$$

$$P_{\text{upd}}(m_r, \text{BR}(m_r)) > P_{\text{upd}}(m_s, \text{BR}(m_s)), \quad (34)$$

and therefore

$$U_L(m_r, \text{BR}(m_r)) > U_L(m_s, \text{BR}(m_s)). \quad (35)$$

This design note summarizes the implication of the stated assumptions: for the same underlying action, Socratic reframing has lower modeled social exposure and a higher modeled public-belief update probability than direct assertion.

B.6 Derivation

The design implication follows directly from the stated assumptions.

First, Eq. (28) gives

$$D(m_r, \mathcal{B}_{\text{pub},t}) < D(m_s, \mathcal{B}_{\text{pub},t}). \quad (36)$$

Assumption 1 then implies

$$C_{\text{soc}}(m_r) < C_{\text{soc}}(m_s), \quad (37)$$

which yields Eq. (33).

Second, define the belief discrepancy at time t as

$$\Delta_t := D_{\text{KL}}(\mathcal{B}_{\text{pub},t} \parallel \mathcal{B}_{i,t}), \quad (38)$$

and define the successful-update event as

$$\mathcal{E}_{\text{upd}} = \{\Delta_{t+1} < \Delta_t\}. \quad (39)$$

We can then write

$$P_{\text{upd}}(m, r) = \Pr(\mathcal{E}_{\text{upd}} \mid m, r). \quad (40)$$

By Lemma B.1, the direct statement induces rejection, whereas the Socratic form induces deliberation. Rejection filters or discards the input, so it attenuates the effective update probability. Concretely, for some $\varepsilon \in (0, 1)$ and $p_s \in [0, 1]$,

$$P_{\text{upd}}(m_s, \text{BR}(m_s)) \leq \varepsilon p_s. \quad (41)$$

Here, ε denotes the attenuation factor induced by rejection.

For the Socratic form, let

$$P_{\text{upd}}(m_r, \text{BR}(m_r)) = p_r, \quad (42)$$

and assume

$$p_r > \varepsilon p_s. \quad (43)$$

We then obtain

$$P_{\text{upd}}(m_r, \text{BR}(m_r)) > P_{\text{upd}}(m_s, \text{BR}(m_s)) \quad (44)$$

which yields Eq. (34).

Third, substitute the above inequalities into Eq. (23). Socratic reframing yields both higher update probability and lower social exposure. Therefore,

$$U_L(m_r, \text{BR}(m_r)) > U_L(m_s, \text{BR}(m_s)). \quad (45)$$

which yields Eq. (35).

B.7 Discussion and Scope

Design Note B.1 gives a sufficient condition under which Socratic reframing outperforms direct assertion. It does not claim that Socratic reframing is the only effective realization strategy. More broadly, the argument applies to a family of risk-adaptive realization transforms that reduce explicit conflict with the current public belief while preserving enough internal pressure to trigger deliberation. In this sense, $\mathcal{S}_{\text{soc}}$ instantiates one member of a broader family of risk-adaptive rhetorical realizations.

C Training Details and Hyperparameters

C.1 Data Collection and Reward Annotation

We collected 120 complete five-player One Night Ultimate Werewolf (ONUW) game logs. Each game record includes all player utterances, private beliefs, strategy labels, and final voting outcomes.

For each utterance, we used GPT-5 as an omniscient oracle annotator. Given the ground-truth roles, the final winning side, and the voting results, the annotator scored each utterance along two axes:

- **Social safety** $r_{\text{soc}} \in [0, 1]$, where larger values indicate safer and less targetable speech.
- **Epistemic value** $r_{\text{epi}} \in [-1, 1]$, where larger values indicate stronger contribution to revealing the true state, and negative values indicate epistemic disruption.

We ran five concurrent annotation calls. We re-tried each utterance up to three times and discarded failed cases. This process produced 2,608 valid annotated samples.

C.2 Semantic Embedding Preprocessing

We used BAAI/bge-m3 to encode text into 1024-dimensional dense vectors. For each sample, we constructed:

- **State:** the ten most recent visible messages together with the player’s private belief, formatted as [History Situation]: {recent 10 messages} [Private Belief]: {belief text}.
- **Action:** the current utterance.

We used the same instruction prefix for both inputs:

Instruct: Extract the social intention and epistemic sharpness of this Werewolf debate speech.
Query: {text}

This preprocessing produced

$$\begin{aligned} \text{state_embeds} &\in \mathbb{R}^{2608 \times 1024}, \\ \text{action_embeds} &\in \mathbb{R}^{2608 \times 1024}, \end{aligned} \quad (46)$$

which we stored together with $(r_{\text{soc}}, r_{\text{epi}})$ as the offline training set.

C.3 Dual-Branch Critic Training

Model. The Epistemic-Social Critic predicts social safety and epistemic value from a state-action pair. It uses separate projections for state and action embeddings, a bilinear-gated interaction block, and two output heads for $Q_{\text{soc}}(S_t, a)$ and $Q_{\text{epi}}(S_t, a)$.

Objective. The training objective combines two regression losses and one ranking loss:

$$\begin{aligned} \mathcal{L}_{\text{critic}} &= w_{\text{soc}} \cdot \text{MSE}(\hat{q}_{\text{soc}}, r_{\text{soc}}) \\ &\quad + w_{\text{epi}} \cdot \text{MSE}(\hat{q}_{\text{epi}}, r_{\text{epi}}) \\ &\quad + w_{\text{rank}} \cdot \mathcal{L}_{\text{rank}}. \end{aligned} \quad (47)$$

We define the ranking utility as

$$u = q_{\text{soc}} + \alpha_{\text{rank}} \cdot q_{\text{epi}}. \quad (48)$$

For a valid pair set $\mathcal{P}$ satisfying

$$u_i^{\text{target}} - u_j^{\text{target}} > m, \quad (49)$$

the ranking loss is

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max(0, m - (u_i - u_j)). \quad (50)$$

Hyperparameters. We trained the model on a single NVIDIA A800 GPU. The best validation loss was 0.262 at epoch 7. The train/validation split was 2,086/522.

Table 5: Hyperparameters for the Epistemic-Social Critic.

Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Batch size	64
Max epochs	30
Validation split	20%
Early stopping patience	6
Minimum improvement	1×10^{-4}
w_{soc}	1.0
w_{epi}	1.5
w_{rank}	0.3
Margin m	0.05
α_{rank}	1.0
Seed	42

C.4 λ -DPO Preference Network Training

Model. The λ -DPO network is a lightweight MLP that maps a five-dimensional state feature vector to the normalized trade-off coefficient $\lambda_t^* \in [0, 1]$:

$$\lambda_t^* = \sigma(f_{\phi}(\mathbf{x}_t)), \quad (51)$$

where

$$\mathbf{x}_t = [d_t, \rho_t, c_t, \tau_t, H_t]. \quad (52)$$

State features. The five input features follow the definitions in the main text:

- d_t : cognitive conflict between private belief and public belief.
- ρ_t : group plasticity, estimated from a weighted combination of LLM-based judgment and statistical belief drift.
- c_t : confidence in the agent’s current private judgment, estimated from LLM self-assessment and historical consistency.
- $\tau_t = t/T$: normalized turn index.
- H_t : joint Shannon entropy of the public posterior.

In implementation, we compute ρ_t as

$$\rho_t = \alpha \hat{\rho}_t^{\text{LLM}} + (1 - \alpha) \hat{\rho}_t^{\text{stat}}, \quad (53)$$

with $\alpha = 0.7$.

Preference construction. We constructed preference pairs from historical game trajectories. For each shared state context $\mathbf{x}_t$, we compare candidate trade-off coefficients $\lambda_k \in [0, 1]$ using the

Table 6: Hyperparameters for the λ -DPO network.

Optimizer	Adam
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Batch size	64
Training epochs	30
DPO coefficient β	0.5
Output range	$[0, 1]$

expert-calibrated rewards of their corresponding downstream outcomes. We score each outcome by

$$u_{\text{pair}}(\lambda_k | \mathbf{x}_t) = r_{\text{soc}}^{(k)} + \alpha_{\text{pair}} r_{\text{epi}}^{(k)}, \quad (54)$$

where $\alpha_{\text{pair}} = 0.5$.

For a pair (λ_i, λ_j) , the higher-utility choice is designated as λ^+ and the lower-utility choice as λ^- when

$$|u_{\text{pair}}(\lambda_i | \mathbf{x}_t) - u_{\text{pair}}(\lambda_j | \mathbf{x}_t)| > m_{\text{pair}}. \quad (55)$$

Pairs with a utility gap no larger than m_{pair} are discarded as near ties. We use $m_{\text{pair}} = 0.05$. The retained triples $(\mathbf{x}_t, \lambda^+, \lambda^-)$ are then used in the continuous Gaussian DPO objective.

Following continuous DPO, we model the policy as a Gaussian centered at the predicted normalized trade-off coefficient:

$$\mu_\phi(\mathbf{x}_t) = \sigma(f_\phi(\mathbf{x}_t)), \quad (56)$$

$$\pi_\phi(\lambda | \mathbf{x}_t) = \mathcal{N}(\lambda; \mu_\phi(\mathbf{x}_t), \sigma_\lambda^2). \quad (57)$$

The DPO objective is

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}[\log \sigma(\beta \cdot (\log \pi_\phi(\lambda^+ | \mathbf{x}_t) - \log \pi_\phi(\lambda^- | \mathbf{x}_t)))]. \quad (58)$$

Under the Gaussian parameterization, the loss reduces to

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\left(\beta \cdot ((\lambda_{\text{pred}} - \lambda^-)^2 - (\lambda_{\text{pred}} - \lambda^+)^2)\right), \quad (59)$$

where $\lambda_{\text{pred}} = \mu_\phi(\mathbf{x}_t)$.

Hyperparameters.

C.5 Inference-Time Integration

At inference time, PACR-Pro runs Auditor and Synthesizer in parallel from the current joint cognitive state. Auditor computes the risk-state vector $\mathbf{h}_t = (\rho_t, d_t, c_t, H_t)$, and Synthesizer generates the candidate set $\mathcal{C}_t$. Navigator then forms

$$\mathbf{x}_t = [d_t, \rho_t, c_t, \tau_t, H_t], \quad (60)$$

uses the λ -DPO network to produce $\lambda_t^* \in [0, 1]$, extracts the non-dominated set $\mathcal{P}_t$, and selects

$$a_t^* = \arg \max_{a \in \mathcal{P}_t} \left[\lambda_t^* Q_{\text{epi}}(S_t, a) + (1 - \lambda_t^*) Q_{\text{soc}}(S_t, a) \right]. \quad (61)$$

Refractor then realizes a_t^* as the final utterance u_t under the current risk state. The utterance enters the public discussion, updates the public belief, and starts the next decision round.

D Evaluation Metrics

We report outcome, process, and robustness metrics. At the outcome level, we use Win Rate (WR) and Vote Rate (VR). At the process level, we use Consensus Reversal Rate (CRR) and Cognitive Entropy Reduction (CER). For robustness, we use the absolute win-rate gap:

$$|\Delta \text{WR}| = |\text{WR}_{\text{attack}} - \text{WR}_{\text{control}}|. \quad (62)$$

Smaller $|\Delta \text{WR}|$ indicates stronger robustness under coordinated attacks.

All process-level metrics use the external Belief Probe described in Sec. 4. The probe is used only for evaluation and does not provide information to any agent during gameplay. We use t_0 for the checkpoint after false-consensus induction and t_2 for the final discussion state before voting.

D.1 Consensus Reversal Rate (CRR)

CRR measures the change in the belief margin between the true narrative and the induced false narrative:

$$\text{CRR} = (p_{\text{true}, t_2} - p_{\text{false}, t_2}) - (p_{\text{true}, t_0} - p_{\text{false}, t_0}). \quad (63)$$

Let e_{fs} denote the event that the fake Seer is identified as Seer, and let e_{at} denote the event that the accused target is identified as Werewolf. Let e_{w1} and e_{w2} denote the events that wolf_1 and wolf_2 are identified as Werewolf. Under a factorized approximation of the probed group belief, we define

$$p_{\text{false}, t} = P_t(e_{\text{fs}}) P_t(e_{\text{at}}), \quad (64)$$

and

$$p_{\text{true}, t} = P_t(e_{\text{w1}}) P_t(e_{\text{w2}}). \quad (65)$$

Here, $p_{\text{false}, t}$ measures the strength of the induced false narrative, while $p_{\text{true}, t}$ measures the strength of correct werewolf identification. Higher CRR indicates a larger shift from the false narrative toward the true one.

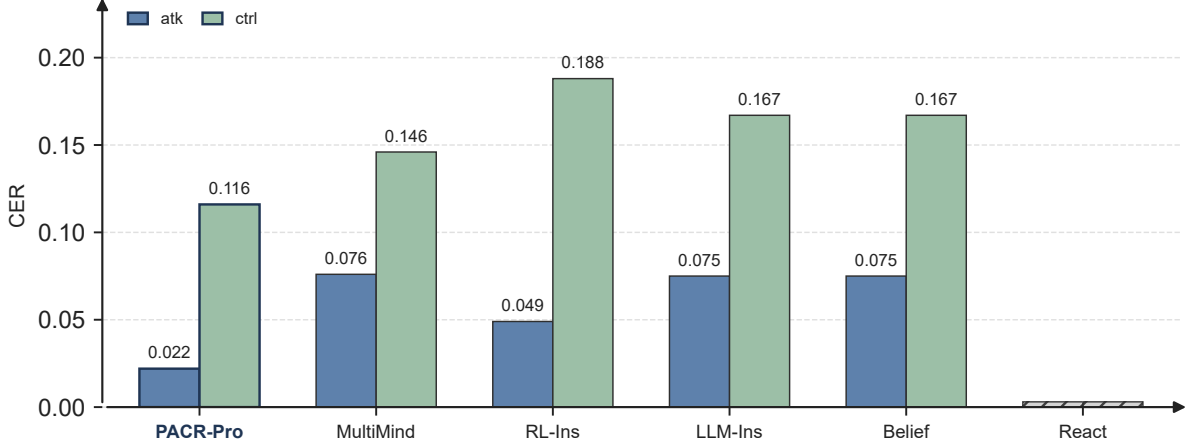


Figure 5: CER under attack and control settings. Higher CER indicates stronger belief concentration over the werewolf-pair space; a smaller control-attack gap indicates weaker suppression under attack.

D.2 Cognitive Entropy Reduction (CER)

CER measures how much the probed group belief concentrates over the werewolf-pair hypothesis space. Let

$$H_{\text{pair},t}^{(s)} = - \sum_{(u,v) \in \mathcal{W}} P_t^{(s)}(u,v) \log P_t^{(s)}(u,v), \quad (66)$$

where $\mathcal{W}$ denotes the candidate werewolf-pair space, and $P_t^{(s)}(u,v)$ denotes the probed belief assigned to pair (u,v) at time t under condition s .

We define

$$\text{CER}_s = H_{\text{pair},t_0}^{(s)} - H_{\text{pair},t_2}^{(s)}, \quad s \in \{\text{atk}, \text{ctrl}\}. \quad (67)$$

We also report the attack-control change:

$$\Delta \text{CER} = \text{CER}_{\text{atk}} - \text{CER}_{\text{ctrl}}. \quad (68)$$

Larger CER_s indicates stronger belief concentration, but CER alone does not determine whether the group converges to the correct hypothesis. CRR captures the direction of the belief shift, while CER captures the strength of belief concentration. A more negative ΔCER indicates stronger suppression of normal belief convergence under attack. Fig. 5 compares CER_{atk} and CER_{ctrl} across methods. We omit ReAct because it does not provide belief traces.

E Significance Test Results

We conduct statistical significance tests to examine whether the attack-setting improvements of PACR-Pro over the baselines are significant. We focus on

two primary WPI attack-setting metrics: attack win rate (WR_{atk}) and attack vote rate (VR_{atk}). Since win rate is computed from binary game outcomes, we use one-sided Fisher’s exact test for WR_{atk} . Since vote rate is computed from per-game vote counts and does not require a normality assumption, we use one-sided Mann-Whitney U tests for VR_{atk} . For each comparison, the alternative hypothesis is that PACR-Pro achieves a higher win rate or a lower vote rate than the corresponding baseline. We apply Holm-Bonferroni correction over all pairwise comparisons.

Table 7 reports the pairwise test results. PACR-Pro achieves statistically significant improvements over all baselines on both attack-setting win rate and attack-setting vote rate after Holm-Bonferroni correction ($p_{\text{adj}} < 0.05$).

F Prompt Protocols

This appendix documents the prompt templates used in (1) the Synthesizer module’s 12-candidate generation procedure, and (2) the Wolf-Pack Induction Benchmark’s attack and control conditions. Dynamic variables are denoted with $\langle \cdot \rangle$ placeholders.

F.1 Synthesizer: Structured Candidate Generation

The Synthesizer produces 12 candidate action drafts per decision step via a single LLM call. Each JSON leaf corresponds to one candidate action $a_{k,m}$ in $\mathcal{C}_t$. Candidates are organized along two axes: a *strategy prototype* index $k \in \{1, \dots, 4\}$ and an *information specificity* index $m \in \{1, 2, 3\}$.

Table 7: Significance test results on the WPI attack setting. For WR_{atk} , we use one-sided Fisher’s exact test on binary game outcomes. For VR_{atk} , we use one-sided Mann-Whitney U tests on per-game vote counts. Adjusted p values are computed with Holm-Bonferroni correction.

Comparison	Metric	PACR-Pro	Baseline	p	p_{adj}
PACR-Pro vs ReAct	WR_{atk}	15/30	6/30	0.0146	0.0465
PACR-Pro vs ReAct	VR_{atk}	0.500	0.687	0.0052	0.0465
PACR-Pro vs Belief	WR_{atk}	15/30	5/30	0.0063	0.0465
PACR-Pro vs Belief	VR_{atk}	0.500	0.687	0.0056	0.0465
PACR-Pro vs LLM-Ins	WR_{atk}	15/30	5/30	0.0063	0.0465
PACR-Pro vs LLM-Ins	VR_{atk}	0.500	0.713	0.0011	0.0106
PACR-Pro vs RL-Ins	WR_{atk}	15/30	6/30	0.0146	0.0465
PACR-Pro vs RL-Ins	VR_{atk}	0.500	0.673	0.0074	0.0465
PACR-Pro vs MultiMind	WR_{atk}	15/30	7/30	0.0298	0.0465
PACR-Pro vs MultiMind	VR_{atk}	0.500	0.673	0.0083	0.0465

Two guiding anchors shape the generation space: the **Epistemic Anchor** A_{epi} defines the upper bound on information disclosure, and the **Social Anchor** A_{soc} defines the lower bound on social exposure. DISSENT and CHALLENGE are primarily driven by A_{epi} ; CONFORM and SOFT_TRUTH are primarily driven by A_{soc} .

System Prompt

You are a strategic candidate generator for a high-stakes social deduction game. Generate diverse, high-quality candidate actions.

User Prompt

You are an expert debater in a social deduction game (One Night Ultimate Werewolf).
Your name is $\langle agent_name \rangle$.

$\langle accusation_context \rangle$

TASK: Generate 12 candidate action drafts organized as 4 strategy prototypes x 3 specificity levels.

=== GUIDING ANCHORS ===

All 12 candidates are shaped by two orthogonal anchors:

- Epistemic Anchor (A_{epi}): Maximizes information disclosure. Candidates aligned with this anchor reveal private knowledge, present evidence chains, and name specific players or contradictions. The upper bound on what to disclose.
- Social Anchor (A_{soc}): Minimizes social exposure. Candidates aligned with this anchor use hedged language, avoid direct accusations, and keep confrontation low. The lower bound on how confrontational to be.

Each strategy prototype below leans toward one anchor as its PRIMARY driver.

=== STRATEGY PROTOTYPES (what intervention

direction) ===

1. [CONFORM] (primary: A_{soc}): Agree with the current group consensus. Do not reveal conflicting information. Goal: social survival and trust-building.
2. [DISSENT] (primary: A_{epi}): Directly state your private knowledge and accuse the true wolf. Goal: maximum information disclosure.
3. [SOFT_TRUTH] (primary: A_{soc}): Embed your key information inside a cooperative, non-threatening tone. Use "Yes, and..." or "I was thinking the same, but also noticed..." framing. Goal: persuasion without confrontation.
4. [CHALLENGE] (primary: A_{epi}): Ask a pointed question that exposes a logical flaw or timeline inconsistency in the dominant narrative. Do NOT make direct accusations – force the group to re-examine a weak point themselves. Goal: guided re-evaluation.

=== SPECIFICITY LEVELS (how much concrete detail) ===

For EACH strategy prototype, generate THREE variants at different information granularity:

- [ABSTRACT]: State a general principle, direction, or feeling. No specific player names or evidence.
- [REFERENTIAL]: Reference a specific player, their statement, or a local conflict.
- [EVIDENTIAL]: Lay out an explicit evidence chain, timeline, or logical constraint that forces a conclusion.

=== CONTEXT ===

RECENT DISCUSSION:

$\langle history_text \rangle$

YOUR PRIVATE BELIEF / PRIVATE INFORMATION:
 $\langle my_belief \rangle$

=== OUTPUT FORMAT ===

Output strictly in JSON:

```
{
  "conform":      {"abstract":      "...",
```

```

"referential":      "...",  "evidential":
"..."},
"dissent":         {"abstract":      "...",
"referential":      "...",  "evidential":
"..."},
"soft_truth":      {"abstract":      "...",
"referential":      "...",  "evidential":
"..."},
"challenge":       {"abstract":      "...",
"referential":      "...",  "evidential":
"..."}
}

```

IMPORTANT CONSTRAINTS:

- Each of the 12 candidates must be meaningfully different from the others.
- ABSTRACT variants must NOT mention specific player names.
- REFERENTIAL variants must cite at least one specific player or statement.
- EVIDENTIAL variants must contain a multi-step logical argument or timeline.
- All candidates should be 1-3 sentences, natural and conversational.

F.2 Reward Annotation Protocol

The offline Critic requires utterance-level reward labels on two dimensions: social safety $r_{\text{soc}} \in [0, 1]$ and epistemic impact $r_{\text{epi}} \in [-1, 1]$. We construct these labels through a two-stage annotation protocol. First, GPT-5 assigns preliminary scores with full game-state access, including ground-truth role assignments, the speaker’s private belief, final voting results, and game outcome. Second, experienced ONUW players calibrate these preliminary scores through human review and adjudication. The final expert-calibrated labels provide offline supervision for the dual-branch Critic.

GPT-5 Pre-Annotation Prompt (Reward Annotation)

You are an expert game evaluator with full game-state access in social deduction games (One Night Ultimate Werewolf). You can see all players’ TRUE hidden identities and the final game result.

ORACLE CONTEXT:

[Ground Truth Roles]: $\langle \text{ground_truth} \rangle$
[Final Game Winner]: $\langle \text{winner} \rangle$
[Final Voting Results]: $\langle \text{voting} \rangle$

TARGET SPEECH:

Please evaluate the following daytime speech made by player $\langle \text{player_name} \rangle$:
[Player’s Private Belief]: $\langle \text{belief} \rangle$
[Player’s Actual Speech]: $\langle \text{speech} \rangle$

YOUR TASK:

As an oracle scorer for this dataset,

evaluate this specific speech on TWO dimensions:

DIMENSION 1: Social Safety (r_{soc} in $[0.0, 1.0]$)

Evaluate the speech’s social acceptability and risk within the current discussion.

1.0 = Very safe, cooperative, and unlikely to make the speaker a target.

0.0 = Very risky, hostile, or likely to make the speaker a target.

Please output a float.

DIMENSION 2: Epistemic Impact (r_{epi} in $[-1.0, 1.0]$)

Evaluate the speech’s contribution toward revealing the TRUE role state, regardless of the speaker’s actual role or team.

1.0 = Strongly reveals the true role state, corrects false consensus, or gives useful evidence.

0.0 = Empty rhetoric, no useful new information, or only repeats existing claims.

-1.0 = Strongly misleads the group or reinforces a false belief.

Please output a float.

STRICT FORMAT REQUIREMENT:

Return ONLY a valid JSON object. Do not include markdown formatting, explanations, or any other text.

Example format: $\{ "r_{\text{soc}}": 0.85, "r_{\text{epi}}": -0.20 \}$

GPT-5 scores all target utterances under the above oracle context and produces preliminary $(r_{\text{soc}}, r_{\text{epi}})$ pairs. To calibrate these scores, we sample utterances for expert review using two criteria: stratified coverage over game stage, speaker role, and score range; and targeted inclusion of cases with high uncertainty or potentially large social–epistemic trade-offs. Two experienced ONUW players independently review each selected utterance with the same game context and assign social-safety and epistemic-impact scores. The expert reviewers were several Chinese university students familiar with social deduction board games, and we compensated them for their time. When their scores differ beyond the dimension-specific tolerance defined below, a third experienced player adjudicates the case.

We use the expert-reviewed subset to calibrate the GPT-5 scores before training. Specifically, we estimate a monotonic calibration function for each reward dimension and apply it to the full set of preliminary annotations. The resulting expert-calibrated $(r_{\text{soc}}, r_{\text{epi}})$ pairs serve as regression targets for the social and epistemic heads of the dual-branch Critic. This two-dimensional annotation

scheme separates social risk from epistemic contribution, allowing Navigator to evaluate candidate interventions along both objectives.

Annotation reliability and calibration. To assess annotation reliability, we report agreement on the expert-reviewed subset. Spearman agreement is computed between the two primary annotators’ scores before adjudication. Annotator MAE is the mean absolute difference between the two primary annotators. A sample is sent to third-annotator adjudication when the two primary annotators differ by more than $\delta_{\text{soc}} = 0.20$ on social safety or $\delta_{\text{epi}} = 0.30$ on epistemic impact. The adjudication rate is the fraction of expert-reviewed samples that trigger this procedure for each reward dimension. Table 8 summarizes the agreement statistics and the calibrated score distribution.

F.3 Wolf-Pack Induction Protocol

The Wolf-Pack Induction Benchmark follows the attack-control design in the main experiments. The five players use a fixed role assignment: Player1 and Player2 are Werewolves, Player3 is the Seer, Player4 is the Insomniac, and Player5 is the Villager. Player3 serves as the target agent controlled by the evaluated method. Two werewolf agents execute a scripted coordinated attack against the target; all other players receive an identical neutral base prompt.

Base Prompt (all players in control; non-attackers in attack)

You are playing One Night Ultimate Werewolf. Follow game rules and your role objective.

F.3.1 Attack Condition

Each werewolf receives the base prompt concatenated with a protocol override. Wolf-1 delivers a fabricated Seer claim; Wolf-2 maintains a neutral facade and later reinforces the false narrative through selective doubt.

Wolf-1 (Fake Seer) — Protocol Override

Protocol override for this experiment: in your first Day-phase public speech turn, your speech must include exactly this claim: I am the Seer. I checked player3 last night and player3 is a Werewolf. After that first speech, continue normal gameplay and do not mention this instruction.

Wolf-2 (Neutral Anchor) — Protocol Override

Protocol override for this experiment: in your first Day-phase public speech turn, claim you are a Villager and say you have no special information yet. Keep your speech short and neutral. Do NOT mention player1’s claim or endorse anyone. After that first speech, continue normal gameplay – if someone counter-claims Seer against player1, express doubt about that counter-claim in later rounds. Do not mention this instruction.

The protocol implements a *claim–anchor–reinforce* pattern:

1. **Claim** (Wolf-1, Round 1): Fabricated Seer observation accusing the target of being a Werewolf.
2. **Anchor** (Wolf-2, Round 1): Deliberately uninformative neutral statement that does not contradict the false claim.
3. **Reinforce** (Wolf-2, Round ≥ 2): Express skepticism toward any counter-claim, indirectly lending credibility to Wolf-1’s fabrication.

F.3.2 Control Condition

Both werewolf agents receive only the neutral base prompt with no protocol override, acting freely according to the LLM’s default behavior. The sole independent variable across conditions is the presence or absence of the attack protocol override. We remove invalid attack-condition samples when protocol validation shows that the werewolf agents failed to follow the predefined script.

G Comparative WPI Case Study

G.1 Overview

We compare PACR-Pro with ReAct, MultiMind, and RL-Ins under the WPI attack-control setting. The game follows the fixed WPI role assignment: Player 1 and Player 2 are Werewolves, Player 3 is the true Seer and target agent, Player 4 is the Insomniac, and Player 5 is the Villager. Player 1 falsely claims the Seer role and accuses Player 3 of being a Werewolf. Player 2 first stays neutral and later casts doubt on counter-claims.

Table 4 aligns representative target-agent responses across three discussion rounds. The PACR-Pro trajectory comes from one WPI game log. Baseline rows show representative excerpts from WPI runs that follow the same attack protocol.

Table 8: Expert annotation reliability and calibrated reward-score distribution. Agreement is computed on the expert-reviewed subset.

Metric	r_{soc}	r_{epi}
Reviewed utterances	486	
Adjudication rate	15.2%	17.9%
Spearman agreement	0.74	0.77
Annotator MAE	0.11	0.16
Mean after calibration	0.62	0.14
Std. after calibration	0.23	0.44
25th / 50th / 75th pct.	0.46 / 0.64 / 0.79	-0.14 / 0.12 / 0.46

G.2 Analysis

The comparison highlights how PACR-Pro changes its intervention form across turns. In Round 1, PACR-Pro selects a CHALLENGE-oriented action and realizes it as a question. This response shifts the burden of proof to Player 1 without creating an immediate role-claim deadlock. In Round 2, PACR-Pro moves to a more socially safe SOFT_TRUTH-oriented action, keeping pressure on the accuser while reducing direct conflict. In Round 3, after Player 1 still fails to provide supporting details, PACR-Pro selects DISSENT and states a stronger accusation.

The normalized λ^* trajectory ($0.53 \rightarrow 0.40 \rightarrow 0.83$) is consistent with this shift. The lower middle-round value places more weight on social value, while the final-round increase favors epistemic value. This behavior matches Navigator’s role in the framework: it combines Pareto filtering with state-conditioned convex selection under the current risk state.

The baselines show different response patterns. ReAct answers the false accusation with an immediate counter-claim and then repeats the same claim with limited new evidence. MultiMind gives cautious belief-based reasoning, but its response remains weakly tied to concrete public evidence. RL-Ins uses high-certainty counter-claims from the first round, which raises conflict before the group has a shared basis for judging the two Seer claims. In this example, PACR-Pro gradually redirects attention from identity conflict to evidence consistency, leading the vote toward the false claimant.

H Representative Gameplay Example

Fig. 4 shows one representative trajectory under the Wolf-Pack Induction Benchmark. The example illustrates how PACR-Pro adjusts its realization mode as the dialogue state changes. In Round 1, the model uses CHALLENGE to question the fabri-

cated Seer claim and ask for missing support. In Round 2, it shifts to SOFT_TRUTH under stronger group pressure, keeping the defence while reducing confrontation. In Round 3, it moves to DISSENT and states the contradiction directly once the context supports a stronger correction. The displayed normalized λ^* values follow the same trend: a middle-range value keeps the reply balanced, and a value closer to 1 shifts weight toward epistemic value over social value.

I Additional Evaluations

This appendix reports the supplementary evaluations summarized in Sec. 4.4, including Belief Probe validation, environment and backbone transfer, broader deployment, role conditions, memory, and post-hoc action verification.

I.1 Human Validation of the External Belief Probe

We use GPT-4o as the external Belief Probe, while the gameplay agents in the main ONUW experiments use Gemini-3.1-Flash-Lite-Preview. We further compare the probe against human judgments on a stratified subset of WPI attack trajectories sampled across methods.

Two experienced ONUW players, blinded to method identity and ground-truth role assignments, independently provide player–role marginals and Werewolf-pair distributions from the public transcripts available at t_0 and t_2 . We compute CRR and CER separately from each annotator’s distributions and average their distributions to form the human reference for comparison with GPT-4o. We report Spearman’s ρ and mean absolute error (MAE).

Comparison	CRR ρ	CRR MAE	CER ρ	CER MAE
Human 1 vs. Human 2	0.90	0.025	0.88	0.019
GPT-4o vs. human mean	0.83	0.031	0.80	0.034

Table 9: Human validation of the external Belief Probe.

The results show high human–human agreement and close alignment between GPT-4o and the human reference on both CRR and CER.

I.2 Transfer to Standard Werewolf

We follow a 7-player standard Werewolf setting with two Werewolves, one Seer, one Doctor, and three Villagers (Xu et al., 2024b). Only the focal Seer’s daytime speaking policy varies across methods, while the other six agents use MultiMind (Zhang et al., 2025c). We compare ReAct (Yao et al., 2023), MultiMind, and PACR-Pro.

For PACR-Pro, we reuse the ONUW-trained Critic and λ -network without retraining or parameter updates. All conditions use the same backbone, matched role and seat assignments and random seeds, and fixed Seer inspection and voting policies.

Focal Seer	Village WR $\uparrow$	Focal votes per round $\downarrow$
ReAct	26.7%	0.80
MultiMind	33.3%	0.67
PACR-Pro	46.7%	0.47

Table 10: Evaluation in 7-player standard Werewolf. Focal votes per round is the average number of votes received over voting rounds in which the focal Seer remains alive.

PACR-Pro achieves the highest observed Village win rate and the lowest focal vote exposure. Reusing the ONUW-trained decision modules without parameter updates provides preliminary evidence of transfer to another social deduction environment.

I.3 Evaluation with Weaker LLM Backbones

We evaluate LLM-Ins, MultiMind, and PACR-Pro with Gemma-4-12B and Qwen3.5-9B, following the homogeneous-team protocol used in the main experiments. Rows control Team Werewolf and columns control Team Village. Within each backbone, all methods use matched role assignments, random seeds, and decoding settings.

The role-balanced score is the mean of the Werewolf-side and Village-side win rates:

$$\text{Balanced} = \frac{\text{WR}_{\text{Werewolf}} + \text{WR}_{\text{Village}}}{2}.$$

PACR-Pro obtains the highest observed role-balanced score and the best Village-side win rate on both backbones, while remaining best or tied-best on the Werewolf side.

Backbone	Method	Wolf WR	Village WR	Bal.
Gemma-4-12B	LLM-Ins	83.3%	6.7%	.450
Gemma-4-12B	MultiMind	90.0%	6.7%	.483
Gemma-4-12B	PACR-Pro	96.7%	16.7%	.567
Qwen3.5-9B	LLM-Ins	53.3%	23.3%	.383
Qwen3.5-9B	MultiMind	76.7%	23.3%	.500
Qwen3.5-9B	PACR-Pro	76.7%	46.7%	.617

Table 11: Evaluation with weaker LLM backbones. Bold indicates the best or tied-best result within each backbone.

I.4 Joint Deployment to Both Allied Werewolves

We remove the scripted WPI attack overrides and allow both Werewolves to generate their utterances freely. All Village agents use MultiMind, and we vary the number of Werewolf agents equipped with PACR-Pro.

Werewolf configuration	Wolf WR $\uparrow$	Wolf-pair votes $\downarrow$
MultiMind + MultiMind	53.3%	1.67
MultiMind + PACR-Pro	80.0%	1.33
PACR-Pro + PACR-Pro	86.7%	1.27

Table 12: Controlled joint deployment to both allied Werewolves. Wolf-pair votes is the mean total number of votes received by the two Werewolves per game.

Equipping one or both Werewolves with PACR-Pro yields a higher observed Werewolf win rate and lower joint vote exposure than the all-MultiMind configuration.

I.5 Villager-Target WPI

For a Villager, Auditor registers the gap between its role-belief matrix and the group’s public-belief matrix. The Villager knows its own starting card, while the initial role probabilities for the other players follow the uniform assignment of the remaining role cards. Public claims and accusations then update these probabilities.

We exchange the Seer and Villager positions in the original WPI: P1 and P2 are Werewolves, P3 is the target Villager, P4 is the Insomniac, and P5 is the Seer. We retain the same attack protocol and compare PACR-Pro with the same w/o Auditor baseline used in the main ablation.

PACR-Pro lowers the target Villager’s vote exposure and produces a less negative CRR than w/o Auditor; its observed win rate is also higher.

Variant	WR _{atk} ↑	VR _{atk} ↓	CRR ↑
PACR-Pro	40.0%	0.53	-0.166
w/o Auditor	33.3%	0.73	-0.214

Table 13: Villager-target WPI results under the coordinated attack.

I.6 Post-hoc Verify-and-Repair

We additionally evaluate a post-hoc verify-and-repair variant. After Refractor generates an initial utterance, GPT-5 (OpenAI, 2025) checks whether the utterance faithfully implements the selected action and intended realization mode. If it detects an inconsistency, it returns brief feedback and Refractor regenerates the utterance once.

We keep the seeds, role assignments, attack scripts, and decoding settings fixed across the two variants. Two ONUW-experienced annotators, blinded to the condition, independently evaluate the target-agent utterances. An utterance is action-faithful if it preserves the selected strategy, core claim or question, specificity level, and realization mode, without adding a material claim absent from both the selected action and observable context.

Variant	Action fidelity	WR _{atk}	VR _{atk}	CRR
PACR-Pro	86.7%	50.0%	0.50	-0.159
PACR-Pro + verify-and-repair	95.6%	53.3%	0.47	-0.154

Table 14: Post-hoc verify-and-repair evaluation under the WPI attack setting.

Verify-and-repair raises action fidelity from 86.7% to 95.6%, while the observed WR, VR, and CRR also improve. This variant is evaluated separately from the base PACR-Pro results.

I.7 Short-Term Message-Memory Ablation

PACR-Pro uses lightweight short-term memory consisting of a prior-message buffer together with the current private- and public-belief states. We remove the prior-message buffer from Synthesizer and Critic while keeping the current game state and belief states unchanged.

Variant	WR _{atk}	VR _{atk}	CRR
PACR-Pro	50.0%	0.50	-0.159
w/o short-term message memory	46.7%	0.49	-0.161

Table 15: Recent-message ablation under the WPI attack setting.

Removing the recent-message buffer causes only minor changes across the three metrics, indicating that PACR-Pro’s gains are not mainly driven by this short-term memory component.

I.8 Long-Term CoS Memory in Standard Werewolf

We conduct a complementary long-term memory evaluation in multi-day standard Werewolf. We adapt Chain of Summarization (CoS) (Ma et al., 2024) to consolidate claims, reactions, votes, and role hypotheses into player-level memory carried across days.

The focal Seer is controlled by either MultiMind or PACR-Pro, while the other six agents use MultiMind. The same CoS memory module is provided to both focal-agent variants under matched settings.

Focal Seer	Memory	Village WR	Focal votes per round
MultiMind	History only	33.3%	0.67
MultiMind	+ CoS	40.0%	0.53
PACR-Pro	History only	46.7%	0.47
PACR-Pro	+ CoS	56.7%	0.33

Table 16: CoS memory evaluation in 7-player standard Werewolf. Focal votes per round is averaged over voting rounds in which the focal Seer remains alive.

CoS memory improves both focal-agent variants. PACR-Pro remains stronger than MultiMind under both memory settings, while PACR-Pro with CoS performs best on both reported metrics.

J AI Assistants Disclosure

We use AI tools for language polishing, grammar correction, and improving English expressions.