PSY-MARL: PSYCHOLOGY-GUIDED MULTI-AGENT REINFORCEMENT LEARNING FOR PSYCHIATRIC INTERVIEWING

Anonymous authors
Paper under double-blind review

ABSTRACT

Psychiatric interviewing plays a central role in diagnosis and treatment, requiring clinicians to explore symptoms while attending to the patient's psychological state and communication needs. Advances in large language models offer new opportunities for psychiatric interviewing, yet clinical dialogue reasoning in this highly context-dependent setting remains challenging. To this end, we propose Psy-MARL, a psychology-guided multi-agent reinforcement learning framework inspired by Theory of Mind and Motivational Interviewing. Psy-MARL comprises three agents: the Belief agent assesses the patient's state, the Strategy agent plans a communication strategy, and the Response agent generates a physician response. Their policies are jointly optimized through reinforcement learning. To obtain informative reward signals and distinguish the contributions of agents, we introduce preference-based coalition credit assignment. This method models role collaboration as a cooperative game, fits coalition utilities from LLM preferences using a Bradley–Terry model, and derives role-specific rewards through Shapley allocation. We train and evaluate Psy-MARL on real-world psychiatric consultation dialogues, using LLM-based pairwise evaluations to assess alignment with recorded physician responses and preference under clinical evaluation criteria. Psy-MARL achieves positive net-win margins against all evaluated single-agent and multi-agent baselines under both protocols, including clinical margins of 6.36–27.19 percentage points against three multi-agent reinforcement learning methods. Ablation studies support the benefits of preference-based coalition utility estimation and role-specific credit allocation. These findings support psychology-guided agent collaboration and preference-based joint training as a promising approach to strengthening clinical reasoning in psychiatric interviewing.

1 INTRODUCTION

Psychiatric interviewing provides a foundation for subsequent diagnosis and treatment, requiring systematic exploration of symptoms and functional impairment while maintaining a collaborative relationship with the patient (Shea, 2016). Clinicians must consider how patients understand their circumstances, their emotional concerns, and their willingness to elaborate when deciding what to clarify and how to pursue it (Miller & Rollnick, 2023). These interdependent decisions require contextual reasoning that connects patient-state understanding with communication planning while maintaining the clinical focus of the interview (Ji et al., 2026; Bi et al., 2026).

Advances in the reasoning and conversational capabilities of large language models (LLMs) have opened new possibilities for interactive psychiatric interviewing (Na et al., 2025). One line of work uses established clinical protocols to guide symptom inquiry (Bi et al., 2025; Tu et al., 2025). Another trains models on domain-specific dialogues to generate doctor responses, with some methods incorporating patient states and dialogue intentions (Yao et al., 2022; Lan et al., 2025; Gu et al., 2026). However, relying on predefined interview procedures or response patterns learned from demonstrations alone is insufficient for adapting clinical inquiry and communication strategies to variations in patients' states and expressions. Strengthening clinical reasoning therefore remains a key challenge for psychiatric interviewing systems, where appropriate decisions depend heavily on the conversational context.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To strengthen clinical reasoning in psychiatric interviewing, we propose **Psy-MARL**, a psychology-guided multi-agent reinforcement learning framework. We draw on Theory of Mind (ToM), which concerns inferring others' mental states, including their beliefs and intentions (Premack & Woodruff, 1978), and Motivational Interviewing (MI), which emphasizes strengthening motivation for change through collaborative communication and respect for autonomy (Miller & Rollnick, 2023). Guided by these complementary perspectives, our multi-agent system (MAS) comprises three sequential roles: the Belief Agent assesses the patient's concerns, emotions, and willingness to elaborate; the Strategy Agent uses this assessment to plan a communication action and its focus; and the Response Agent integrates both intermediate outputs with the dialogue context to generate the physician response.

We then jointly optimize the three agents through reinforcement learning (RL) on a corpus of real-world psychiatric consultations to further strengthen the system's clinical reasoning. To address the challenges of credit assignment across agents and obtaining informative reward signals, we develop **preference-based coalition credit assignment**, modeling role collaboration as a cooperative game. We estimate coalition utilities from LLM preference judgments through convex optimization of a Bradley–Terry model (Bradley & Terry, 1952). Drawing on Shapley values (Shapley, 1952), we aggregate each role's marginal contributions across coalitions to derive role-specific rewards for joint reinforcement learning. Figure 1 provides an overview of the framework.

We evaluate Psy-MARL on a corpus of real-world psychiatric consultations, assessing alignment with recorded physician responses and preference under clinical evaluation criteria. Psy-MARL achieves positive net-win margins under both protocols against all evaluated single-agent and multi-agent baselines, including task-adapted multi-agent RL methods. Compared with ablation variants using direct scoring or shared team rewards, Psy-MARL achieves stronger performance, supporting the effectiveness of its preference-based coalition credit assignment.

Our main contributions are as follows:

- **Psychology-guided multi-agent design.** We design a sequential architecture informed by Theory of Mind and Motivational Interviewing, in which specialized roles for patient-state assessment and communication planning jointly guide physician response generation.
- **Preference-based coalition credit assignment.** We develop a training approach that fits coalition utilities from LLM preference judgments using a Bradley–Terry model and applies Shapley allocation to derive role-specific rewards. This connects feedback on final responses to the joint optimization of intermediate reasoning and communication decisions.
- **Empirical evaluation and analysis.** We evaluate both reference alignment and clinical-rubric preference, and examine the training design through ablations comparing preference-based coalition utility estimation with direct scoring and role-specific credit allocation with shared team rewards.

2 Related Work

2.1 Dialogue Systems for Psychiatric Interviews

Research on psychiatric dialogue has developed resources for distress assessment and depression-oriented interviewing (Gratch et al., 2014; Yao et al., 2022), with recent work extending to psychiatric comorbidity and multimodal clinical assessment (Wan et al., 2026; Chen et al., 2026b). Beyond these resources, studies explore diagnostic modeling from structured patient information (Sha et al., 2025) and the organization of interview interactions. TRUST combines cooperative modules with a clinical dialogue-act schema for trauma interviewing (Tu et al., 2025), while other approaches investigate memory and reflection in simulated consultations (Lan et al., 2024) and unobtrusive symptom probing during diagnostic dialogue (Cao et al., 2025). In related mental health support and counseling tasks, specialized agents support care providers and coordinate CBT-oriented consultation (Kampman et al., 2024; Xu et al., 2025). Empathy-R1 combines structured empathy reasoning with reinforcement learning in a single language model (Yao et al., 2025), while DeepSAGE learns a therapeutic-intention policy to guide LLM responses within a staged CBT protocol (Zhang et al., 2026). Psy-MARL draws on Theory of Mind and Motivational Interviewing (Premack & Woodruff, 1978; Miller & Rollnick, 2023) to organize patient-state assessment, communication planning, and

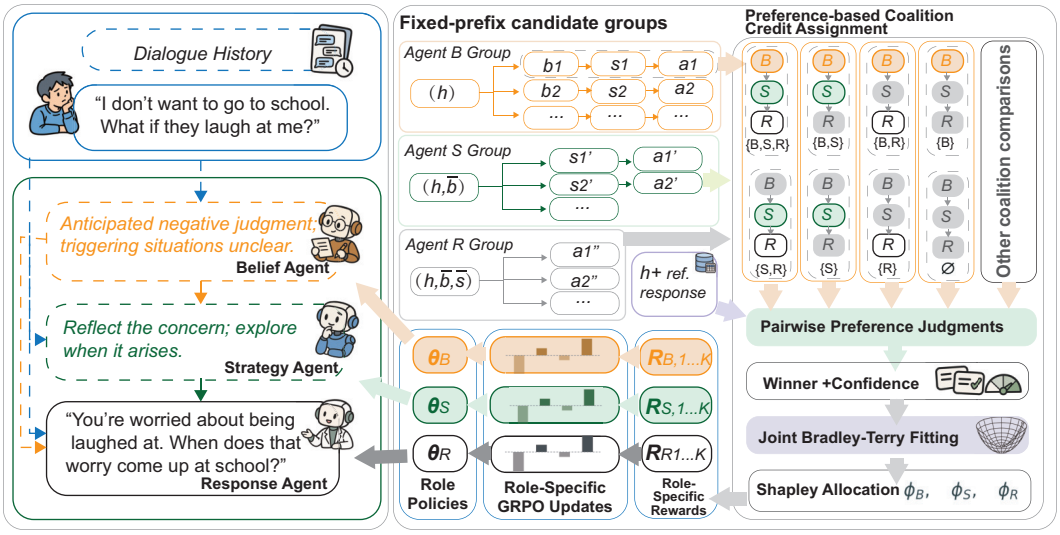

Figure 1: Overview of Psy-MARL. Left: psychology-guided Belief–Strategy–Response collaboration. Right: joint training of the three agents through preference-based coalition credit assignment and role-specific GRPO updates. Gray-filled coalition nodes denote frozen reference policies.

response generation into sequential roles with independently trainable policies. It uses feedback on the final response to jointly optimize these interdependent roles, linking intermediate patient-state and communication decisions to their effects on the resulting doctor response.

2.2 Multi-Agent Reinforcement Learning and Credit Assignment

Recent work applies reinforcement learning to collaborative LLM workflows, including multi-agent discussion (Park et al., 2025), generator–critic refinement (Pulici et al., 2026), and conversational self-play (Jiang et al., 2026). MARFT (Liao et al., 2025) provides a general framework for reinforcement fine-tuning of multi-agent systems. Beyond workflow design, optimization must account for differences in agents' roles and interaction histories. Dr. MAS (Feng et al., 2026) introduces agent-wise advantage normalization to accommodate different reward distributions, while STRONGER-MAS (Zhao et al., 2026) develops agent- and turn-wise grouping for policy optimization. Alongside these advances in optimization, credit assignment addresses how feedback on a collective outcome can guide the learning of individual roles.

Classical cooperative MARL approaches construct individual learning signals through value decomposition (Sunehag et al., 2017; Rashid et al., 2020) or counterfactual advantages (Foerster et al., 2018). For LLM-based systems, C3 (Chen et al., 2026a) estimates decision-level credit by comparing alternative actions under fixed interaction histories, while SHARP (Li et al., 2026) combines Shapley-inspired marginal credit with outcome and tool-process feedback. In SHARP, marginal credit is derived from changes in task accuracy under agent ablation. Applying coalition-based credit assignment to psychiatric interviewing additionally requires defining how open-ended responses are valued against recorded physician replies. Psy-MARL constructs policy coalitions that preserve the sequential collaboration workflow and estimates their utilities from pairwise reference-similarity judgments using a Bradley–Terry model. It then computes exact Shapley values over the resulting estimated game to obtain role-specific training rewards. This separates the estimation of coalition utility from the allocation of individual credit, linking comparative feedback on final responses to the contributions of intermediate patient-state assessment and communication planning.

3 METHODOLOGY

3.1 PSYCHOLOGY-GUIDED MULTI-AGENT ARCHITECTURE

Given available patient information x and a multi-turn dialogue history H_t, we denote their combination by $h = (x, H_t)$. The history includes physician, patient, and guardian utterances when present. Each rollout produces one external physician response ($T = 1$); intermediate agent outputs are internal computations rather than additional patient turns. The reference physician response y_t^* provides training supervision and is excluded from all agent inputs. Figure 1 summarizes the architecture and training framework.

We draw on Theory of Mind (ToM) (Premack & Woodruff, 1978) to motivate reasoning about the patient's perspective, and on Motivational Interviewing (MI) (Miller & Rollnick, 2023) to inform collaborative communication that respects the patient's autonomy and willingness to engage. These considerations motivate three complementary roles that separate patient-state assessment, communication planning, and physician response generation.

Belief Agent: patient-state assessment. The Belief Agent produces an assessment b of the patient's concerns, emotions, perspective, and willingness to elaborate from h. It distinguishes reported information from tentative interpretations and identifies unresolved information to guide downstream planning and response generation.

Strategy Agent: communication planning. Given h and b, the Strategy Agent generates a plan s specifying a communication action and a dialogue-grounded focus. Actions draw on MI-informed techniques such as open-ended exploration, reflective listening, affirmation, and summarization, while allowing specific clarification when needed. Plans are concise action-and-focus instructions without a fixed set of action labels.

Response Agent: strategy-conditioned response generation. The Response Agent integrates h, b, and s to generate a natural, concise physician response a that implements the selected strategy while remaining grounded in the dialogue.

Let $\mathcal{N} = \{B, S, R\}$ denote the three roles. Their sequential dependencies define the joint policy

$$\pi_\theta(b, s, a \mid h) = \pi_{\theta_B}(b \mid h)\, \pi_{\theta_S}(s \mid h, b)\, \pi_{\theta_R}(a \mid h, b, s), \tag{1}$$

where $\theta = (\theta_B, \theta_S, \theta_R)$ denotes the role-specific trainable parameters and role prompts are implicit in the policy notation. All three agents retain access to the original dialogue context h. The complete role prompts are provided in Appendix A.1.

3.2 PREFERENCE-BASED COALITION UTILITY ESTIMATION

In psychiatric interviewing, the usefulness of a patient-state assessment or communication plan depends on how downstream agents interpret and realize it, making intermediate outputs difficult to evaluate in isolation. Joint training from final-response feedback thus raises two challenges: (i) **allocating credit to roles whose effects on the final response are interdependent**, and (ii) **estimating discriminative and comparable utilities for different role-policy configurations**.

Cooperative game and policy coalitions. To address the first challenge, we model role collaboration as a cooperative game $(\mathcal{N}, v)$, with the agent roles as players. A coalition $S \subseteq \mathcal{N}$ is a subset of these players, and the characteristic function $v : 2^{\mathcal{N}} \to \mathbb{R}$ assigns a utility to each coalition. Within this framework, we use the Shapley value (Shapley, 1952) to allocate utility gains to individual roles. For role i and a coalition S excluding i, the marginal contribution $v(S \cup \{i\}) - v(S)$ measures the utility change associated with joining coalition S. Its Shapley credit aggregates these changes across all coalitions that exclude i:

$$\phi_i(v) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left[v(S \cup \{i\}) - v(S) \right], \tag{2}$$

The weights correspond to a uniform average of marginal contributions over all possible player orderings, accounting for how a role's contribution varies with its collaborators. This yields separate credit values $\phi_B(v)$, $\phi_S(v)$, and $\phi_R(v)$ for the three roles.

Coalition evaluation requires each configuration to produce a complete physician response. Directly removing nonmember roles would disrupt the sequential workflow and, for coalitions excluding the Response Agent, leave no final response to evaluate. We therefore preserve the complete Belief–Strategy–Response workflow and construct coalitions through policy substitution, using frozen reference policies for nonmember roles. This allows role-policy combinations to be evaluated while retaining the downstream interactions through which intermediate outputs affect the final response. For each coalition S, role policies are selected and outputs are sampled sequentially as follows:

$$\pi_i^S = \begin{cases} \pi_{\theta_i}, & i \in S, \\ \pi_i^{\mathrm{ref}}, & i \notin S, \end{cases} \qquad i \in \mathcal{N}, \tag{3}$$
$$b^S \sim \pi_B^S(\cdot \mid h), \quad s^S \sim \pi_S^S(\cdot \mid h, b^S), \quad a^S \sim \pi_R^S(\cdot \mid h, b^S, s^S).$$

Here, π_i^{ref} denotes the frozen backbone policy without a LoRA adapter, and b^S, s^S, and a^S denote the coalition's patient-state assessment, communication plan, and final physician response, respectively.

Preference-based coalition utility evaluation. To address the second challenge, we develop a preference-based estimator of coalition utility. Unlike tasks assessed through exact-answer checks or executable tests, psychiatric dialogue lacks a simple rule-based correctness signal. LLM-based scoring offers a flexible alternative, but direct numerical ratings may concentrate on a few discrete values and obscure subtle differences between outputs (Liu et al., 2023). This limitation is particularly relevant to Shapley allocation, which depends on utility differences between related coalitions: insufficient score discrimination may mask the marginal effects of role-policy changes. We therefore combine comparative LLM judgments (Liusie et al., 2024) with joint Bradley–Terry utility fitting.

We elicit preferences between coalitions that differ in one role's policy, matching the marginal comparisons required by Shapley allocation. Specifically, we define $\mathcal{D} = \{(S \cup \{i\}, S) : i \in \mathcal{N},\ S \subseteq \mathcal{N} \setminus \{i\}\}$. For each pair $(S_a, S_b) \in \mathcal{D}$, the judge compares the complete responses a^{S_a} and a^{S_b} under the same dialogue context h and reference physician reply y_t^*. It selects the response whose content, wording, and communicative intent more closely match the reference. The selected winner and reported confidence are converted into a soft preference target $q_{ab} \in (0, 1)$ for S_a over S_b. Details of coalition evaluation, rollout reuse, preference mapping, and BT fitting, together with training judge prompts, are provided in Appendix A.2.

Bradley–Terry utility fitting. These preference judgments provide local comparisons, whereas Shapley allocation requires numerical utilities on a common scale. We jointly fit a Bradley–Terry (BT) model (Bradley & Terry, 1952) over the comparison graph to obtain such utilities. The model relates preference to utility differences through $p_{ab}(v) = \sigma(v(S_a) - v(S_b))$, where σ is the logistic function. For each sampled collection of coalition responses under history h, we solve

$$\hat{v}_h = \operatorname*{arg\,min}_{v:\, v(\emptyset)=0} - \sum_{(S_a, S_b) \in \mathcal{D}} \left[q_{ab} \log p_{ab}(v) + (1 - q_{ab}) \log(1 - p_{ab}(v)) \right]. \tag{4}$$

Joint fitting estimates coalition utilities that best match the pairwise judgments under the BT model, providing a common scale for computing marginal contributions. We fix $v(\emptyset) = 0$ to set the reference level without changing utility differences or Shapley credits. With a connected comparison graph and soft preference targets $q_{ab} \in (0, 1)$, the anchored objective is strictly convex and admits a unique finite minimizer. The fitted utilities define the estimated coalition game $\hat{v}_h$, from which Equation 2 computes role-specific credits.

3.3 Role-Specific Policy Optimization

From credit to training reward. Given the estimated coalition game $\hat{v}_h$, we convert each role's Shapley credit into a bounded training reward:

$$r_i = \mathrm{clip}(\tau \phi_i(\hat{v}_h), -r_{\max}, r_{\max}), \tag{5}$$

where $\tau > 0$ controls reward scaling and $r_{\max} > 0$ is the clipping threshold. Their values are reported in Appendix A.4.

Fixed-prefix rollout groups. In the sequential workflow, downstream inputs depend on sampled upstream outputs. Comparing candidates generated from different prefixes would therefore mix variation in a role's behavior with variation in its input. Inspired by AT-GRPO (Zhao et al., 2026), we construct each role's update group from K candidates sharing the same complete input: h for Belief, (h, b) for Strategy, and (h, b, s) for Response. Shared upstream prefixes are selected randomly from sampled trajectories. Each candidate is completed into a physician response and receives a reward derived from the target role's Shapley credit in its corresponding fitted coalition game (Appendix A.2). Only candidates belonging to the corresponding fixed-prefix group contribute to that role's update.

Group-relative policy updates. For each role i, let $\{r_{i,k}\}_{k=1}^{K}$ denote the rewards within an update group. We compute group-relative advantages as

$$\hat{A}_{i,k} = \frac{r_{i,k} - \mu_i}{\sigma_i + \epsilon}, \qquad k = 1, \ldots, K, \tag{6}$$

where μ_i and σ_i are the group's reward mean and standard deviation, respectively, and $\epsilon > 0$ ensures numerical stability. We optimize each role's policy using its corresponding advantages in a clipped GRPO objective (Shao et al., 2024). Coalition-based credit assignment incurs additional training overhead, which we mitigate through batched preference evaluation and rollout reuse (Appendix A.7).

4 Experiments

4.1 Experimental Setup

Dataset. We use an internally collected corpus of 350 de-identified Mandarin Chinese outpatient psychiatric consultations involving patients aged 10–18. We partition the corpus at the patient level before constructing turn-level examples, keeping all consultations from the same patient in a single split. Each example pairs the available patient information and preceding dialogue history with the recorded physician response as its target. The resulting training and test sets contain 11794 and 1830 examples, respectively. Ethics approval, privacy safeguards, and data-access arrangements are described in the Ethics Statement.

Implementation. Psy-MARL uses Qwen3-8B (Yang et al., 2025) with frozen backbone weights and independent LoRA (Hu et al., 2021) adapters (rank 32, $\alpha = 64$), initialized from role-specific SFT checkpoints. Qwen3.5-27B (Qwen Team, 2026) provides reference-similarity feedback during training. Psy-MARL and its trained ablation variants use $K = 4$ candidates per fixed-prefix GRPO group and are trained for two epochs. Training and decoding settings are provided in Appendix A.4.

Evaluation. We evaluate reference-response alignment and clinical-rubric preference using GPT-5.6 Luna (OpenAI, 2026). The recorded physician response serves as the reference for alignment evaluation. Each response pair is evaluated in both A/B orders, with outcomes determined separately for each order and aggregated over valid judgments. We report win rates (Win, %) and net-win margins (Net = Win minus Loss, in percentage points), with positive Net favoring the first-named method. To assess the clinical appropriateness of generated responses, we define three task-specific evaluation dimensions informed by psychiatric interviewing principles (Shea, 2016) and Motivational Interviewing (Miller & Rollnick, 2023): HDE (hypothesis-driven exploration), MEO (motivational engagement and openness), and SSA (stage-specific alignment). Dimension scores are comparison-conditioned means on the original 0–100 scale. BLEU-2 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) provide auxiliary lexical-overlap measures. Rubric rationale, evaluation prompts, and aggregation procedures are detailed in Appendix A.6.

Baselines. All baselines use the same backbone and data split. Base Single and Base MAS are frozen, prompt-only single-agent and role-decomposed configurations, respectively. SFT Single is trained on the full training set, whereas SFT MAS uses 500 training examples with LLM-annotated Belief and Strategy targets and recorded physician responses. The resulting role-specific checkpoints initialize Psy-MARL and both ablation variants; annotation details are provided in Appendix A.4. We additionally compare with task-adapted MAPoRL (Park et al., 2025), MARFT (Liao et al., 2025), and MADA-RL (Pulici et al., 2026), whose implementations are described in Appendix A.5.

Table 1: Reference-response alignment. Pairwise comparison directions follow the column headings; positive Net (Win−Loss) favors the first-named method. BLEU-2 and ROUGE-L are auxiliary lexical-overlap metrics. "–" denotes self-comparison.

Method	BLEU-2	ROUGE-L	Ours vs. Method		Method vs. Base Single	
			Win (%)	Net (pp)	Win (%)	Net (pp)
Base Single	6.64	14.69	44.77	+21.72	–	–
Base MAS	6.09	13.64	43.51	+14.42	42.10	+19.40
SFT Single	**7.74**	**17.40**	49.91	+20.15	39.63	+14.15
SFT MAS	7.37	16.76	46.61	**+24.44**	36.80	+11.21
Ours	7.48	17.21	–	–	**44.77**	**+21.72**

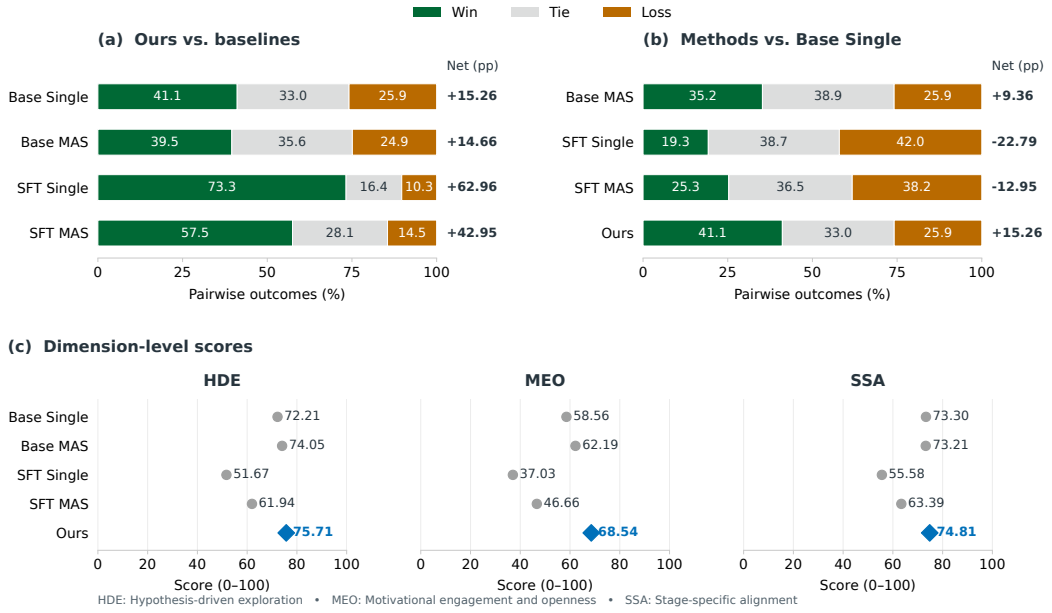

Figure 2: LLM-based clinical evaluation. (a) Psy-MARL versus each baseline. (b) Each method versus Base Single. Wins favor the first-named method; bars show win, tie, and loss percentages, with Net equal to Win−Loss in percentage points. (c) Comparison-conditioned HDE, MEO, and SSA means on a 0–100 scale. Each baseline's means come from its comparison with Psy-MARL; Psy-MARL pools its scores across the baseline comparisons shown.

4.2 Main Results

Alignment with real-world physician responses. Base MAS achieves a $+19.40$ net-win margin over Base Single without parameter updates (Table 1), supporting the role-decomposed configuration. SFT Single obtains the highest BLEU-2 and ROUGE-L, indicating stronger lexical overlap with the recorded physician responses. However, Psy-MARL achieves a positive net-win margin in their direct reference-alignment comparison, where the judge also considers semantic similarity and content coverage. Psy-MARL obtains the largest net-win margin against Base Single ($+21.72$) and a $+24.44$ margin over SFT MAS. Because Psy-MARL starts from the SFT MAS checkpoints and retains the same role structure, the latter comparison supports the benefit of subsequent joint RL, which uses feedback on the final response to optimize the collaborating agents.

Evaluation on clinical metrics. Psy-MARL achieves positive clinical net-win margins against all baselines (Figure 2), including $+14.66$ against Base MAS and $+42.95$ against its SFT MAS initialization. The larger margin against SFT MAS is consistent with the comparisons against Base Single: Base MAS achieves a positive clinical net-win margin, whereas both SFT baselines have negative margins. Together with their higher lexical overlap, this suggests that SFT may favor reproduction of familiar response patterns without reliably adapting them to the current interview

Table 2: Comparison with multi-agent RL baselines. Pairwise results report **Psy-MARL vs. the row method**; positive Net favors Psy-MARL. Dimension means come from these comparisons, with Psy-MARL's scores pooled across the listed opponents.

	Reference alignment		Clinical		Dimension means (0–100)		
Method	Win (%)	Net (pp)	Win (%)	Net (pp)	HDE	MEO	SSA
MAPoRL	42.30	+24.65	51.17	+25.83	65.68	63.15	66.19
MARFT	51.35	+28.40	59.91	+27.19	67.35	56.45	71.07
MADA-RL	36.12	+7.89	34.28	+6.36	71.76	62.14	70.21
Psy-MARL	–	–	–	–	**74.91**	**69.45**	**76.13**

context. Psy-MARL also achieves the highest reported means across HDE, MEO, and SSA. Among the reported aggregate dimension means, the largest separation occurs on MEO: Psy-MARL scores 68.54, while Base MAS has the highest baseline mean of 62.19. These clinical gains suggest that joint RL may help the agents jointly reason about patient states and communication strategies using dialogue context, leading to responses that better fit the psychiatric interview situation.

4.3 Comparison with Multi-Agent RL Baselines

We compare Psy-MARL with task-adapted MAPoRL, MARFT, and MADA-RL. Table 2 reports the results; adaptation details are provided in Appendix A.5.

Comparison with MAPoRL and MADA-RL. MAPoRL adaptation coordinates homogeneous response agents through candidate exchange and revision, while MADA-RL trains critics to revise responses after observing three generator outputs. Psy-MARL explicitly separates patient-state assessment, communication planning, and response generation, making what to explore and how to engage the patient explicit inputs to the response policy. It achieves positive margins against both methods under both evaluation protocols, including clinical net-win margins of $+25.83$ and $+6.36$, respectively. MADA-RL is the closest competitor in these comparisons; access to multiple candidate formulations may facilitate its response refinement. Psy-MARL's positive margins support its complete design, which jointly trains the intermediate assessment and planning roles using feedback on the final response.

Comparison with MARFT. MARFT adaptation uses the same sequential role structure and SFT initialization as Psy-MARL, but trains the agents with PPO, three role-specific value estimators, and a shared terminal reward. Psy-MARL achieves net-win margins of $+28.40$ in reference alignment and $+27.19$ in clinical evaluation. MARFT computes each role's advantage using its value estimate and the shared reward, whereas Psy-MARL assigns separate rewards based on estimated marginal contributions across policy coalitions. This differentiated feedback may help coordinate the agents' updates according to their different effects on the final response. Differences between PPO and GRPO may also contribute to the performance gap; the Shared-Reward Ablation examines reward allocation more directly within the same GRPO framework.

4.4 Ablation Study

Ablation settings. Both ablations share Psy-MARL's role structure, SFT initialization, GRPO setup, and reward scaling and clipping. Shared-Reward Ablation retains pairwise–BT utility estimation but broadcasts the utility of the full coalition $\{B, S, R\}$ to all agents. Direct-Score Ablation retains Shapley allocation but uses independent 1–5 reference-similarity ratings from the same Qwen3.5-27B judge, linearly normalized to $[0, 1]$, as coalition utilities. Base MAS provides an untrained reference. Configuration details are given in Appendix A.4.

Effect of role-specific credit assignment. Psy-MARL achieves net-win margins of $+9.66$ in reference alignment and $+12.40$ in clinical evaluation against Shared-Reward Ablation (Table 3). The reported MEO means also favor Psy-MARL (66.49 vs. 53.51), indicating better engagement and openness under the clinical rubric. These results suggest that differentiated rewards provide more

Table 3: Ablation results. Pairwise results report **Psy-MARL vs. the row variant**; positive Net favors Psy-MARL. Dimension means come from these comparisons, with Psy-MARL's scores pooled across the listed opponents. The Base MAS comparison is reused from the main evaluation.

	Reference alignment		Clinical		Dimension means (0–100)		
Variant	Win (%)	Net (pp)	Win (%)	Net (pp)	HDE	MEO	SSA
Base MAS	43.51	+14.42	39.54	+14.66	74.05	62.19	73.21
Shared-Reward Ablation	38.29	+9.66	41.20	+12.40	74.20	53.51	72.42
Direct-Score Ablation	57.39	+23.68	45.88	+15.10	70.43	50.80	68.60
Psy-MARL (Ours)	–	–	–	–	**76.53**	**66.49**	**75.10**

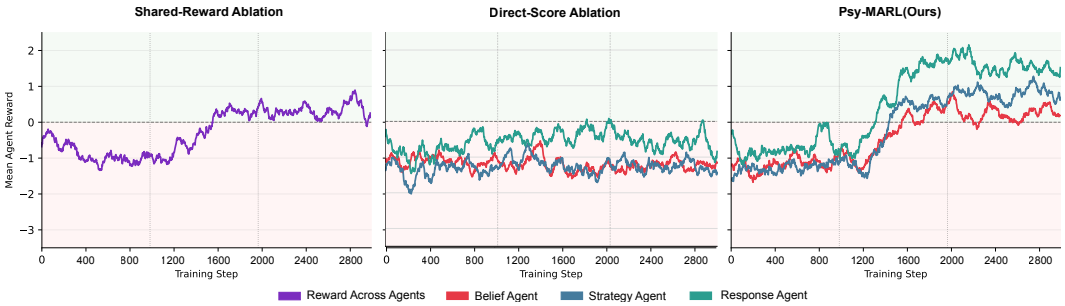

Figure 3: Training reward dynamics. Curves show mean training rewards: a shared team reward for Shared-Reward Ablation and role-specific rewards for Direct-Score Ablation and Psy-MARL. Absolute reward magnitudes are not directly comparable across configurations because reward definitions and utility scales differ.

targeted guidance for joint training, linking updates of the assessment, planning, and response roles to their estimated contributions to the final reply.

Effect of preference-based utility estimation. Psy-MARL achieves net-win margins of $+23.68$ in reference alignment and $+15.10$ in clinical evaluation against Direct-Score Ablation. These gains support preference-based coalition-utility estimation when Shapley allocation is retained. Independent ratings may not reliably distinguish similar coalition responses, obscuring differences relevant to marginal credit. Pairwise judgments explicitly compare these alternatives, while BT fitting integrates the preferences into comparable utilities within each example. This may make differences between coalition responses more informative for role-specific reward allocation.

Reward dynamics. Both Shapley-based variants exhibit role-wise reward separation, with the Response Agent generally receiving higher mean rewards (Figure 3). Psy-MARL's three role means rise during the middle of training and remain above their early-training levels, whereas Direct-Score Ablation shows fluctuations without sustained late-training growth. These trajectories describe how the assigned rewards evolve and differ across roles. The benefit of preserving role-specific feedback for joint training is supported by Psy-MARL's final evaluation gains over Shared-Reward Ablation, rather than by differences in absolute reward levels across configurations.

5 Conclusion

To address the complex reasoning demands of psychiatric interviewing, Psy-MARL combines a multi-agent architecture guided by Theory of Mind and Motivational Interviewing with preference-based coalition credit assignment. This enables feedback on final responses to guide the joint optimization of patient-state assessment, communication planning, and response generation. Experiments show positive net-win margins in reference-response alignment and clinical-rubric preference, while ablation studies support the benefits of preference-based coalition utility estimation and role-specific credit allocation.

AI USE STATEMENT

During the preparation of this manuscript, a large language model was utilized to aid in polishing the grammar and improving the clarity of the text. LLMs were also used to annotate intermediate targets for supervised initialization and to provide training and evaluation feedback, as described in the experimental setup and appendix. The authors reviewed and edited all outputs to ensure the final content accurately reflects our original ideas and are fully responsible for all statements and conclusions presented.

ETHICS STATEMENT

The study was approved by the Institutional Review Board of [*** Hospital] (Approval No. KYS-***). The institution's name and approval number are withheld for double-blind review. The collection and use of the clinical data for model training and evaluation in this study were approved by the collaborating hospital's institutional review board. Written informed consent was obtained from legal guardians, and assent was obtained from participating minors. Transcripts were manually reviewed, and personally identifying information was removed or replaced with placeholders before analysis. This study uses de-identified text for offline model training and evaluation and does not involve prospective clinical deployment. The findings do not establish clinical efficacy or support replacing professional psychiatric care.

Data Availability. Upon acceptance, the de-identified corpus will be made available through a controlled-access process, subject to applicable ethics requirements and a data-use agreement. Instructions for requesting access will be provided at that time. Unredacted clinical transcripts will not be released.

REPRODUCIBILITY STATEMENT

The main text and appendix describe the model architecture, preference-based coalition credit assignment, training settings, baseline adaptations, and evaluation procedures. We plan to release the implementation and training configurations to support reproducibility. Data access arrangements are described in the Data Availability statement.

REFERENCES

Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. MAGI: Multi-agent guided interview for psychiatric assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24898–24921. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1278. URL https://aclanthology.org/2025.findings-acl.1278/.

Guanqun Bi, Zhoufu Liu, Zhuang Chen, Dazhen Wan, Xiyao Xiao, and Minlie Huang. S^4: Operationalizing speech act theory for strategic semi-structured psychiatric interview. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38881–38899. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1802. URL https://aclanthology.org/2026.acl-long.1802/.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Jieming Cao, Chen Huang, Yanan Zhang, Ruibo Deng, Jincheng Zhang, and Wenqiang Lei. Breaking the stigma! unobtrusively probe symptoms in depression disorder diagnosis dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 182–200, 2025.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026a.

Zhuang Chen, Guanqun Bi, Wen Zhang, Jiawei Hu, Aoyun Wang, Xiyao Xiao, Kun Feng, and Minlie Huang. Unveiling the landscape of clinical depression assessment: from behavioral signatures to psychiatric reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1748–1756, 2026b.

Lang Feng, Longtao Zheng, Shuo He, Fuxiang Zhang, and Bo An. Stronger-MAS: Multi-agent reinforcement learning for collaborative LLMs. *arXiv preprint arXiv:2602.08847*, 2026.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3123–3128, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1421/.

Yiyang Gu, Yougen Zhou, Qin Chen, Ningning Zhou, Jie Zhou, Aimin Zhou, and Liang He. Enhancing depression-diagnosis-oriented chat with psychological state tracking. In Xian-Ling Mao, Zhaochun Ren, and Muyun Yang (eds.), *Natural Language Processing and Chinese Computing*, volume 16104 of *Lecture Notes in Computer Science*, pp. 107–119, Singapore, 2026. Springer. doi: 10.1007/978-981-95-3349-7_9. URL https://doi.org/10.1007/978-981-95-3349-7_9.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Hongru Ji, Yuyin Fan, Meng Zhao, Xianghua Li, Lianwei Wu, and Chao Gao. STRIDE-ED: A strategy-grounded stepwise reasoning framework for empathetic dialogue systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7088–7102. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.321. URL https://aclanthology.org/2026.acl-long.321/.

Bowen Jiang, Taiwei Shi, Ryo Kamoi, Yuan Yuan, Camillo J Taylor, Longqi Yang, Pei Zhou, and Sihao Chen. One model, all roles: Multi-turn, multi-agent self-play reinforcement learning for conversational social intelligence. *arXiv preprint arXiv:2602.03109*, 2026.

Onno P Kampman, Ye Sheng Phang, Stanley Han, Michael Xing, Xinyi Hong, Hazirah Hoosainsah, Caleb Tan, Genta Indra Winata, Skyler Wang, Creighton Heaukulani, et al. A multi-agent dual dialogue system to support mental health care providers. *arXiv preprint arXiv:2411.18429*, 2024.

Kunyao Lan, Bingrui Jin, Zichen Zhu, Siyuan Chen, Shu Zhang, Kenny Q Zhu, and Mengyue Wu. Depression diagnosis dialogue simulation: Self-improving psychiatrist with tertiary memory. *arXiv preprint arXiv:2409.15084*, 2024.

Kunyao Lan, Tianyi Sun, Cong Ming, Binwei Yao, Yanli Ding, Yiming Yan, Yan Li, Chao Luo, Lu Chen, Jianhua Chen, and Mengyue Wu. Towards reliable and empathetic depression-diagnosis-oriented chats. *Science China Technological Sciences*, 68:2120406, 2025. doi: 10.1007/s11431-025-3036-5. URL https://doi.org/10.1007/s11431-025-3036-5.

Yanming Li, Xuelin Zhang, WenJie Lu, Ziye Tang, Maodong Wu, Haotian Luo, Tongtong Wu, Zijie Peng, Hongze Mi, Yibo Feng, et al. Who deserves the reward? SHARP: Shapley credit-based optimization for multi-agent system. *arXiv preprint arXiv:2602.08335*, 2026.

Junwei Liao, Muning Wen, Jun Wang, and Weinan Zhang. MARFT: Multi-agent reinforcement fine-tuning. *arXiv preprint arXiv:2504.16129*, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 139–151. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.8. URL https://aclanthology.org/2024.eacl-long.8/.

William R. Miller and Stephen Rollnick. *Motivational Interviewing: Helping People Change and Grow*. Guilford Press, 4 edition, 2023. ISBN 9781462552795.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7362–7376. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.385. URL https://aclanthology.org/2025.findings-acl.385/.

OpenAI. GPT-5.6 Luna. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed: 2026-09-21.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30215–30248, 2025.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512.

Martino ML Pulici, Cuong Xuan Chu, Evgeny Kharlamov, Zifeng Ding, Volker Tresp, and Yunpu Ma. MADA-RL: Multi-agent debate-aware reinforcement learning for parameter-efficient reasoning in compact models. *arXiv preprint arXiv:2607.18006*, 2026.

Qwen Team. Qwen3.5-27B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-27B. Accessed: 2026-09-21.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Yuyang Sha, Hongxin Pan, Wei Xu, Weiyu Meng, Gang Luo, Xinyu Du, Xiaobing Zhai, Henry HY Tong, Caijuan Shi, and Kefeng Li. MDD-LLM: Towards accuracy large language models for major depressive disorder diagnosis. *Journal of Affective Disorders*, 388:119774, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Lloyd S. Shapley. A value for N-person games. Technical Report P-295, RAND Corporation, 1952. URL https://www.rand.org/pubs/papers/P295.html.

Shawn Christopher Shea. *Psychiatric Interviewing: The Art of Understanding*. Elsevier, 3 edition, 2016. ISBN 9781437716986. URL https://shop.elsevier.com/books/psychiatric-interviewing/shea/978-1-4377-1698-6.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Sichang Tu, Abigail Powers, Stephen Doogan, and Jinho D Choi. TRUST: An LLM-based dialogue system for trauma understanding and structured assessments. *arXiv preprint arXiv:2504.21851*, 2025.

Tianxi Wan, Jiaming Luo, Siyuan Chen, Kunyao Lan, Jianhua Chen, Haiyang Geng, and Mengyue Wu. From medical records to diagnostic dialogues: A clinical-grounded approach and dataset for psychiatric comorbidity. In *International Conference on Learning Representations*, volume 2026, pp. 53029–53057, 2026.

Ancheng Xu, Di Yang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, Wanxin Qiu, Mingchen Ma, Haihong Wu, Bingyu Li, et al. AutoCBT: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling. *arXiv preprint arXiv:2501.09426*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. *arXiv preprint arXiv:2205.11764*, 2022.

Xianrong Yao, Dong She, Chenxu Zhang, Yimeng Zhang, Yueru Sun, Noman Ahmed, Yang Gao, and Zhanpeng Jin. Empathy-R1: A chain-of-empathy and reinforcement learning framework for long-form mental health support. *arXiv preprint arXiv:2509.14851*, 2025.

Qi Zhang, Heajun An, Prakriti Dumaru, Sang Won Lee, Lifu Huang, Pamela J Wisniewski, and Jin-Hee Cho. DeepSAGE: Stage-aware reinforcement learning for structured CBT counseling dialogue. *arXiv preprint arXiv:2608.22615*, 2026.

Yujie Zhao, Lanxiang Hu, Yang Wang, Minmin Hou, Hao Zhang, Ke Ding, and Jishen Zhao. Stronger-MAS: Multi-agent reinforcement learning for collaborative LLMs. In *International Conference on Learning Representations*, volume 2026, pp. 150619–150651, 2026.

A APPENDIX

A.1 AGENT ROLES AND PROMPTS

We provide English translations of the Chinese role prompts for the architecture described in Section 3.1. All agent outputs remain in Chinese.

Belief Agent: Patient-State Assessment

System message.
You are an assistant for assessing the patient's state and perspective in psychiatric interviewing.
Using the available patient information and dialogue history, produce a concise assessment covering:

1. The current concern and the statements relevant to it.
2. The patient's possible emotions, concerns, expectations, and understanding of the situation.
3. The patient's expressed willingness to elaborate and the information that still requires clarification.

Ground interpretations in the dialogue and indicate uncertainty when evidence is insufficient. Do not treat assumptions as established facts. Do not infer concealment or noncooperation solely from brief answers, silence, or disagreement.
Output a concise assessment in Chinese. Do not select a communication strategy, provide treatment advice, or generate a physician response.

User message.
Assess the patient's current state using the following information.
Patient information: {`patient_info`}
Dialogue history: {`history_text`}
Output: Directly provide a concise patient-state assessment in Chinese.

Strategy Agent: Communication Planning

System message.
You are an assistant for planning communication strategies in psychiatric interviewing.
Using the patient information, dialogue history, and patient-state assessment, identify the main objective of the current exchange and select an appropriate communication strategy.
Balance the information that requires clarification with the patient's perspective and willingness to engage. Draw on collaborative communication, respect for autonomy, and context-appropriate techniques such as open-ended exploration, reflective listening, affirmation, summarization, or specific clarification. Ground affirmations in the patient's expressed experiences or efforts. Do not force every technique into each response or introduce change motivation when no relevant goal has been established.
Generate a concise action-and-focus instruction; strategies are not restricted to a predefined set of labels. Output only JSON with a single `strategy` field containing this instruction in Chinese. Do not generate a complete physician response or add explanations.

User message.
Choose a communication strategy using the following information.
Patient information: {`patient_info`}
Dialogue history: {`history_text`}
Patient-state assessment: {`patient_state`}
Output: JSON containing only the `strategy` field, with a brief Chinese communication action and its focus.

Response Agent: Strategy-Conditioned Response Generation

System message.
You are an assistant for generating physician responses in psychiatric interviewing.
Integrate the patient information, dialogue history, patient-state assessment, and communication strategy to produce a natural, concise physician response appropriate to the current exchange.
Implement the selected strategy while remaining grounded in the dialogue. Do not present tentative upstream interpretations as facts to the patient.

Use questions, reflections, affirmations, or brief summaries according to the communicative need. Not every response must include a question. When exploration is needed, use clear, non-leading language; direct questions remain appropriate for specific clarification. Avoid mechanical repetition and repeatedly returning to the same content. If asking a follow-up, focus on one main issue at a time.
Output only the physician response in Chinese. Do not reveal internal assessments, strategy labels, role names, or reasoning. Do not output JSON or additional explanations.

User message.
Formulate a physician response using the following information.
Patient information: {`patient_info`}
Dialogue history: {`history_text`}
Patient-state assessment: {`patient_state`}
Communication strategy: {`strategy`}
Output: Directly provide the physician response in Chinese.

Input and output interfaces. All three agents receive the available patient information and retain access to the original dialogue context. Belief assessments are passed as text. The Strategy Agent's `strategy` field is extracted and passed to the Response Agent; if JSON parsing fails, the generated strategy text is used directly. Reference physician responses are excluded from all agent prompts and supplied only to the scorer.

A.2 Coalition Construction and Utility Fitting

Policy coalitions. Let $\mathcal{N}$ denote the role set, with $n = |\mathcal{N}|$, and let $2^{\mathcal{N}}$ contain all policy coalitions. For each coalition $S \subseteq \mathcal{N}$, members use their current role-specific LoRA adapters, while nonmembers use the frozen Qwen3-8B backbone without an adapter under the same role prompts. Every coalition executes the complete Belief–Strategy–Response sequence, with downstream inputs constructed from the corresponding upstream outputs.

Candidate evaluation and rollout reuse. Each candidate in a role-specific update group is completed into a physician response before reward assignment. Its completed trajectory supplies the full-coalition response, while the remaining coalition responses are obtained through policy substitution. For role i and candidate $k \in \{1, \ldots, K\}$, let $\hat{v}_{h,i,k}$ denote the game fitted from the corresponding coalition comparisons. Its role-specific credit $\phi_i(\hat{v}_{h,i,k})$ is converted into the candidate's training reward using Equation 5. Fixed-prefix group construction is described in Appendix A.4.

Previously sampled outputs may be shared when their role policies and complete conditioning inputs match the required computation. Changes to upstream outputs require regenerating the affected downstream outputs: changing a Belief output can affect both Strategy and Response, while changing a Strategy output affects the subsequent Response. The K candidates within each update group remain separately sampled. Coalition utilities are fitted jointly from the comparison targets associated with each candidate; a reused response can therefore receive different fitted utilities in different candidate evaluations.

Pairwise judgments and utility fitting. We compare adjacent coalitions using the comparison set $\mathcal{D}$ defined in the main text. The number of comparisons is $m = |\mathcal{D}| = n2^{n-1}$. All m response pairs are explicitly listed in a single judge request together with the dialogue context and reference physician response. The judge returns a winner and confidence for each pair.

Let w denote the selected winner and c the reported confidence, clipped to $[p_0, 1]$. The soft preference target for candidate A is

$$q_{ab} = \begin{cases} p_0 + \lambda(c - p_0), & w = A, \\ p_0 - \lambda(c - p_0), & w = B, \\ p_0, & \text{otherwise}, \end{cases}$$

where $p_0 = \frac{1}{2}$ denotes neutral preference and $\lambda \in (0, 1)$ controls the magnitude of the deviation from neutrality. Ties and unavailable or unrecognized judgments receive p_0; for unavailable judgments, this is a computational fallback. The mapping defines soft fitting targets rather than calibrated preference probabilities.

Targets are clipped to $[\epsilon_{\mathrm{BT}}, 1 - \epsilon_{\mathrm{BT}}]$, where $\epsilon_{\mathrm{BT}} \in (0, p_0)$ is a numerical safeguard. We then jointly fit the coalition utilities using Equation 4, fixing $v(\emptyset) = 0$ and adding no regularizer. Utilities are initialized at zero. Optimization uses adaptive-step gradient descent with initial step size η_0 and terminates when the gradient norm falls below δ_g, the step size falls below $\eta_{\min}$, or the iteration limit $L_{\max}$ is reached. Here, δ_g is the gradient tolerance and $\eta_{\min}$ is the minimum step size. Parameter values and step-size settings are reported in Appendix A.4.

Training judge prompts. The following boxes present English renderings of the Chinese training judge prompts. Batch size and confidence bounds are expressed symbolically: {num_pairs} is instantiated as m, and p_0 denotes the neutral confidence bound defined above.

Training Judge: Pairwise Reference Alignment

You are a text-similarity evaluation assistant. For each of the following {num_pairs} response pairs, determine which response is more similar to the reference response.

Dialogue context
{context}

Reference response (ground truth)
{reference}

Response pairs to compare
{response_pairs}

Judgment criteria
Compare only similarity: select the response whose content, wording, and intent are closer to the reference. Do not evaluate response quality; assess only similarity to the reference.

Mandatory rule
Unless A and B in a pair are exactly identical, you must select a winner (A or B) and must not output a tie. Even when the difference is very small, compare carefully and select the response closer to the reference.

Output format
Output only a JSON array containing one object per response pair, with no additional text. Each object must contain:

- pair: the index of the response pair;
- winner: the selected winner;
- confidence: certainty in the judgment, within $[p_0, 1]$.

Training Judge: Direct Similarity Scoring

You are a text-similarity evaluation assistant. Evaluate how similar the candidate response is to the reference response.

Dialogue context
{context}

Reference response (ground truth)
{reference}

Candidate response
{prediction}

Scoring rubric (1–5)

- **5:** The meaning is almost identical; wording and intent closely match the reference.
- **4:** The main content matches, with minor wording differences but consistent intent.
- **3:** Some content is similar to the reference, but clear differences remain.
- **2:** The response has low similarity to the reference, with only limited overlap.
- **1:** The response is almost unrelated to or completely different from the reference.

Evaluate only semantic similarity to the reference, not the quality of the response itself.

Output format
Output only a JSON object containing a single `score` field with a number from 1 to 5.

A.3 Properties of Utility Fitting and Credit Allocation

The following properties apply to a fixed estimated coalition game $\hat{v}$ obtained from a single candidate evaluation.

BT utility fitting. For fixed preference targets, the objective in Equation 4 is convex and invariant to a common additive shift of coalition utilities. With a connected comparison graph, fixing $v(\emptyset) = 0$ removes this ambiguity and makes the objective strictly convex. Interior targets $q_{ab} \in (0, 1)$ ensure that the anchored objective admits a unique finite minimizer. Numerical optimization follows the procedure in Appendix A.2. These properties concern fitting the supplied preferences and do not establish the accuracy of the judge's decisions.

Shapley allocation. The raw Shapley credits computed by Equation 2 satisfy efficiency, symmetry, the null-player property, and additivity (Shapley, 1952). In particular,

$$\sum_{i \in \mathcal{N}} \phi_i(\hat{v}) = \hat{v}(\mathcal{N}) - \hat{v}(\emptyset) = \hat{v}(\mathcal{N}). \tag{7}$$

Thus, the estimated utility gain over the all-reference system is fully allocated among the roles. Symmetry assigns equal credit to roles with identical marginal contributions, null players receive zero credit, and additivity makes the allocation linear in the utility function.

All coalitions are enumerated, so the allocation is exact for the fitted game, while its utilities remain estimates derived from sampled responses and judge feedback. Negative credits are permitted. Player orderings determine the allocation weights without changing the Belief–Strategy–Response execution sequence.

A.4 Training and Implementation Details

SFT annotation. The intermediate Belief and Strategy targets used for role-specific SFT are generated by GPT-5.6 Luna, and 20% of the annotated examples are manually reviewed.

Training settings. RL training uses eight A800 GPUs, with the training judge hosted on a separate two-A800 node. Each role uses an independent AdamW optimizer with a learning rate of 5×10^{-6} and a training batch size of eight dialogue examples. Input and output lengths are capped at 8,192 and 256 tokens, respectively, and thinking mode is disabled for all roles. Training rollouts use temperatures of $(0.7, 0.9, 0.8)$ and top-p values of $(0.90, 0.95, 0.95)$ for the Belief, Strategy, and Response agents, respectively.

Fixed-prefix sampling. For each downstream role, the selected full trajectory supplies one candidate, and the remaining $K - 1$ candidates are sampled with the corresponding upstream prefix held fixed. Each selected rollout batch is used for one optimization pass.

Reward and utility-fitting parameters. Psy-MARL and its two ablation variants use $\tau = 10$ and $r_{\max} = 5$. For preference-based utility estimation, each candidate evaluation includes $m = 12$ adjacent-coalition comparisons, with $\lambda = 0.8$ and $\epsilon_{\mathrm{BT}} = 0.01$. The BT solver starts with $\eta_0 = 0.5$; the step size is multiplied by 1.1 after an accepted update, up to a maximum of 2.0, and halved after a rejected update. The stopping parameters are $\delta_g = 10^{-6}$, $\eta_{\min} = 10^{-10}$, and $L_{\max} = 200$.

A.5 Adaptations of Multi-Agent RL Baselines

All three adapted baselines use the frozen Qwen3-8B backbone and the same training and test splits. We adapt their reward interfaces to physician-response generation using reference-similarity feedback from Qwen3.5-27B. Reference physician responses are accessible only to the training scorer and are excluded from agent inputs.

MAPoRL. We adapt MAPoRL (Park et al., 2025) using homogeneous response agents that share a single LoRA adapter initialized on the base model without prior SFT. We use two agents during training and three during evaluation, with three internal rounds. Agents first generate independent candidates and then revise their responses after observing the preceding round's candidates. Training uses PPO with a learned value estimator and GAE, with policy updates based on second-round responses; the third round is used for subsequent generation without an additional policy-update objective. For agent i, the adapted reward is

$$R_i = s_{\text{sim}}(a_i^{(2)}, y^*) + \frac{\gamma}{M-1} \sum_{j \neq i} s_{\text{sim}}(a_j^{(2)}, y^*), \qquad \gamma = 0.5,$$

where $a_i^{(2)}$ is the agent's second-round response, M is the number of agents participating in training reward computation, and s_{sim} is the judge's similarity score relative to the reference physician response y^*. At inference, each agent votes for one final-round candidate using the dialogue history and candidate texts. Candidate authorship is hidden and presentation order is randomized. The candidate receiving the most votes is returned without rewriting; ties are resolved uniformly at random using a fixed seed. No reference response or external judge is used for selection.

MARFT. We adapt MARFT (Liao et al., 2025) to the same sequential Belief–Strategy–Response workflow as Psy-MARL. The agents use three independent LoRA adapters (rank 32, $\alpha = 64$), initialized from the same role-specific SFT checkpoints as Psy-MARL. The normalized reference-similarity score of the final response supplies a shared terminal team reward. Training uses PPO and GAE with three role-specific value estimators, one for each agent, without BT–Shapley reward redistribution. Although the terminal reward is shared, each role's advantages are computed using its respective value estimates. At inference, the Response Agent produces the final physician response.

MADA-RL. We adapt MADA-RL (Pulici et al., 2026) using its two-stage generator–critic procedure. Three generators and three critics each use an independent LoRA adapter (rank 32, $\alpha = 64$), initialized on the base model without prior SFT. The training set is partitioned into three disjoint subsets for independent generator training with GRPO. Original task rewards are replaced by reference-similarity scores normalized to $[0, 1]$. The trained generators are then frozen and produce three candidate responses for each dialogue history. Each critic receives the history and all three candidates and generates a revised response. The resulting critic-training dataset is partitioned into three disjoint subsets. The mean generator score provides a fixed baseline for each dialogue input:

$$b_G(h) = \frac{1}{3} \sum_{j=1}^{3} s_{\text{sim}}(g_j, y^*), \qquad A_{C,k} = s_{\text{sim}}(c_k, y^*) - b_G(h),$$

where h denotes the complete dialogue input, g_j is the response from generator j, c_k is the k-th revised response sampled by the critic being trained for that input, and s_{sim} denotes the normalized reference-similarity score. The critic advantage is used directly in the clipped policy objective, without within-prompt mean subtraction that would cancel the generator baseline. Both stages use $K = 4$ sampled completions per prompt and require no learned value estimator. At inference, the three generators produce candidates, followed by three critics that independently revise them after observing all generator outputs. One final critic response is selected uniformly at random using a fixed seed, without reference access or judge-based selection.

All discussion and revision rounds are internal computations and do not introduce additional patient turns.

A.6 Evaluation Protocol and Prompts

Clinical evaluation dimensions. To assess clinically relevant response behavior beyond reference alignment, we define three task-specific dimensions informed by psychiatric interviewing principles and Motivational Interviewing (MI). **Hypothesis-driven exploration (HDE)** reflects purposeful clinical information gathering, assessing whether exploration follows clinically relevant cues and hypotheses grounded in the dialogue (Shea, 2016). **Motivational engagement and openness (MEO)** draws on MI's emphasis on collaborative engagement and eliciting the patient's own perspective, assessing whether communication encourages further expression (Miller & Rollnick, 2023). **Stage-specific alignment (SSA)** reflects the dynamic organization of psychiatric interviewing, assessing

whether a response fits the current conversational task, such as alliance-building, assessment, or intervention (Shea, 2016). These dimensions operationalize selected interviewing principles for evaluating final responses in context.

The following boxes provide concise English summaries of the Chinese evaluation prompts, including their scoring criteria and required output fields.

Reference-Response Alignment Evaluation

Evaluate two candidate responses in psychiatric interviewing by comparing each with the reference physician response. Assign each candidate an integer similarity score from 0 to 100.
Scoring criteria. Consider semantic similarity, wording and style, and coverage of the reference response's main points. Evaluate similarity to the reference rather than the intrinsic quality of the response.
Decision rule. The higher-scoring candidate wins. An absolute score difference of at most 5 is a tie.

Dialogue context: {`context`}
Reference physician response: {`reference`}
Candidate A: {`a`}
Candidate B: {`b`}

Output format. Return only JSON containing `score_a`, `score_b`, `winner`, and `reason`. Scores are integers from 0 to 100; `winner` is `"A"`, `"B"`, or `"tie"`.

Clinical-Rubric Evaluation

Evaluate two candidate physician responses in child and adolescent psychiatric interviewing. Score each candidate separately on the following dimensions, using integer scores from 0 to 100.
Hypothesis-driven exploration (HDE). High scores indicate targeted exploration of core symptoms, risk indicators, or cognitive biases disclosed by the patient or guardian, guided by clinical hypotheses. Low scores indicate generic small talk or exploration unrelated to the clinical concerns.
Motivational engagement and openness (MEO). High scores indicate open-ended exploration that encourages the patient to elaborate. Low scores indicate communication that restricts elaboration or uses leading language.
Stage-specific alignment (SSA). High scores indicate alignment with the current interview stage, such as alliance-building, assessment, or intervention, and appropriate progression toward its clinical goals. Low scores indicate a stage mismatch or discontinuity with the dialogue context.
Decision rule. Sum the three dimension scores for each candidate. The candidate with the higher total wins; an absolute total-score difference of at most 10 is a tie.

Dialogue context: {`context`}
Candidate A: {`a`}
Candidate B: {`b`}

Output format. Return only JSON containing `dim1_a` and `dim1_b` for HDE, `dim2_a` and `dim2_b` for MEO, and `dim3_a` and `dim3_b` for SSA. Include `total_a`, `total_b`, `winner`, and `reason`. Each total is the sum of the corresponding candidate's three dimension scores; `winner` is `"A"`, `"B"`, or `"tie"`.

Outcome aggregation. Each response pair is evaluated in both A/B orders. Outcomes are determined separately for each order using the thresholds above and mapped back to model identities before aggregation. Win, tie, and loss rates are computed with equal weight over valid order-specific judgments; failed or invalid judgments are excluded. If only one order yields a valid judgment, that judgment is retained. Scores are not averaged across orders before determining outcomes. The two order-specific judgments belong to the same test example.

Dimension-score aggregation. HDE, MEO, and SSA means are computed from valid scores mapped back to model identities, including judgments with win, tie, and loss outcomes. Each baseline's reported means come from its comparison with Psy-MARL. Within each table or figure, Psy-MARL's means pool its valid scores across comparisons with all opponents shown. These comparison-conditioned means may therefore differ across evaluation blocks. All dimension scores are reported on their original 0–100 scale.

A.7 Training Overhead

Coalition-based credit assignment introduces additional training overhead through response generation under multiple policy coalitions and LLM-based preference evaluation. We reduce redundant computation in two ways. First, all adjacent-coalition comparisons within each candidate evaluation are submitted in a single judge request, sharing the dialogue context and reference response. Second, compatible sampled prefixes and coalition responses are reused, while downstream computations affected by changed inputs are regenerated. The reuse conditions are detailed in Appendix A.2.

These measures reduce repeated generation and evaluation requests but do not eliminate the additional cost of coalition evaluation. Coalition evaluation and reward computation are confined to training; at inference, Psy-MARL generates responses through the Belief–Strategy–Response workflow without coalition evaluation or reward judging.