MIND: An Evidence-State Decision Interface for Criteria-Grounded Psychiatric Inquiry and Diagnosis

Anonymous ACL submission

Abstract

Psychiatric consultation is a multi-turn dialogue task in which agents must elicit discriminative evidence, map uncertain patient narratives to diagnostic criteria, and decide when evidence suffices for screening. Existing dialogue and retrieval-augmented systems condition the policy on raw histories or appended passages, leaving observed evidence, unresolved criterion checks, active differentials, and support reliability entangled across turns. We introduce MIND, an evidence-state decision interface for criteria-grounded psychiatric inquiry. MIND represents each turn as a typed evidence state over observed evidence, missing checks, differentials, criterion-linked support, and reliability metadata, and makes this state the shared object for action selection, process rewards, information-gain scoring, and trajectory rectification. A training-split Psychiatric Reasoning Bank turns retrieval into state construction through typed role assignment and reliability gating, rather than unstructured prompt injection. Under EMR-grounded simulator protocols, MIND improves accuracy by 8.8 and 7.3 points over strong inference-only, RAG-based, and RL-based baselines, with gains transferring to public MDD-5k diagnostic dialogues. Matched controlled interventions show that the gains come from the shared evidence-state interface rather than longer prompts, retrieval text, or formatting alone. These results suggest that explicit evidence-state interfaces are an effective abstraction for long-horizon, criteria-grounded medical dialogue.

1 Introduction

Psychiatric consultation is a demanding form of multi-turn medical dialogue: the doctor must elicit decisive evidence from subjective, incomplete narratives, weigh competing diagnoses against criteria such as DSM-5, and decide when to commit to a screening category (Organization et al., 2022; Yin et al., 2025; Feng et al., 2026). Unlike symptom-driven dialogue, relevant cues are often metaphorical or temporally fragmented, comorbidity is common, and decisive signals—functional impairment, exclusion checks, safety risks—are easy to miss (Lamers et al., 2011). An effective agent must therefore track at every turn *what is known*, *what is still missing*, and *which alternative explanations remain live*.

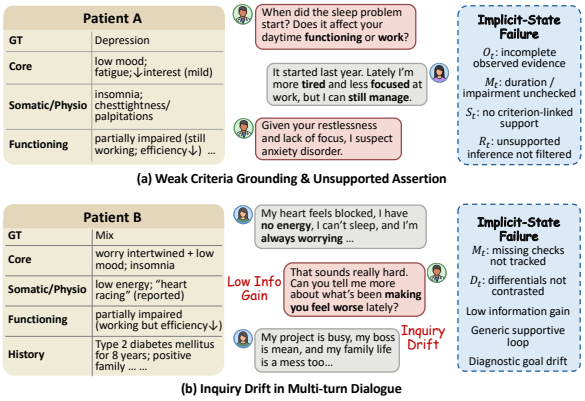

Figure 1: Failure modes of implicit clinical state. (a) Without typed evidence, missing checks, and reliability gating, the model makes unsupported assertions. (b) Without tracked missing checks and active differentials, inquiry drifts to generic low-yield questions.

Recent work attacks this problem from several angles: sequential decision methods that reduce diagnostic uncertainty (Zhu et al., 2025; Winston et al., 2024), retrieval-augmented methods that ground responses in external clinical knowledge (Sun et al., 2025; Jeong et al., 2024), and joint inquiry–diagnosis frameworks that couple question selection with label prediction (Jia et al., 2025b; Dou et al., 2025; Aich et al., 2025). In nearly all of them the policy still conditions on the raw dialogue history H_t, optionally augmented with retrieved passages, so which criteria are already satisfied, which are still uncertain, which differentials are active, and how reliable each retrieved support is

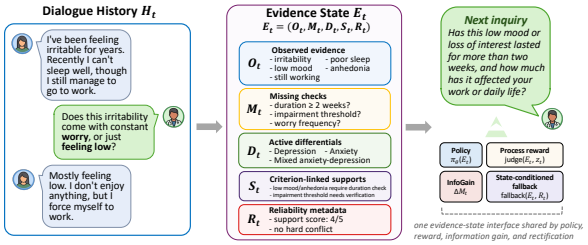

Figure 2: MIND in action. (a) Criteria-grounded PRB retrieval drives informative follow-ups. (b) State-conditioned rectification curbs inquiry drift.

must be re-derived from text each turn. Two failure modes recur: confident assertions not traceable to the patient report (Figure 1(a)), and inquiry that drifts into generic or repeated questions instead of cues separating competing diagnoses (Figure 1(b)).

We attribute these failures to using H_t as the policy state: the quantities a consultation agent decides over—observed evidence, missing checks, active differentials, and the reliability of external support—are entangled in an unstructured token sequence. We instead introduce an *evidence-state decision interface* (ESDI): a typed, reusable control state that replaces raw dialogue history as the operational decision object. ESDI constructs

$$H_t \xrightarrow{f_\phi} \mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t), \qquad \mathcal{E}_t \to \{\pi_\theta, R_{\text{proc}}, R_{\text{info}}, \Gamma_{\text{recover}}\}, \tag{1}$$

where $\mathcal{O}_t$ records patient-reported evidence, $\mathcal{M}_t$ unresolved criterion and exclusion checks, $\mathcal{D}_t$ active differential hypotheses, $\mathcal{S}_t$ criterion-linked support entries, and $\mathcal{R}_t$ reliability metadata for those supports. $\mathcal{E}_t$ is not a one-off summary or a prompt template: it is the shared object that the policy, process-reward judge, information-gain scorer, and trajectory-recovery module read and write, so retrieval, criterion checking, reliability assessment, and fallback all act as operators on the same state.

We instantiate this view in MIND, a criteria-grounded reinforcement learning framework for psychiatric inquiry and diagnosis. MIND builds a Criteria-Grounded Psychiatric Reasoning Bank (PRB) from training-split consultations and clinical criteria, and uses reliability-gated PRB retrieval to populate $\mathcal{S}_t$ and $\mathcal{R}_t$; each retrieved item is admitted only after being mapped to a typed role in $\mathcal{E}_t$, and the admitted fields are reused with identical semantics by the policy, reward judge, information-gain scorer, and fallback. Training combines terminal diagnostic rewards with rubric-based process rewards over $\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t$, and a state-conditioned rectification module triggers self-retry or PRB-guided fallback when $\mathcal{R}_t$ or repetition checks indicate a low-yield turn. To our knowledge, this provides a controlled study showing that sharing a typed clinical evidence state across policy learning, reward computation, information-gain scoring, and trajectory recovery improves psychiatric inquiry.

The contributions of this paper are as follows: (1) we introduce an evidence-state decision interface for psychiatric inquiry, replacing raw dialogue history with typed clinical decision variables shared by policy action selection, process rewards, information-gain estimation, and recovery; (2) we operationalize retrieval as state construction rather than prompt augmentation, where PRB supports are admitted only after typed role assignment and reliability gating; (3) we provide controlled empirical evidence that MIND's gains come from the shared typed evidence state rather than prompt formatting or retrieval volume alone, with matched comparisons and representation-level controls showing improvements in diagnostic accuracy, inquiry quality, and support faithfulness.

2 Problem Definition

We formulate psychiatric consultation as a budgeted task-oriented dialogue $\mathcal{P} = \{X^{ex}, X^{im}, d^*\}$, with explicit cues X^{ex}, latent cues X^{im} elicited through questioning, and gold label d^*. At turn t, the agent observes history H_t, takes an inquiry or diagnosis action a_t, and receives a patient response; the episode terminates on a diagnosis or at turn budget L, yielding $\tau = (H_1, a_1, \ldots, H_T)$ with $T \leq L$ and

$$d_p = \pi_\theta^{\text{diag}}(H_T), \qquad T \leq L. \tag{2}$$

The task therefore requires both eliciting diagnostically useful evidence within L turns and mapping the accumulated state to d^*.

Evidence-state MDP formulation. Instead of conditioning the policy directly on H_t, we treat psychiatric inquiry as an evidence-state MDP in which the operational state is constructed before action selection:

$$H_t \xrightarrow{f_\phi} \mathcal{E}_t, \quad a_t \sim \pi_\theta(\mathcal{E}_t, H_t), \quad r_t = R(\mathcal{E}_t, a_t, \mathcal{E}_{t+1}). \tag{3}$$

$\mathcal{E}_t$ is the shared decision interface: the policy reads all typed fields, process rewards score field-consistent reasoning, information gain is defined

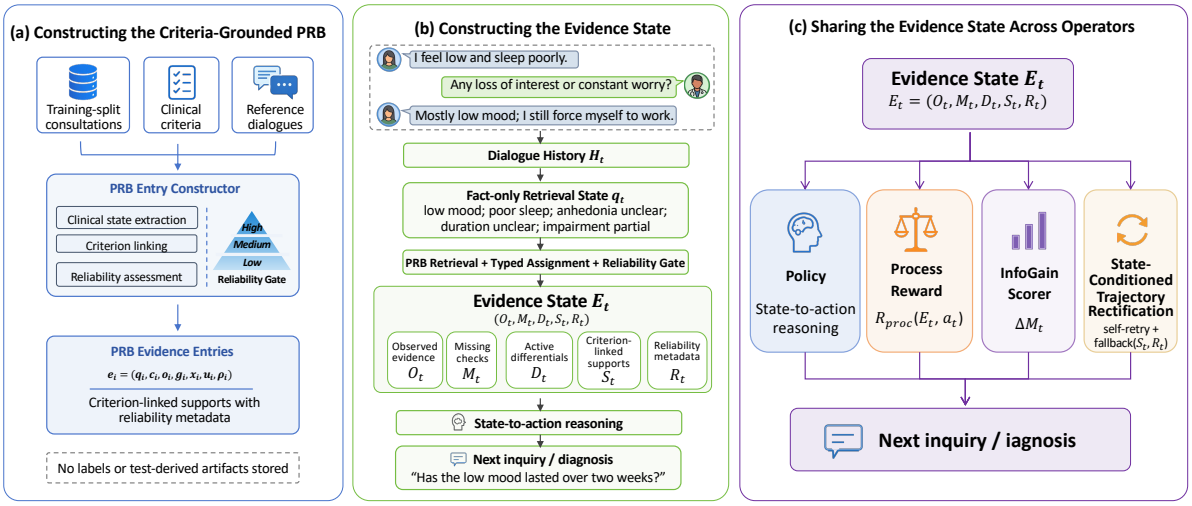

Figure 3: MIND architecture. (a) PRB is built offline from training-split consultations as label-free typed entries $e_i = (q_i, c_i, o_i, g_i, x_i, u_i, \rho_i)$. (b) Each turn maps H_t to $\mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t)$, with q_t filling observed/missing/differential fields and reliability-gated PRB retrieval filling support/reliability fields. (c) Policy, process reward, information gain, and rectification all read $\mathcal{E}_t$.

over the reduction of $\mathcal{M}_t$, and recovery triggers depend on $\mathcal{R}_t$, repetition, and $\mathcal{M}_t$ stagnation. Because individual fields can be masked, shuffled, or gated at inference time, this formulation exposes representation-level intervention points unavailable when the state is raw text.

3 Methodology

MIND realizes the evidence-state interface introduced above (Figure 3). At turn t, a constructor maps H_t to

$$\mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t), \tag{4}$$

where $\mathcal{O}_t$ holds observed patient facts, $\mathcal{M}_t$ unresolved criterion/exclusion checks, $\mathcal{D}_t$ active differentials, $\mathcal{S}_t$ PRB-retrieved criterion-linked supports, and $\mathcal{R}_t$ their reliability scores and conflict flags. The policy, reward judge, and recovery operator all consume $\mathcal{E}_t$, so the same typed object drives action selection, process supervision, and trajectory rectification. The design has three components: (i) a *Criteria-Grounded Psychiatric Reasoning Bank* (PRB) that populates $(\mathcal{S}_t, \mathcal{R}_t)$ from training-split consultations and clinical criteria; (ii) a retrieval-conditioned policy that performs state-to-action reasoning over $\mathcal{E}_t$; and (iii) SFT→RL training with terminal rewards, process rewards, and state-conditioned rectification.

Why a typed state rather than retrieved text. H_t plus retrieved passages entangles four control variables—observed evidence, missing checks, active differentials, and retrieval reliability—and forces the policy to re-derive their decomposition as the history grows. Typed factoring makes these variables explicit and keeps $|\mathcal{E}_t|$ bounded while $|H_t|$ grows with T. Section 4.1 reports a backbone-, retrieval-, and reward-matched comparison (Table 5) isolating this interface from prompt formatting and retrieval content.

Evidence-state decision interface. A framework satisfies the *evidence-state decision interface* if: (i) it constructs a typed state

$$\mathcal{E}_t = f_\phi(H_t, \mathrm{PRB}) = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t), \tag{5}$$

from history and admissible retrieval; (ii) each field has a distinct decision role; (iii) action selection, reward, information gain, and recovery all read the same state,

$$\begin{aligned} &a_t \sim \pi_\theta(\mathcal{E}_t, H_t), \quad r_t = R(\mathcal{E}_t, a_t, \mathcal{E}_{t+1}), \\ &\tilde{a}_t = \Gamma_{\mathrm{recover}}(\mathcal{E}_t, a_t); \end{aligned} \tag{6}$$

and (iv) retrieved supports affect decisions only after typed-field assignment and reliability gating. $\mathcal{E}_t$ is therefore an explicit *intervention point*: replacing it with raw history or untyped retrieval under otherwise matched training changes the decision process (Table 5).

Operational semantics. Sharing $\mathcal{E}_t$ across policy, reward, and recovery distinguishes the interface from a structured prompt template, which only organizes text for a single generation call. The

Table 1: Decision-object comparison: MIND moves the policy/reward/recovery shared object from raw or retrieved text to a typed evidence state $\mathcal{E}_t$. Columns: how evidence enters the loop, the reward variable, and the recovery variable. Each row lists 1–2 anchor works.

Method type (anchor works)	State object	Retrieval role	Reward object	Recovery object
Raw-history RL (Feng et al., 2026; Jia et al., 2025b)	Dialogue history H_t	None	Raw trajectory	None or rules
Standard RAG (Asai et al., 2024)	H_t plus passages	Prompt context	Raw trajectory	None
Structured prompt (Chen et al., 2024)	Labeled text sections	Formatted context	Raw trajectory	None
MIND (ours)	Typed evidence state $\mathcal{E}_t$	State construction	Evidence-state reward	Evidence-state recovery

Structured-prompt RAG RL variant in Table 5 uses the same typed layout in the policy prompt but does not share fields with the reward or fallback. Two medically trained annotators validate f_ϕ on 100 turns, reporting 91.8% partial-or-better field accuracy (77.4% fully correct, Cohen's $\kappa = 0.75$).

Complexity argument. Since $\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t$ live in a fixed schema and $|\mathcal{S}_t|$ is bounded by retrieval depth k, $|\mathcal{E}_t|$ stays roughly constant while $|H_t|$ grows linearly in T, shrinking the effective hypothesis space on long-horizon dialogues—consistent with the larger PsySim-Adapt drop when PRB, reliability gating, or thinking supervision is removed (Table 6).

3.1 PRB as a State-Construction Operator

PRB is a state-construction operator, not a knowledge bank, answer memory, or generic case retriever: it instantiates the support and reliability fields of $\mathcal{E}_t$ from the current retrieval state. Entries are built offline with an off-the-shelf LLM, but the contribution is the *typed schema and operational contract*: (i) each entry is decomposed into criterion-indexed fields $(c_i, o_i, g_i, x_i, u_i, \rho_i)$ rather than free-form text; (ii) candidates influence the policy only after reliability gating and typed-slot assignment; and (iii) the same entry is reused with role-consistent semantics across action selection, reward, information-gain scoring, and fallback. Swapping the construction LLM leaves downstream performance nearly unchanged (Table 14), indicating the gain comes from the interface contract rather than the extractor.

Formally,

$$g_{\mathrm{PRB}}(q_t) \to (\mathcal{S}_t, \mathcal{R}_t), \tag{7}$$

and the full evidence state combines g_{PRB} with fields parsed from q_t:

$$f_\phi(H_t,\mathrm{PRB}) = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, g^{\mathcal{S}}_{\mathrm{PRB}}(q_t), g^{\mathcal{R}}_{\mathrm{PRB}}(q_t)). \tag{8}$$

Each PRB entry is a typed record

$$e_i = (q_i, c_i, o_i, g_i, x_i, u_i, \rho_i), \tag{9}$$

with q_i a compact retrieval state, c_i the associated criterion or exclusion check, o_i the patient-evidence anchor, g_i the missing or unclear check, x_i exclusion/safety cues, u_i a next-inquiry rationale, and ρ_i reliability scores and hard-issue flags. PRB is built only from the training split—no test EMRs, test-derived states, or test support notes—and entries omit diagnostic labels, so it cannot act as an answer memory. Unlike standard RAG, PRB retrieval populates typed fields of $\mathcal{E}_t$ that are reused across inquiry, diagnosis, process reward, and rectification.

3.1.1 Clinical Retrieval State

For each training-split dialogue, the pipeline distills a fact-only retrieval state q_i that explicitly marks missing fields as "not mentioned" or "unclear" and covers chief complaint, duration, severity, sleep, impairment, safety risk, psychosis/mania cues, stressors, and substance use. q_i is then parsed into observed-evidence anchors o_i and missing-check candidates g_i, so retrieval updates $\mathcal{O}_t$ and $\mathcal{M}_t$ rather than injecting an undifferentiated paragraph.

3.1.2 Clinical Support Synthesis

Given q_i, the pipeline synthesizes criterion-indexed fields from the dialogue and clinical references: c_i names the criterion or exclusion check, x_i flags safety/exclusion cues, and u_i records the next-inquiry rationale. These can be rendered as a natural-language note r_i, but the policy consumes them through their typed roles in $\mathcal{E}_t$. The LLM acts only as an offline preprocessor; the policy never queries it at inference time.

3.1.3 Reliability Assessment

An LLM judge produces ρ_i from dialogue consistency, reference relevance, and any unsupported clinical content, yielding a 1–5 score and hard-issue flags. Only entries above the reliability threshold enter $\mathcal{S}_t$ and $\mathcal{R}_t$; otherwise MIND falls back to null support.

3.2 Explicit Clinical Reasoning with Process Supervision

Each turn proceeds in two retrieval-conditioned steps: build a retrieval state for PRB lookup, then produce support-conditioned state-to-action reasoning under process supervision.

3.2.1 Step 1: Retrieval-State Construction and PRB Lookup

Given H_t, the policy first verbalizes the observable and missing-check fields of $\mathcal{E}_t$ as a retrieval-oriented query state:

$$y_t^{(1)} = \texttt{<think>}\ z_t^{(1)}\ \texttt{</think>}\ \texttt{<rag_query>}\ q_t\ \texttt{</rag_query>}. \tag{10}$$

The trace $z_t^{(1)}$ identifies facts, missing fields, and retrieval focus, and is excluded from the core process reward. The resulting q_t approximates $(\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t)$ in text and drives PRB lookup. We retrieve the top-k entries by semantic similarity,

$$\begin{aligned} &\mathrm{PRB}(q_t) = \mathrm{Top}\text{-}k(\cos(q_t, q_i)), \\ &\widetilde{\mathcal{S}}_t = \{(c_i, o_i, g_i, x_i, u_i)\}_{i=1}^{k}, \end{aligned} \tag{11}$$

and only candidates above the reliability threshold form $\mathcal{S}_t$, with their scores and hard-issue flags forming $\mathcal{R}_t$; otherwise the policy uses null support.

3.2.2 Step 2: State-to-Action Reasoning and Response Generation

Conditioned on $(H_t, \mathcal{E}_t)$, the policy emits a reasoning trace and the next action:

$$y_t^{(2)} = \texttt{<think>}\ z_t^{(2)}\ \texttt{</think>}\ \texttt{<answer>}\ a_t\ \texttt{</answer>}. \tag{12}$$

The trace $z_t^{(2)}$ links observed evidence, missing checks, active alternatives, and reliable supports to the next inquiry or diagnosis.

3.2.3 Process Reward and RL Objective

We score the reasoning quality of $z_t^{(2)}$ with an LLM judge along three psychiatric-reasoning dimensions: *symptom analysis* (S_t^{sym}, whether the trace organizes symptoms, onset, severity, and functional impact), *differential consideration* (S_t^{diff}, whether it weighs competing hypotheses and surfaces unresolved criteria), and *decision logic* (S_t^{dec}, whether it justifies the chosen action as informative under the current state). Grounding rewards in these stages avoids generic "bad reasoning" penalties and encourages the observation–formulation–decision progression clinicians follow:

$$\mathbf{R}_t^{\mathrm{proc}} = \left(S_t^{\mathrm{sym}}, S_t^{\mathrm{diff}}, S_t^{\mathrm{dec}}\right), \tag{13}$$

with each component scored on a 0–2 rubric and normalized before aggregation. The turn-level reward combines $\mathbf{R}_t^{\mathrm{proc}}$ with an information-gain term that credits turns reducing $|\mathcal{M}_t|$. RL then maximizes the episode return

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} r_t + r^{\mathrm{term}}\right], \tag{14}$$

where r_t aggregates process reward, information-gain reward, format compliance, and operational penalties, and r^{term} is the terminal diagnostic reward. Optimization uses GRPO.

3.2.4 State-Conditioned Trajectory Rectification

At both training and inference time, MIND monitors each turn for low-utility behaviors—malformed output, repeated inquiries, generic low-yield questions, low retrieval reliability, budget violations, and unsafe content. On a trigger, the agent first attempts a constrained self-retry conditioned on the failure mode; if the retry fails and the fallback cap is not reached, the action is replaced by a structured reference inquiry from the nearest reliable PRB entry. The rule is *state-conditioned* because every trigger is a deterministic function of evidence-state fields (e.g., $|\mathcal{M}_t|$ stagnation, semantic repetition, or low $\mathcal{R}_t$) and requires no separate value network or rollout-level reranking.

4 Experiments

Datasets. We use 1,600 de-identified Chinese psychiatric EMRs (2019–2023) over four ICD-10-style screening categories (Organization, 1992) (Depression, Anxiety, Mixed, Other; adjudicated $\kappa = 0.78$). EMR-level splits are taken once before any artifact is produced: 1,200 training EMRs feed PRB construction, SFT, and GRPO; 400 held-out EMRs drive PsySim-Std/PsySim-Adapt closed-loop evaluation. No test EMR, test-derived profile, simulated dialogue, or support note enters the PRB or any training stage. Statistics, ethics, release, and the public MDD-5k set (Yin et al., 2025) are in Appendix B, C; code and a de-identified sample: `https://anonymous.4open.science/r/MIND-83DF/`.

Baselines. We compare inference-only LLMs (GPT-4o, DeepSeek-V3.2, Qwen3-4B/8B/32B, Baichuan-M2), retrieval/criteria-guided systems (DDT, MRD-RAG), and fine-tuned/RL consultation models (Qwen3-4B/8B†, HuatuoGPT-o1-8B, DoctorAgent-RL, DDO). † marks the same SFT

Table 2: Per-class precision/recall/F1 (%) on four psychiatric categories under PsySim-Std and PsySim-Adapt. "†" shares our SFT warm-start data.

PatientSim	Setting	Method	Depression			Anxiety			Mix			Other			Acc
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	
PsySim-Std	Inference-only	GPT-4o	60.0	61.0	60.5	49.5	52.0	50.7	44.0	46.0	45.0	36.0	38.0	37.0	49.5
		DeepSeek-V3.2	50.0	42.0	45.6	30.0	38.0	33.5	48.0	44.0	45.9	60.0	58.0	59.0	46.0
		Qwen3-4B	48.0	44.5	46.2	25.0	23.0	24.0	13.5	15.0	14.2	43.5	42.0	42.7	32.8
		Qwen3-8B	63.0	62.0	62.5	38.5	42.0	40.2	30.0	32.0	31.0	30.5	32.0	31.2	43.0
		Qwen3-32B	63.0	62.5	62.7	38.5	36.5	37.5	32.0	34.0	33.0	51.5	49.0	50.2	47.2
		Baichuan-M2	60.5	60.3	60.4	33.0	36.5	34.7	32.5	30.0	31.2	45.5	46.0	45.7	44.7
		DDT	56.0	51.0	53.4	36.0	39.0	37.4	41.0	37.0	38.9	50.0	55.0	52.4	46.5
		MRD-RAG	59.0	62.0	60.5	49.0	52.0	50.5	41.0	39.0	40.0	32.0	30.0	31.0	46.0
	Fine-tuned	Qwen3-4B†	57.6	55.4	56.5	25.4	23.6	24.5	31.2	28.7	29.9	47.1	52.0	49.4	35.5
		Qwen3-8B†	67.4	64.1	65.7	65.8	71.4	68.5	60.2	58.6	59.4	65.3	66.0	65.6	53.5
		HuatuoGPT-o1-8B	55.0	53.4	54.2	54.7	50.9	52.7	27.3	25.8	26.5	32.8	33.5	33.1	55.5
		DoctorAgent-RL	73.2	72.6	72.9	67.8	72.9	70.3	65.1	60.4	62.7	65.5	67.0	66.2	60.2
		DDO	76.4	73.8	75.1	73.5	70.1	71.8	65.7	62.4	64.0	67.2	70.4	68.8	60.2
		Ours (MIND-4B)	61.3	60.8	61.0	43.6	49.9	46.5	42.7	38.0	40.2	58.5	58.4	58.4	52.0
		Ours (MIND-8B)	**79.6**	**77.0**	**78.3**	**72.4**	**68.7**	**70.5**	**66.3**	**67.1**	**66.7**	**73.0**	**78.7**	**75.7**	**69.0**
PsySim-Adapt	Inference-only	GPT-4o	55.0	56.4	55.7	39.4	36.7	38.0	21.6	23.0	22.3	6.4	4.9	5.6	40.5
		DeepSeek-V3.2	40.6	32.9	36.3	19.5	27.3	22.7	42.0	41.6	41.8	58.9	65.4	62.0	36.8
		Qwen3-4B	42.1	41.7	41.9	18.5	18.7	18.6	25.4	27.0	26.2	36.7	35.2	35.9	24.3
		Qwen3-8B	52.0	52.3	52.1	27.4	27.5	27.4	19.5	21.1	20.3	12.6	11.5	12.0	33.5
		Qwen3-32B	54.3	55.0	54.6	30.7	30.0	30.3	23.6	25.3	24.4	38.2	46.3	41.9	37.8
		Baichuan-M2	52.7	51.5	52.1	28.6	28.0	28.3	21.5	23.0	22.2	33.4	41.0	36.8	35.9
		DDT	44.0	44.7	44.3	30.9	30.0	30.4	33.5	28.6	30.9	46.7	44.7	45.7	37.0
		MRD-RAG	50.4	51.6	51.0	39.5	46.6	42.8	33.0	27.3	29.9	22.4	20.5	21.4	36.5
	Fine-tuned	Qwen3-4B†	41.3	40.9	41.1	12.6	11.4	12.0	22.5	23.8	23.1	36.4	34.9	35.6	27.0
		Qwen3-8B†	53.4	53.7	53.5	47.6	47.0	47.3	39.8	40.6	40.2	48.1	55.7	51.6	44.0
		HuatuoGPT-o1-8B	58.7	59.4	59.0	51.0	49.9	50.4	41.5	35.4	38.2	30.4	28.5	29.4	43.5
		DoctorAgent-RL	60.0	59.0	59.5	55.3	55.0	55.1	41.8	44.0	42.9	49.7	57.2	53.2	52.5
		DDO	61.2	58.5	59.8	51.7	58.9	55.1	49.7	46.2	47.9	52.7	52.9	52.8	49.4
		Ours (MIND-4B)	53.7	52.9	53.3	49.0	48.3	48.6	41.5	40.9	41.2	47.4	47.9	47.6	45.0
		Ours (MIND-8B)	**70.2**	**67.4**	**68.8**	**55.6**	**52.1**	**53.8**	**45.1**	**47.0**	**46.0**	**65.0**	**71.5**	**68.1**	**59.8**

Table 3: Public MDD-5k four-way diagnosis from released dialogue history. Set is excluded from PRB, SFT, RL, reward, simulator, and hyperparameter tuning.

Model	Acc	Macro-F1
Qwen3-8B	34.8	32.9
Qwen3-8B†	42.5	40.6
MRD-RAG	38.6	36.8
DoctorAgent-RL	46.8	45.0
DDO	48.6	47.0
MIND-8B	**53.8**	**52.0**

Table 4: Patient simulator audit: profile factual consistency, unsupported symptom rate, label leakage, and answerability over all simulator turns.

Metric	PsySim-Std	PsySim-Adapt
Profile factual consistency (%) ↑	94.2	91.6
Unsupported symptom rate (%) ↓	3.1	5.4
Label leakage rate (%) ↓	0.0	0.0
Answerability rate (%) ↑	92.7	88.9

warm-start data as MIND. Implementation and fairness controls are in Appendix D, R.2.

Experimental Setup. MIND uses Qwen3 backbones with a PRB-derived SFT warm start and GRPO, optimizing terminal correctness, process supervision, information gain, retrieval reliability, and operational penalties. All systems share a turn budget ($L = 10$), final-diagnosis protocol, and four-way label space, and are evaluated under two EMR-grounded simulators (PsySim-Std, PsySim-Adapt) plus the simulator-independent public MDD-5k. We report accuracy, macro-F1, support faithfulness, and human ratings with 95% paired bootstrap CIs against the strongest baseline; full hyperparameters and statistical protocols are in Appendix B.

Patient Simulator Validation. Since simulator and policy may share a Qwen3 (Yang et al., 2025) backbone, we audit the simulator first. Profiles exclude gold labels and forbid out-of-profile symptoms; an automatic audit covers profile factual consistency, unsupported symptom rate, label leakage, and answerability, and two medically trained annotators validate 100 responses (Cohen's κ=0.81). Label leakage is zero by construction and factual consistency exceeds 90% under both simulators (Table 4). Re-running under GLM 4.7 (Team et al., 2025a) and Llama-3-8B-Instruct (Grattafiori et al., 2024) simulators keeps the MIND-over-DDO gap at +8.5 to +8.9 (Appendix Q), and non-matching simulator–policy pairings preserve the ordering (Appendix O.2); full prompts and protocol are in Appendix B.

Table 5: Matched-variable intervention on the policy-input representation (Qwen3-8B, same SFT/GRPO/rewards/retrieval/turn budget; only representation differs). First six rows disable fallback to isolate representation; two middle rows share the typed state with only policy or only reward to isolate policy–reward consistency.

Variant	PsySim-Std		PsySim-Adapt	
	Acc	Macro-F1	Acc	Macro-F1
Raw-history RL	63.4	64.5	53.9	54.4
Standard RAG RL	66.0	67.6	56.5	56.9
Structured-prompt RAG RL	66.7	68.7	57.2	57.5
Policy=ES, Reward=Raw	67.2	69.4	57.8	58.0
Policy=Raw, Reward=ES	66.4	68.1	56.9	57.0
Evidence-state RL (both)	68.2	71.8	59.0	58.5
MIND	**69.0**	**72.9**	**59.8**	**59.2**

Table 6: MIND component ablation under two simulators. Support is the FC/SG/PF average from the support-faithfulness rubric. Generalization of PRB to non-MIND backbones (Qwen3-32B / Baichuan-M2 ± PRB) is in Appendix U.

Model	PsySim-Std			PsySim-Adapt		
	Support	Acc	F1	Support	Acc	F1
MIND w/o SFT	8.0	62.7	64.0	7.8	53.7	54.5
MIND w/o InfoGain reward	8.2	66.5	67.7	8.0	57.5	57.8
MIND w/o Reasoning reward	8.1	65.3	66.9	7.9	56.0	56.4
MIND w/o PRB	8.0	65.7	66.1	7.7	56.3	57.0
MIND w/o reliability gate	8.1	66.5	67.2	7.8	57.1	57.7
MIND w/o thinking	7.5	61.5	60.6	7.0	51.9	52.3
MIND w/o fallback	8.2	67.1	67.8	8.1	58.0	58.5
MIND	**8.3**	**69.0**	**72.9**	**8.3**	**59.8**	**59.2**

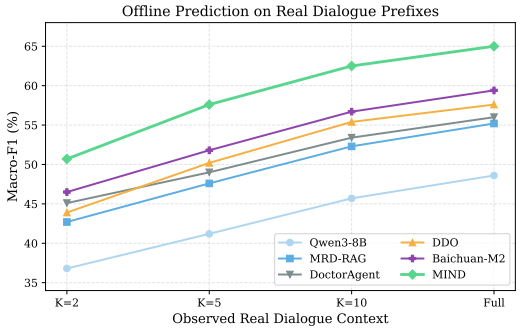

Figure 4: Offline screening on public MDD-5k prefixes: MIND scales best with longer public histories, confirming evidence-state aggregation transfers outside simulator rollout.

4.1 Results and Analysis

Overall Performance. Table 2 reports per-class P/R/F1 under both simulators. MIND is strongest overall and per category. The MIND-over-DDO gap is largest when early narratives admit multiple labels and the agent must judge which criterion is unresolved, when to stop, and which retrieved cue is reliable, indicating that criteria-grounded PRB retrieval, reliability gating, and process-supervised state-to-action reasoning—rather than turn-budget effects—drive the gain.

State-Representation Comparison. All variants share the Qwen3-8B backbone, SFT data, GRPO setup, reward weights, retrieval content, and turn budget; only the policy-input representation differs (Table 5). Raw history and standard RAG trail MIND, prompt-only structuring yields a modest gain, one-sided sharing of the typed state is intermediate, and the largest jump occurs when policy and reward both operate on the evidence state—identifying *policy–reward consistency over a shared typed state*, rather than retrieval content or prompt layout, as the main source of improvement.

Public MDD-5k Diagnostic Evaluation. On released MDD-5k (Yin et al., 2025) dialogues, with no follow-up and a fixed history, MIND raises accuracy from 48.6 to 53.8 and macro-F1 from 47.0 to 52.0 over the strongest baseline (Table 3), indicating that the evidence-state policy transfers beyond closed-loop interaction. Protocol and prefix-based next-question results are in Appendix J.1, J.2.

Ablation Study. Table 6 decomposes MIND into training, retrieval, reasoning, and recovery. Across both simulators, removing SFT weakens the two-step turn format, dropping reasoning or information-gain rewards reduces evidence acquisition, and removing PRB or reliability gating lowers both support quality and accuracy; disabling fallback has a smaller but stable effect, matching its role as low-yield-turn repair rather than a representation change. PRB also transfers to non-MIND backbones (Qwen3-32B, Baichuan-M2) under both simulators (Appendix U). Field masking, token-matched RAG, random retrieval, and shuffled reliability further show $\mathcal{E}_t$ is not a prompt template (Appendix K.1, T); matched state-abstraction, separated-judge, scale-behavior, and paired-significance results are in Appendix M.2, N, O.1, P.

Offline Scaling on Public Prefixes. On increasing prefix lengths from public MDD-5k dialogues (Figure 4), MIND's accuracy rises monotonically and its gap over the strongest baseline widens, showing that the evidence-state policy converts longer public histories into screening accuracy. Reliability-threshold τ_{rel} sensitivity (gold-label peak at 0.2) and an independent GPT-4o cross-judge for faithfulness—separate from the LLM producing ρ_i—are in Appendix L, G.

Inquiry-Quality Indicator. MIND attains the

Table 7: Human evaluation on 160 cases (40 per category) by five reviewers (one licensed psychiatrist, four medically trained graduate students).

Model	Professional Reliability	Naturalness	Preference (%)
GPT-4o	2.95	3.32	45.0
Qwen3-8B	2.15	2.80	22.5
Baichuan-M2	2.50	3.05	32.0
DDO	3.15	3.20	52.5
Ours (MIND)	**4.05**	**3.92**	**78.5**

highest Missing-Check Resolution (MCR-shared 49.6 vs. DDO 41.2) and the lowest repetition (8.0 vs. 12.1), reflecting more targeted evidence acquisition (Appendix F).

Support Faithfulness Evaluation. Each dialogue is paired with PRB-retrieved turn-level supports; DeepSeek-V3.2 scores anonymized outputs on factual consistency (FC), support grounding (SG), and patient faithfulness (PF). MIND attains the top average (8.3), $+0.8$ over DDO and 1.5–2.5 over inference-only and RAG baselines, with the largest gap on SG ($+0.8$), consistent with support-conditioned decisions (Appendix Table 8). An independent GPT-4o judge preserves the ranking (Spearman 0.93); human–judge correlations are in Appendix G, S.

Human Evaluation. Five reviewers blindly rate 160 anonymized cases on Professional Reliability, Naturalness, and pairwise Preference, with large PR disagreements adjudicated by the psychiatrist. MIND leads on PR (4.05), Naturalness (3.92), and preference (78.5%), with pairwise agreement 0.79 and pre-adjudication Krippendorff's $\alpha = 0.74$ (Table 7); a balanced diagnostic–dialogue radar is given in Appendix W.

5 Related Work

5.1 Psychiatric Consultation as Multi-turn Inquiry

Multi-turn medical consultation has been cast as sequence generation, follow-up selection, and uncertainty-reducing planning over patient narratives (Valizadeh and Parde, 2022; Tang et al., 2016; Wei et al., 2018; Kao et al., 2018; Chen et al., 2022; Wang et al., 2023; Hou et al., 2023; Zhao et al., 2024; Hu et al., 2024; Chopra and Shah, 2025; Chen et al., 2025; Xu et al., 2026). Psychiatric consultation differs in that diagnostic thresholds, exclusion checks, functional impairment, and safety risks are seldom explicit, causing drift to low-yield questions or unsupported assertions (Xia et al., 2020; Xu et al., 2019; Liu et al., 2022; Zhong et al., 2022; Yin et al., 2025).

5.2 Medical LLMs for Sequential Decision-Making

Medical LLMs have advanced history taking, diagnosis, and treatment planning via prompting, domain adaptation, CoT, and instruction tuning (Zhou et al., 2025; Kim et al., 2024; Shi et al., 2024; Johri et al., 2024; Jia et al., 2025a; Rose et al., 2025; Zheng et al., 2024; Liu et al., 2024b; Tian et al., 2024; Wang et al., 2025), with recent moves toward proactive consultation, multi-agent collaboration, and RL-optimized inquiry (Tu et al., 2025; Ma et al., 2025; Zhang et al., 2024; Liu et al., 2023; Jia et al., 2025b; Dou et al., 2025; Pan et al., 2025; Lai et al., 2025; Shao et al., 2024; Chen et al., 2024; Feng et al., 2026). In these systems policy, reward, and recovery read raw history or appended retrieval independently, yielding sparse long-horizon supervision; MIND changes *what is optimized* by sharing a typed evidence state across all three.

5.3 Retrieval-Augmented Medical Reasoning

RAG grounds medical LLMs in guidelines, references, or case knowledge (Jeong et al., 2024; Asai et al., 2024; Li et al., 2023), but unfiltered passages can mislead inquiry in metaphorical or underspecified psychiatric narratives. MIND turns retrieval into state construction: a clinical retrieval state queries the Psychiatric Reasoning Bank, and admitted entries populate typed fields shared by policy, reward, and recovery (Yin et al., 2025; Tu et al., 2025). Unlike POMDP belief states (Young et al., 2013; Williams and Young, 2007) or slot-based DST (Wu et al., 2019), $\mathcal{E}_t$ is a typed decision interface carrying criterion-linked supports and reliability metadata for joint policy and reward use.

6 Conclusion

MIND is a criteria-grounded RL framework that recasts psychiatric inquiry as evidence-state decision making. A training-split Psychiatric Reasoning Bank turns retrieval into a shared state for policy learning, process rewards, information-gain scoring, and trajectory recovery. Across EMR-grounded simulators and MDD-5k, MIND improves accuracy, support faithfulness, and inquiry quality over inference-only, RAG, and RL baselines. Matched-variable interventions further show that these gains come from the shared evidence-state interface rather than prompt format or retrieval volume, supporting evidence-state design for long-horizon criteria-grounded medical dialogue.

Limitations

This work has several limitations. First, we focus on a four-way psychiatric screening setup rather than the full granularity of clinical coding. This design keeps the dialogue objective and error analysis controlled, but it also means that the reported labels should be interpreted at the screening-category level. The finer composition of the Other category is summarized in Appendix V.

Second, our main experiments use English-rendered prompts and paper-facing terminology for a dataset constructed from Chinese clinical materials. The release will therefore include the original-language prompts and evaluation scripts so that independent reproductions can avoid ambiguity introduced by translation or notation differences. Finally, several appendix analyses use compact tables and sampled audits for readability; these analyses are intended to characterize model behavior rather than replace the primary accuracy, macro-F1, support-faithfulness, and human-evaluation results reported in the main text.

7 Ethical Considerations

Large language models may generate fluent but incorrect or unsupported clinical content, especially in high-stakes settings such as psychiatric consultation. Our work is intended for academic research on decision-making and training signals in diagnostic dialogue. It does not provide medical advice and should not be used as a stand-alone system for real-world clinical decision-making. In practice, diagnosis and treatment must be determined by qualified clinicians based on comprehensive assessment and appropriate clinical procedures.

To reduce potential misuse, we focus on criteria-grounded supports as decision cues rather than clinical evidence, and we evaluate our method primarily in a controlled simulation environment. The patient simulator and reconstructed profiles are used for research and benchmarking only, and do not represent real patients. Any deployment would require rigorous prospective validation, human oversight, safety monitoring, and compliance with applicable regulations and institutional review processes.

Safety safeguards. MIND is intended only for research on psychiatric inquiry behavior and screening-level decision support. It is not a diagnostic device and must not be used for autonomous clinical decision-making. We evaluate the framework on five high-stakes scenarios (suicidal ideation, self-harm intent, psychosis crisis, direct medication requests, and emergency risk escalation; Appendix X); on each scenario the system declines to act as a stand-alone clinician, refers to licensed clinicians or crisis services when indicated, and abstains from prescribing medication or specifying dosages. Released artifacts contain only de-identified synthetic trajectories and do not include real patient health information. Bias across demographic groups, longitudinal validity, and prospective clinical safety remain open questions and must be addressed before any operational use.

References

Ankit Aich, Avery Quynh, Pamela Osseyi, Amy Pinkham, Philip Harvey, Brenda Curtis, Colin Depp, and Natalie Parde. 2025. Using llms to aid annotation and collection of clinically-enriched data in bipolar disorder and schizophrenia. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 181–192.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2025. Cod, towards an interpretable medical agent using chain of diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14345–14368.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440.

Harshita Chopra and Chirag Shah. 2025. Feedback-aware monte carlo tree search for efficient information seeking in goal-oriented conversations. *arXiv preprint arXiv:2501.15056*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, et al. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. 2026. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. In *ICASSP 2026-2026 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 16952–16956. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhenyu Hou, Yukuo Cen, Ziding Liu, Dongxue Wu, Baoyan Wang, Xuanhe Li, Lei Hong, and Jie Tang. 2023. Mtdiag: an effective multi-task framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14241–14248.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei W Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215.

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

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025a. medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9278–9298.

Zhihao Jia, Mingyi Jia, Junwen Duan, and Jianxin Wang. 2025b. Ddo: Dual-decision optimization for llm-based medical consultation via multi-agent collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26380–26397.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Femke Lamers, Patricia van Oppen, Hannie C Comijs, Johannes H Smit, Philip Spinhoven, Anton JLM van Balkom, Willem A Nolen, Frans G Zitman, Aartjan TF Beekman, and Brenda WJH Penninx. 2011. Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the netherlands study of depression and anxiety (nesda). *Journal of clinical psychiatry*, 72(3):341.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Zhou, and Zuozhu Liu. 2024b. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. " my nose is running."" are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. *arXiv preprint arXiv:2204.13953*.

Mingyu Derek Ma, Xiaoxuan Wang, Yijia Xiao, Anthony Cuturrufo, Vijay S Nori, Eran Halperin, and Wei Wang. 2025. Memorize and rank: Elevating large language models for clinical diagnosis prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24786–24794.

World Health Organization. 1992. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*, volume 1. World Health Organization.

World Health Organization et al. 2022. Mental health and covid-19: early evidence of the pandemic's impact: scientific brief, 2 march 2022. Technical report, World Health Organization.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. 2025. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13803–13826.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22294–22314.

Penglei Sun, Yixiang Chen, Xiang Li, and Xiaowen Chu. 2025. The multi-round diagnostic rag framework for emulating clinical reasoning. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4073–4076. IEEE.

Kai-Fu Tang, Hao-Cheng Kao, Chun-Nan Chou, and Edward Y Chang. 2016. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS workshop on deep reinforcement learning*.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025b. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Guoxin Wang, Minyu Gao, Shuai Yang, Ya Zhang, Lizhi He, Liang Huang, Hanlin Xiao, Yexuan Zhang, Wanyue Li, Lu Chen, et al. 2025. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*.

Huimin Wang, Wai-Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023. Coad: Automatic diagnosis through symptom and disease collaborative generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6348–6361.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Caleb Winston, Cleah Winston, Claris Winston, and Chloe Winston. 2024. Medical question-generation for pre-consultation with LLM in-context learning. In *GenAI for Health: Potential, Trust and Policy Compliance*.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 808–819.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1062–1069.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Shihao Xu, Tiancheng Zhou, Jiatong Ma, Yanli Ding, Yiming Yan, Ming Xiao, Guoyi Li, Haiyang Geng, Yunyun Han, Jianhua Chen, et al. 2026. Lingxidiagbench: A multi-agent framework for benchmarking llms in chinese psychiatric consultation and diagnosis. *arXiv preprint arXiv:2602.09379*.

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

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. 2025. Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25715–25723.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966.

Xuehan Zhao, Jiaqi Liu, Yao Zhang, Zhiwen Yu, and Bin Guo. 2024. Haiformer: Human-ai collaboration framework for disease diagnosis via doctor-enhanced transformer. In *ECAI 2024*, pages 1495–1502. IOS Press.

Zi'ou Zheng, Christopher Malon, Martin Renqiang Min, and Xiaodan Zhu. 2024. Exploring the role of reasoning structures for constructing proofs in multi-step natural language reasoning with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15299–15312.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, 38(16):3995–4001.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, et al. 2025. Large language models for disease diagnosis: A scoping review. *npj Artificial Intelligence*, 1(1):9.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

Algorithm 1: MIND turn execution with typed evidence-state construction and state-conditioned rectification

Input: Dialogue history H_t, PRB, policy π_θ, thresholds τ_{rel}, turn budget L
Output: Doctor action a_t

Generate clinical retrieval state q_t from H_t
Parse q_t into observed evidence $\mathcal{O}_t$, missing checks $\mathcal{M}_t$, and active differentials $\mathcal{D}_t$
Retrieve candidate PRB entries $\widetilde{\mathcal{S}}_t$ from PRB
Compute reliability $\text{Rel}(q_t, e_i)$ for each entry
Set $\mathcal{S}_t = \{e_i : \text{Rel}(q_t, e_i) \geq \tau_{\text{rel}}\}$ and store scores/flags in $\mathcal{R}_t$
Construct $\mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t)$
Generate reasoning and action $a_t \sim \pi_\theta(\mathcal{E}_t, H_t)$
if *a_t violates format, repetition, safety, retrieval, or budget checks* **then**
  Retry once with the detected failure mode specified
  if *retry still violates checks and fallback cap is not reached* **then**
    Replace a_t with a structured reference inquiry from the nearest reliable PRB entry
return a_t

Table 8: Support-faithfulness scores (0–10) on factual consistency (FC), support grounding (SG), and patient faithfulness (PF); DeepSeek-V3.2 judge. An independent GPT-4o cross-judge preserves the ranking (Table 12, Appendix G).

Model	FC	SG	PF	Avg.	Model	FC	SG	PF	Avg.
Qwen3-8B	6.0	5.6	5.9	5.8	Qwen3-8B†	7.0	7.1	6.8	7.0
Baichuan-M2	6.1	5.7	5.9	5.9	Qwen3-32B†	7.3	7.4	7.1	7.3
MRD-RAG	6.4	6.5	5.9	6.3	DoctorAgent-RL	7.4	7.5	7.0	7.3
DDO	7.5	7.7	7.3	7.5	Ours (MIND)	**8.3**	**8.5**	**8.0**	**8.3**

A Turn-level Execution Procedure

Algorithm 1 gives the turn-level execution procedure used by MIND. At each turn, the dialogue history is first converted into a clinical retrieval state, whose observed evidence, missing criterion checks, and active differentials determine which PRB entries are admitted into the typed evidence state. The same state then conditions action generation, process-level validation, and controlled rectification, so retrieval is not an auxiliary context but the decision interface shared by policy and safeguards. This operationalizes the main paper's abstraction: long-horizon psychiatric inquiry is guided by criterion-linked evidence and reliability metadata rather than by unstructured retrieved text or prompt-only heuristics.

B The Details of Experimental Setup

Datasets.

Our dataset starts from 1,600 de-identified EMRs containing structured psychiatric descriptions, including chief complaints, symptom clusters, course/onset, severity and functional impact, and risk/exclusion cues. Each case is assigned to one of four ICD-10-style psychiatric screening categories (Organization, 1992): Depression, Anxiety, Mixed Anxiety–Depression, and Other. The Mixed category denotes cases where both depressive and anxiety symptoms are clinically salient and the primary syndrome is difficult to determine from available evidence, while Other includes remaining psychiatric conditions or cases that cannot be confidently assigned to the first three categories. The raw EMR label distribution is 320/536/424/320 for Anxiety/Depression/Mix/Other. EMRs are split at the EMR level into 1,200 training and 400 held-out test EMRs; the EMRs themselves provide the clinical basis and screening labels, while simulated dialogues instantiate controlled multi-turn interactions for policy learning and evaluation.

Following MDD-5k-style construction, we reconstruct a structured patient profile from each EMR and simulate multi-turn consultations conditioned on the profile. Importantly, we enforce EMR-level split isolation *before* profile reconstruction, dialogue simulation, PRB construction, and model training: the 1,200 training EMRs are the sole source of SFT data, RL trajectories, and PRB entries, while the 400 test EMRs feed only the closed-loop evaluation. All artifacts derived from a test EMR—reconstructed profiles, simulated dialogues, retrieval states, support notes, and clinical rationales—are excluded from the PRB used during both training and inference, preventing test-derived information from leaking into retrieval, reward computation, or policy optimization.

The resulting augmented data are summarized in Table 9. The three rows are reported under *different counting units* that match how each split is consumed: SFT is counted at the *dialogue-turn* level—every (history, doctor-response) pair from the simulated training-EMR consultations is a separate supervised example—while RL and Test are counted at the *dialogue* level, with each entry corresponding to one closed-loop trajectory rolled out from a training or test EMR under PsySim-Std/PsySim-Adapt with the standard turn budget $L = 10$. The SFT row is therefore numerically larger than the underlying 1,200 training EMRs (an average of $\sim$11.6 doctor turns per simulated consultation), the RL row reflects multi-seed roll-outs over the same 1,200 training EMRs across both simulators, and the Test row is the 602 dialogue-level evaluation

Table 9: Dataset split and augmentation statistics by diagnostic category. SFT counts are at the *dialogue-turn* level (each doctor-response turn from a simulated training-EMR consultation is one example); RL counts are at the *dialogue* level, obtained as multi-seed PsySim-Std/PsySim-Adapt roll-outs over the same 1,200 training EMRs; Test counts are dialogue-level evaluation instances generated from the 400 held-out EMRs.

Split	Category				Total
	Anxiety	Depression	Mix	Other	
SFT	2,410	4,120	3,640	3,720	**13,890**
RL	884	1,929	1,141	1,485	**5,399**
Test	115	182	140	165	**602**

instances obtained from the 400 held-out EMRs. Since high-quality psychiatric consultation data are scarce and costly to construct, we use this four-way label space as a controlled testbed for studying multi-turn psychiatric screening and differential inquiry, rather than as a replacement for fine-grained clinical diagnosis. The released resources follow the language of the available clinical materials and expert evaluations.

C Planned Release of De-identified Consultation Trajectories

We will release the Chinese data, source code, trained models, evaluation scripts, construction protocols including PRB construction, and expert-reviewed, de-identified synthetic consultation trajectories generated by MIND to facilitate reproducible research on criteria-grounded psychiatric consultation. Each released instance contains the reconstructed patient profile, multi-turn inquiry trajectory, final ICD-10-style category, retrieved PRB supports, and basic metadata such as turn length and evidence tags. The release is intended as a research resource for studying inquiry behavior and diagnosis in controlled criteria-grounded settings, rather than as a substitute for real clinical data.

Quality Control. To improve the reliability of the released trajectories, domain experts reviewed the construction protocol and audited representative samples for profile consistency, diagnostic-label consistency, criteria coverage, and inappropriate clinical claims. During generation, each trajectory is constrained by the reconstructed patient profile and the ICD-10-style category, and expert feedback is used to refine prompts and filtering rules for inconsistent or unsupported dialogues. To prevent leakage, all splits are performed at the original EMR level before profile reconstruction, dialogue simulation, PRB construction, SFT/RL training, and evaluation. Consequently, any artifact derived from a held-out EMR, including patient profiles, simulated dialogues, retrieval states, PRB support notes, and rationales, is excluded from the training split and from the PRB used during inference.

D Baselines and Implementation Details

Baselines.

We organize baselines into four categories.

General-purpose LLMs. GPT-4o (Hurst et al., 2024) and DeepSeek-V3.2 (Liu et al., 2024a) are evaluated in an inference-only multi-turn consultation setting, serving as strong references for general instruction following and reasoning without task-specific training.

Medical LLMs. We include Qwen3-4B (Yang et al., 2025), Qwen3-8B, Qwen3-32B, and Baichuan-M2 (Dou et al., 2025) as representative open medical-capable LLMs. We include Baichuan-M2 because it is a publicly available medical-capable model with accessible checkpoints and reproducible inference settings at the time of our experiments. More recent or unavailable variants are not included when their checkpoints, inference interface, or licensing conditions prevent controlled reproduction. These models are evaluated directly under the same psychiatric consultation protocol to measure medical-domain capability without additional tuning.

Retrieval-augmented consultation baselines. DDT (Yin et al., 2025) is a guideline or criteria-tree grounded dialogue baseline that asks questions following structured diagnostic checkpoints. MRD-RAG (Sun et al., 2025) augments multi-turn consultation with retrieved supports to reduce unsupported clinical claims and improve evidence-conditioned decisions.

Fine-tuned and RL-based consultation models. Qwen3-4B† and Qwen3-8B† are instruction-tuned with the same warm-start data as our method for a controlled comparison. HuatuoGPT-o1-8B is included as a medical reasoning baseline. DoctorAgent-RL (Feng et al., 2026) and DDO (Jia et al., 2025b) optimize multi-turn inquiry policies with reinforcement-learning-style objectives to improve question selection and final diagnosis. Finally, we report ours, MIND-4B and MIND-8B.

"†" denotes using the same instruction-tuning data as our supervised warm-start for a controlled

comparison. All methods are evaluated under the same multi-turn protocol, and results are reported under the most appropriate setting for each baseline.

Experimental Setup.

Our framework follows a multi-turn doctor–patient dialogue paradigm. The doctor agent is built on Qwen3 backbones and warm-started with supervised fine-tuning on consultation instruction data, yielding MIND-4B and MIND-8B. The warm-start data are derived from the same PRB construction pipeline, which produces retrieval states, typed support entries, and reference inquiry rationales from training-split cases. We then optimize the doctor agent with GRPO on 8 NVIDIA A100 GPUs.

During each consultation turn, MIND summarizes the dialogue into a clinical retrieval state, retrieves criterion-relevant PRB entries, assigns them to the typed fields of $\mathcal{E}_t$, and admits them only when retrieval reliability passes the predefined gating threshold. The policy then generates explicit state-to-action reasoning and produces either the next inquiry or the final diagnosis. The maximum number of interaction turns is capped at $L = 10$. We allow early diagnosis when a model explicitly produces a valid final label. However, to ensure fair comparison across models with different stopping behaviors, we use the same maximum turn budget and evaluate the final valid prediction. Since early stopping is not the focus of this work and many baselines are not trained with calibrated stopping policies, we do not use average stopping turn as a primary metric.

The RL reward combines both rule-based and LLM-based signals. Rule-based signals include terminal diagnostic correctness, format validation, pollution sanitization, repeated-question penalties, retrieval gating, and information-gain rewards computed from newly extracted symptoms. LLM-based signals are used for higher-level rubric scoring of intermediate reasoning quality, including symptom analysis, differential considerations, and decision logic, as well as auxiliary interaction-quality checks. Thus, the reward is not a model-based judge reward alone: the largest-weight terminal diagnostic reward is computed from the ground-truth screening label, while the LLM judge contributes only rubric-based process and auxiliary quality signals. Retrieval reliability is used for support gating and reward shaping, while a state-conditioned rectification mechanism triggers self-retry or PRB-guided fallback when low-utility turns are detected.

Unless otherwise specified, MIND refers to the 8B variant in the following experiments. We evaluate MIND and all baselines under the same multi-turn protocol and report accuracy and macro-average F1 as primary metrics. Experiments are conducted under two patient simulators, PsySim-Std and PsySim-Adapt, to assess robustness under simulator variation. In addition, we conduct support-faithfulness evaluation using DeepSeek-V3.2 as an automatic judge along three dimensions: factual consistency, support grounding of key clinical claims, and faithfulness to patient context. Human evaluation further assesses consultation quality under a standardized rubric. Key hyperparameters for SFT, RL training, and inference/simulation are summarized in Tables 16, 17, and 20, respectively. The data distribution of the PRB support library is shown in Table 15.

Statistical Testing. For accuracy and macro-F1, we estimate uncertainty with non-parametric bootstrap resampling over test cases. Specifically, we sample test cases with replacement for $B = 1{,}000$ bootstrap replicates, recompute each metric on every replicate, and report the 2.5th and 97.5th percentiles as the 95% confidence interval. To compare MIND with the strongest baseline under the same evaluation cases, we use paired bootstrap resampling: for each replicate, we resample cases once and compute the metric difference between the two systems on the identical resampled set. A one-sided p-value is estimated as the fraction of replicates where the difference is less than or equal to zero. We mark improvements as statistically significant when $p < 0.05$. In the current run, MIND improves over the strongest non-MIND baseline with statistically significant margins under both simulators. Under PsySim-Std, the accuracy gain is +8.8 points (95% CI: [5.4, 12.1], $p < 0.01$), and the macro-F1 gain is +4.6 points (95% CI: [1.9, 7.2], $p < 0.01$). Under PsySim-Adapt, the accuracy gain is +7.3 points (95% CI: [3.9, 10.6], $p < 0.01$), and the macro-F1 gain is +6.4 points (95% CI: [3.1, 9.5], $p < 0.01$).

E State-Representation Comparison

To isolate whether the gain comes from the evidence-state interface rather than from retrieval text or prompt formatting alone, we compare four policy-input variants under the same Qwen3-8B backbone, SFT warm-start data, GRPO configuration, reward weights, patient simulators, and

turn budget. The first four rows disable state-conditioned fallback so that the comparison focuses on the representation supplied to the policy. All variants use the same training cases, rollout budget, terminal reward, process-reward prompt, and evaluation scripts. Thus, this experiment changes the object of policy conditioning while holding the optimization and interaction environment fixed.

Raw-history RL conditions directly on the dialogue history H_t and does not use PRB retrieval at inference time. *Standard RAG RL* conditions on H_t plus top-k retrieved PRB support notes, but it appends the notes as prompt context and does not expose explicit observed-evidence, missing-check, differential-hypothesis, or reliability fields to the policy. *Structured-prompt RAG RL* uses the same typed layout as $\mathcal{E}_t$—organizing retrieved supports into labeled observed/missing/differential/support/reliability sections in the policy prompt—but restricts this structure to the policy input alone; the process-reward judge and fallback operator receive raw history rather than the typed state. This variant isolates whether the gain comes from prompt formatting or from sharing the typed state across decision operators. *Evidence-state RL* conditions on $\mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t)$: $\mathcal{O}_t$ contains patient-reported facts, $\mathcal{M}_t$ contains missing or unclear criterion checks, $\mathcal{D}_t$ contains active differential hypotheses, $\mathcal{S}_t$ contains reliability-gated support entries, and $\mathcal{R}_t$ contains support scores and hard-issue flags. The same typed state is shared with the process-reward judge and information-gain calculator. The final row reports full MIND with the same evidence-state interface plus state-conditioned rectification.

The comparison shows a monotonic improvement as the policy input becomes more structured, but the gains are not uniform across the structuring steps. Raw-history RL performs reasonably but must repeatedly infer known facts, missing criteria, active alternatives, and reliability from a growing dialogue context. Standard RAG RL improves both diagnostic and inquiry metrics by exposing the policy to relevant supports, but it still treats retrieved text as unstructured context and cannot explicitly suppress weakly matched evidence or bind it to a missing criterion. Structured-prompt RAG RL yields a modest further gain (+0.7 Acc, +1.1 F1 under PsySim-Std) by organizing the prompt into typed sections, confirming that structured formatting helps the policy parse retrieved information. However, the gap from Structured-prompt RAG RL to Evidence-state RL (+1.5 Acc, +3.1 F1 under PsySim-Std; +1.8 Acc, +1.0 F1 under PsySim-Adapt) is substantially larger than the formatting-only gain, indicating that the main benefit comes from sharing the typed state with the reward judge and information-gain calculator, not from prompt layout alone. The full MIND row adds state-conditioned rectification on top of the evidence-state interface, yielding the best repetition rate and the strongest overall performance. These results support the central claim that the main novelty is not retrieval content or prompt formatting, but retrieval-as-state-construction with shared operational semantics across policy, reward, and recovery.

F Automatic Missing-Check Resolution Analysis

We compute MCR using the structured retrieval state produced at each turn, without introducing additional human labels. Each retrieval state follows the fields used by the PRB construction prompt, including chief complaint, symptoms and duration, severity, sleep, impairment, safety risk, psychosis/mania cues, stressors, and substance use. A field is treated as a missing check when it is explicitly marked as "not mentioned", "unclear", or an equivalent null-status phrase in q_t. After the next patient response, we reconstruct q_{t+1} and count the turn as resolving a missing check if at least one previously missing field becomes observed, explicitly negated, or otherwise clinically specified. Repetition is computed by semantic duplicate detection between the current inquiry and previous doctor questions in the same trajectory.

Formally, let $\mathcal{M}_t$ be the set of missing checks extracted from q_t, let Δ_t be the subset of those checks whose status changes after the next patient response, and let T be the number of inquiry turns. We compute MCR only over turns that contain at least one missing check:

$$\mathrm{MCR} = \frac{\sum_{t=1}^{T} \mathbb{I}(|\mathcal{M}_t|>0)\mathbb{I}(|\Delta_t|>0)}{\sum_{t=1}^{T} \mathbb{I}(|\mathcal{M}_t|>0)}, \tag{15}$$

where the denominator is the number of valid inquiry turns. This metric is intentionally defined over missing *checks* rather than formal diagnostic criteria: the extracted gaps are operational fields in the evidence state, not manually annotated ICD-10 or DSM criterion labels. Thus, MCR measures whether the next question helps fill an explicit evidence gap recorded by the system.

Table 10: Matched-variable intervention on the policy-input representation. Acc/F1 are reported under the two patient simulators; MCR and repetition are averaged over inquiry turns. The first four rows isolate the representation by using the same backbone, rewards, and turn budget without state-conditioned fallback. "Structured-prompt RAG RL" uses the same typed layout as $\mathcal{E}_t$ in the policy prompt but does not share it with the reward judge or fallback operator.

Variant	PsySim-Std		PsySim-Adapt		Inquiry Quality	
	Acc	Macro-F1	Acc	Macro-F1	MCR ↑	Rep. ↓
Raw-history RL	63.4	64.5	53.9	54.4	43.5	14.5
Standard RAG RL	66.0	67.6	56.5	56.9	48.3	12.2
Structured-prompt RAG RL	66.7	68.7	57.2	57.5	49.5	11.5
Evidence-state RL	68.2	71.8	59.0	58.5	53.0	8.8
MIND	**69.0**	**72.9**	**59.8**	**59.2**	**54.2**	**8.0**

Table 11: Missing-check resolution under two rubrics. *MCR-shared* applies MIND's missing-check schema as a unified rubric to every system; *MCR-self* uses each system's own retrieval state (or extracted symptom mentions for baselines without one). Repetition is averaged over inquiry turns.

Method	MCR-shared ↑	MCR-self ↑	Repetition ↓
Qwen3-8B	29.4	24.5	18.4
MRD-RAG	32.8	35.4	15.6
Qwen3-8B†	36.5	39.2	13.7
DoctorAgent-RL	39.6	43.2	12.7
DDO	41.2	44.7	12.1
MIND-8B	**49.6**	**54.2**	**8.0**

Table 12: Support-faithfulness evaluation with GPT-4o as an independent judge from a different model family. DeepSeek-V3.2 is excluded from the evaluated systems because it serves as the primary judge in Table 8. The ranking under GPT-4o is consistent with the DeepSeek-V3.2-judged results.

Model	FC	SG	PF	Avg.
Qwen3-8B	5.9	5.4	5.8	5.7
Qwen3-8B†	6.9	7.0	6.7	6.9
Qwen3-32B	6.0	5.7	5.9	5.9
Qwen3-32B†	7.2	7.3	7.0	7.2
Baichuan-M2	6.0	5.6	5.8	5.8
MRD-RAG	6.3	6.4	5.8	6.2
DoctorAgent-RL	7.3	7.4	6.9	7.2
DDO	7.4	7.6	7.2	7.4
MIND	**8.1**	**8.2**	**7.9**	**8.1**

G Judge Robustness for Support-Faithfulness Evaluation

The main support-faithfulness evaluation uses DeepSeek-V3.2 as the automatic judge. To reduce the risk that the conclusion is specific to the judge used for PRB construction or evaluation, we repeat the same evaluation with GPT-4o (Hurst et al., 2024), which is from a different model family and is not used for PRB construction, SFT, RL rewards, model selection, or hyperparameter tuning. The prompt, sampled outputs, retrieved supports, patient context, and scoring dimensions are kept identical to Appendix Table 8. GPT-4o assigns scores for factual consistency (FC), support grounding (SG), and patient faithfulness (PF) on the same 0–10 scale.

As shown in Tables 12 and 13, GPT-4o yields the same overall and per-dimension ordering as DeepSeek-V3.2, with MIND obtaining the highest average support-faithfulness score (8.1) and a 0.7-point margin over the strongest baseline DDO; absolute scores from GPT-4o are uniformly 0.1–0.3 points lower than those from DeepSeek-V3.2 across systems, so the rank-level gaps in Table 8 are reproduced rather than compressed. The high Pearson and Spearman correlations indicate that the relative differences among systems are stable across judge families, while the rounded-score weighted κ values indicate substantial agreement. These results suggest that the support-faithfulness gains are not merely an artifact of a single evaluator's preferences.

H PRB Construction and Policy Prompts

For reproducibility we list the offline prompts used to populate the typed PRB schema described in Section 3.1 from training-split dialogues, together with the Step 2 reasoning prompt used by the doctor policy at inference time.

Table 13: Agreement between DeepSeek-V3.2 and GPT-4o support-faithfulness judgments. Pearson and Spearman correlations are computed at the sample level; weighted κ is computed after rounding scores to integer bins.

Dimension	Pearson	Spearman	Weighted κ
FC	0.93	0.91	0.71
SG	0.94	0.92	0.74
PF	0.92	0.90	0.69
Avg.	0.95	0.93	0.76

Clinical Retrieval State q_i. ***Prompt****: Compress the dialogue history into a fact-only clinical retrieval state q_i. Mark missing information explicitly and do not provide a diagnosis. Cover: chief complaint, symptoms and duration, severity, sleep, impairment, safety risk, psychosis/mania cues, stressors, and substance use.* ***Output****: One concise paragraph: "Patient reports ...; duration ...; impairment ...; risk ...; psychosis/mania cues ...; stressors ...; substances ...".*

Clinical Support Synthesis. ***Prompt****: Given the dialogue history and clinical references, construct a criterion-indexed evidence entry.* ***Output****: (Criterion/check) ...; (Observed evidence) ...; (Missing or unclear check) ...; (Exclusion/safety cue) ...; (Next-inquiry rationale) Asking "next_question" helps clarify ... and guides the next diagnostic decision.*

Reliability Assessment. ***Prompt****: Verify whether the support is consistent with the dialogue context and clinically relevant to the references. Score reliability from 1 to 5 and mark hard issues.* ***Output****: (Score) 4/5. (Hard issues) H1–H3: NO. (Rationale) ...*

Doctor Policy Step 2 (state-to-action reasoning). ***Prompt****: Given the dialogue history and the typed evidence state (observed evidence, missing checks, active differentials, reliable support entries, and reliability flags), reason about the current diagnostic state and decide the next action.* ***Output****: <think> The patient reports ..., while ... remains unclear. The key differential issue is .... The next question should clarify ... because it helps distinguish ... from ....* </think> <answer> ... </answer>.

I Sensitivity to PRB Construction LLMs

Although the main method only assumes an offline PRB construction procedure, we evaluate whether the choice of the off-the-shelf LLM used in this preprocessing step affects downstream performance.

Table 14: Sensitivity to off-the-shelf LLMs used in the offline PRB construction pipeline. Scores are close across construction LLMs, with Kimi-K2 slightly higher in this run.

Construction LLM	Support	PsySim-Std		PsySim-Adapt	
		Acc	F1	Acc	F1
DeepSeek-V3.2 (Liu et al., 2024a)	8.22	68.3	72.2	59.1	58.5
GLM-4.7 (Zeng et al., 2025)	8.19	68.0	71.8	58.8	58.2
Kimi-K2 (Team et al., 2025b)	**8.28**	**69.0**	**72.9**	**59.8**	**59.2**

Table 15: Distribution of PRB support quality scores (1–5).

Score	1	2	3	4	5
Count	2,318	27,597	15,459	47,628	8,676
Ratio	2.28%	27.14%	15.20%	46.84%	8.53%

We instantiate the same PRB construction pipeline with Kimi-K2 (Team et al., 2025b), DeepSeek-V3.2 (Liu et al., 2024a), and GLM-4.7 (Zeng et al., 2025). All settings keep the same Qwen3-8B doctor backbone, PRB construction protocol, SFT data size, GRPO configuration, and inference budget. The only changed component is the preprocessing LLM used to construct retrieval states, support notes, and reference inquiry rationales from training-split cases.

As shown in Table 14, performance remains stable across construction LLMs, indicating that the main gains come from criteria-grounded PRB usage, process supervision, state-conditioned rectification, and RL rather than a particular offline construction model. Kimi-K2 provides a small advantage, likely because its constructed supports are slightly more complete under our prompts, but the margin is modest.

J External Evaluation on MDD-5k

J.1 Diagnostic Evaluation on Public MDD-5k Dialogues

To further reduce dependence on our closed-loop simulator, we conduct an external dialogue-level diagnostic evaluation on the publicly released dialogue portion of MDD-5k. Although the original EMRs used to construct MDD-5k are not public, its released diagnostic dialogues provide a suitable public benchmark for testing whether a model can infer the final psychiatric screening category from dialogue evidence.

For each held-out dialogue, unless otherwise specified, the model receives the complete released dialogue history and predicts one of the four screening categories: Depression, Anxiety, Mixed

Hyperparameters	Value
Backbone model	Qwen3
Training method	LoRA
LoRA rank (r)	64
LoRA alpha (α)	32
Learning rate	1e-4
Warmup ratio	0.03
Effective batch size	128
Per-device batch size	2
Gradient accumulation steps	8
Training epochs	3
Max sequence length	4096
Gradient clipping	1.0

Table 16: Key hyperparameters for the SFT warm-start stage. The effective batch size is obtained with gradient accumulation over 8 GPUs.

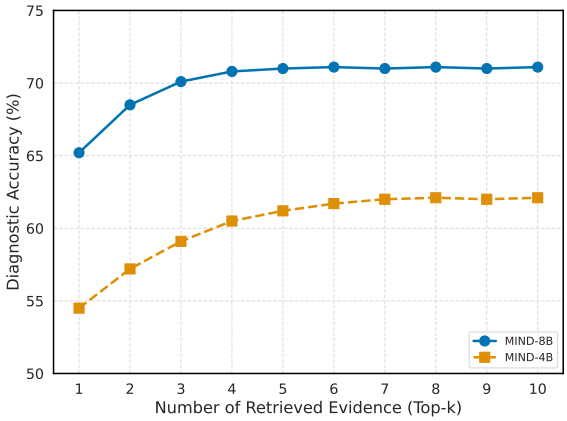

Figure 5: Retrieval-depth sensitivity. Accuracy of MIND-4B/8B vs. Top-k.

Anxiety–Depression, or Other. The prediction is compared with the released diagnostic label provided by MDD-5k. All reported scores are produced by our re-evaluation under a unified prompt, label mapping, and evaluation protocol, rather than copied from the original MDD-5k report: released depressive diagnoses are mapped to Depression, anxiety diagnoses to Anxiety, anxiety–depression or co-occurring depressive/anxiety diagnoses to Mixed Anxiety–Depression, and all remaining diagnoses to Other. We report classification accuracy and macro-F1. This evaluation set is used only for external testing and is excluded from PRB construction, SFT, RL training, reward design, simulator tuning, and hyperparameter selection.

The absolute scores are lower than the EMR-grounded simulator results in Table 2, which is expected because the models cannot ask additional

Hyperparameters	Value
Algorithm	GRPO
Compute	8× NVIDIA A100
Actor learning rate	5e-6
KL coefficient (β)	0.02
Clip ratio (ϵ)	0.10–0.18
Entropy coefficient	1e-4
Rollout backend	vLLM
Rollout temperature	1.0
Top-p	1.0
Generations per prompt	4
Policy updates	1,000–2,000
Episodes per update	128
Mini-batch size	64
Micro-batch size per GPU	1–2
PPO epochs per update	1
Gradient clipping	1.0
Max prompt length	4096
Max response length	1024
Reward weights	
Terminal screening reward	5.0
Information gain	0.005
Clinical reasoning	0.01
Format compliance	0.1

Table 17: Key hyperparameters for RL policy optimization with GRPO. The update range reflects early stopping on validation performance and training stability.

follow-up questions and must diagnose from fixed evidence. The ranking is nevertheless consistent with the main results: retrieval-augmented, tuned, and RL-based baselines improve over inference-only generation, while MIND obtains the strongest external dialogue-level performance.

J.2 Next-Question Evaluation on Public MDD-5k Dialogue Prefixes

We further evaluate whether MIND can select clinically useful follow-up questions from public real-dialogue evidence. We sample dialogue prefixes from the publicly released MDD-5k diagnostic dialogues. For each case, the model is given the chief complaint or the first K=10 doctor–patient exchanges and is asked to generate one next doctor question. No simulated patient response is provided, and no further rollout is performed. This setting isolates the immediate inquiry decision from the downstream simulator response. It therefore tests whether a model can identify a useful next evidence gap from public partial consultation context, rather than whether it can exploit simulator-

Table 18: Operational triggers and caps for state-conditioned trajectory rectification. These checks are rule-based and are used both as intervention triggers and as penalty signals during RL.

Trigger	Operational definition	Action
Malformed output	Missing required tags, invalid diagnosis format, or unparseable retrieval state.	Self-retry
Repeated inquiry	Semantic duplicate of a recent question or repeated check without new evidence.	Self-retry
Generic low-yield question	Question does not target symptoms, duration, impairment, safety, exclusion cues, or differential evidence.	Self-retry
Low retrieval reliability	No candidate support satisfies $\mathrm{Rel}(q_t, r_i) \geq \tau_{\mathrm{rel}}$.	Null support / fallback
Budget violation	Action ignores the remaining turn budget or fails to provide a final label at the last turn.	Penalty / forced diagnosis
Unsafe or off-track response	Response contains irrelevant, non-clinical, or unsafe patient-facing content.	Self-retry / fallback
Self-retry cap	At most once per turn.	Skip to fallback
Fallback cap	At most twice per episode.	Force diagnosis next turn
Total intervention cap	At most three rectification interventions per episode.	Force diagnosis next turn

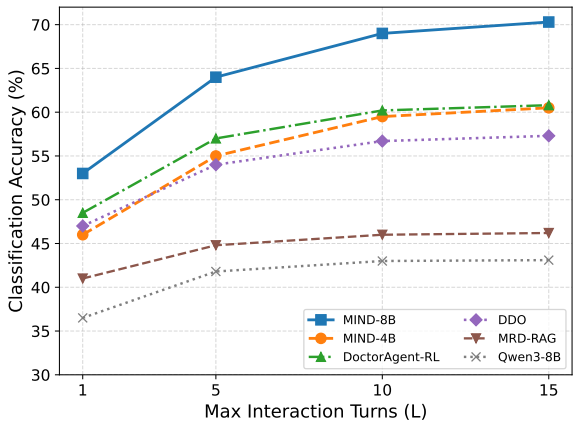

Figure 6: Dialogue-horizon sensitivity. Accuracy of MIND and baselines vs. max turns L.

specific response patterns. The same prefix is given to all systems, and PRB-based methods use only train-split PRB entries with reliability-gated support injection.

The generated questions are evaluated by reviewers with medical or clinical training under an anonymized protocol. Each reviewer is shown the same dialogue prefix and the candidate questions from different systems, but not the model identities. We use a pre-defined rubric tailored to the four-category psychiatric screening setting. The rubric covers four aspects: clinical relevance, expected information gain, safety and exclusion awareness, and patient-facing naturalness. Clinical relevance measures whether the question targets symptoms or contextual factors that are meaningful for the observed prefix. Expected information gain measures whether the answer would likely reduce uncertainty among Depression, Anxiety, Mixed Anxiety–Depression, and Other. Safety and exclusion awareness measures whether the question appropriately checks risk, functional impairment, manic or psychotic cues, substance use, or other exclusion factors when such checks are indicated. Naturalness measures whether the question is understandable, non-leading, and acceptable in a patient-facing consultation.

Scores are assigned on a 1–5 scale for each dimension and averaged across reviewers. When reviewers disagree substantially, the case is discussed and, if necessary, adjudicated by an additional senior reviewer. This evaluation measures the quality of the immediate inquiry decision rather than final diagnostic accuracy.

This evaluation complements the external diagnostic task in Appendix J.1. The former tests label prediction from released dialogue evidence, whereas this experiment tests whether a model selects an appropriate next inquiry from a public partial consultation. Together, the two analyses provide additional evidence about model behavior outside simulator-based interaction.

K Evidence-State Analyses

K.1 Evidence-State Field Masking Analysis

To verify that each field of the evidence-state interface has a distinct causal role in the decision process, we conduct a field-masking experiment at inference time. Starting from the full MIND system, we independently mask (replace with an empty placeholder) each of the five fields—$\mathcal{O}_t$, $\mathcal{M}_t$, $\mathcal{D}_t$, $\mathcal{S}_t$, $\mathcal{R}_t$—while keeping all other components unchanged. The same test cases, turn budget, and evaluation protocol are used. This experiment goes beyond standard component ablation (e.g., "w/o PRB") by isolating the contribution of individual state fields rather than entire modules.

Table 24 shows that masking any single field degrades performance, but the degradation patterns differ across fields, confirming that the fields serve distinct operational roles rather than being redundant descriptions of the same information.

$\mathcal{M}_t$ **(missing checks)** produces the largest MCR drop (-13.1 points), consistent with its designed role as the inquiry-target selector: without explicit missing-check information, the policy loses its primary signal for choosing which criterion to pursue next, leading to more generic and repeated questions.

$\mathcal{O}_t$ **(observed evidence)** produces the largest accuracy drop (-3.6 points), because the policy can no longer anchor its diagnostic hypotheses in con-

Table 19: Rubric for turn-level clinical process reward. Scores are assigned on a 0–2 scale for each dimension and normalized before reward aggregation.

Dimension	**0**	**1**	**2**
Symptom analysis	Missing or contradicting key symptoms	Mentions some relevant symptoms but omits important fields	Correctly identifies confirmed, negated, and missing symptom evidence
Differential focus	No clear differential or irrelevant alternatives	Partially considers competing categories	Explicitly distinguishes plausible categories and relevant exclusions
Decision logic	Next action is unsupported or generic	Action is somewhat relevant but weakly justified	Action directly targets the most informative missing evidence

Hyperparameters	Value
Doctor agent (MIND)	
Sampling temperature	1.0
Top-p	1.0
Max response length	2048
Reliability threshold τ_{rel}	0.2
Patient simulator	
Backbone model	Qwen3 (fixed)
Temperature	0.0
Max interaction turns (L)	10
Max response length	512

Table 20: Hyperparameters for inference and patient-agent simulation.

Table 21: External held-out MDD-5k diagnostic evaluation. Models receive the full released dialogue and predict the four-way category; the set is excluded from PRB, SFT, RL, reward, simulator, and hyperparameter tuning.

Model	Accuracy	Macro-F1
Qwen3-8B	34.8	32.9
Qwen3-8B†	42.5	40.6
MRD-RAG	38.6	36.8
DoctorAgent-RL	46.8	45.0
DDO	48.6	47.0
MIND-8B	**53.8**	**52.0**

firmed patient reports and must re-derive observed facts from the raw dialogue history.

$\mathcal{S}_t$ **(criterion supports)** produces the largest Support faithfulness drop (-0.7 points), confirming that criterion-linked support entries are the primary driver of grounded clinical claims. Without them, the policy generates more unsupported assertions.

$\mathcal{D}_t$ **(active differentials)** has a moderate effect on accuracy but a smaller effect on MCR, suggesting that differential hypotheses primarily influence which diagnostic category the policy considers rather than which question it asks next.

Table 22: Rubric for next-question evaluation on public MDD-5k dialogue prefixes. Scores are assigned on a 1–5 scale, where higher scores indicate better inquiry quality.

Dimension	**Scoring criterion**
Clinical relevance	Whether the question targets symptoms, context, duration, impairment, or other evidence relevant to the observed prefix.
Expected information gain	Whether the answer would likely reduce uncertainty among Depression, Anxiety, Mixed Anxiety–Depression, and Other.
Safety/exclusion awareness	Whether the question checks safety risk, functional impairment, psychosis/mania cues, substance use, or other exclusion factors when indicated.
Naturalness	Whether the question is understandable, non-leading, concise, and acceptable in a patient-facing consultation.

Table 23: Next-question evaluation on public MDD-5k dialogue prefixes. Each model receives the same public chief complaint or first $K{=}10$ doctor–patient exchanges and generates one follow-up question. Scores are averaged across reviewers on a 1–5 scale.

Method	Rel.	Info.	Safety	Nat.
Qwen3-8B	3.15	3.05	2.85	3.45
MRD-RAG	3.62	3.58	3.45	3.50
DoctorAgent-RL	3.68	3.61	3.48	3.60
DDO	3.70	3.66	3.55	3.62
MIND-8B	**4.18**	**4.10**	**4.02**	**3.93**

$\mathcal{R}_t$ **(reliability metadata)** has the smallest individual effect, but its removal still degrades Support faithfulness by -0.6 points, confirming that reliability gating prevents weakly matched supports from influencing the policy even when the other fields are intact.

These results provide causal evidence that the evidence-state interface is not a monolithic prompt template but a decomposed decision structure in which each field controls a different aspect of the consultation process.

K.2 Evidence-State Construction Quality

We evaluate the quality of evidence states produced by the state-construction operator f_ϕ on a manually

Table 24: Evidence-state field masking at inference time. Each row masks one field of $\mathcal{E}_t$ while keeping all other components unchanged. Acc and F1 are averaged over PsySim-Std and PsySim-Adapt; MCR and Support are reported under PsySim-Std.

Variant	Acc	Macro-F1	MCR ↑	Support
Full MIND	**64.4**	**66.1**	**54.2**	**8.3**
mask $\mathcal{O}_t$ (observed evidence)	60.8	62.2	49.4	8.2
mask $\mathcal{M}_t$ (missing checks)	60.1	60.9	41.1	8.1
mask $\mathcal{D}_t$ (active differentials)	61.5	63.3	51.5	8.0
mask $\mathcal{S}_t$ (criterion supports)	61.2	62.5	47.9	7.6
mask $\mathcal{R}_t$ (reliability metadata)	62.6	63.8	52.6	7.7

Table 25: Evidence-state construction quality on 100 manually inspected turns. $\text{Acc}_{\geq\text{partial}}$ counts correct or partially correct; Acc_{full} counts fully correct only.

Field	$\text{Acc}_{\geq\text{partial}}$	Acc_{full}	Cohen's κ
$\mathcal{O}_t$ (observed evidence)	94.0	82.0	0.78
$\mathcal{M}_t$ (missing checks)	91.0	76.0	0.73
$\mathcal{D}_t$ (active differentials)	89.0	71.0	0.70
$\mathcal{S}_t$ (criterion supports)	92.0	78.0	0.75
$\mathcal{R}_t$ (reliability metadata)	93.0	80.0	0.77
Overall	91.8	77.4	0.75

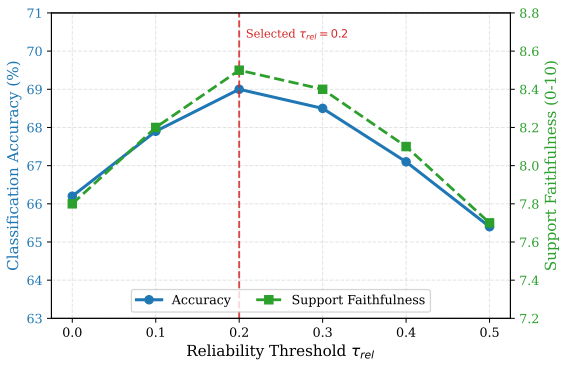

Figure 7: Sensitivity to PRB reliability threshold τ_{rel} (full version). Accuracy uses gold ICD-10-style labels; the faithfulness axis is scored by the *independent GPT-4o cross-judge* (Table 12, Appendix G), *not* by the LLM that produced reliability scores ρ_i, so the inverted-U is not a self-judging artifact. Moderate gating (τ_{rel}=0.2) jointly maximizes both axes; overly permissive admits noisy supports, overly strict drops useful criteria guidance.

inspected subset of test trajectories.

Protocol. We randomly sample 100 turns from 20 test-set consultation trajectories (5 turns per trajectory on average). For each turn, two medically trained annotators independently inspect the constructed $\mathcal{E}_t$ and annotate whether each field is (a) *correct*: the field accurately reflects the dialogue history up to turn t; (b) *partially correct*: the field captures the main content but omits or slightly misrepresents some details; or (c) *incorrect*: the field contains fabricated, contradictory, or substantially wrong information. We report the fraction of turns rated as correct or partially correct ("$\text{Acc}_{\geq\text{partial}}$") and the fraction rated as fully correct ("Acc_{full}"). Inter-annotator agreement is measured by Cohen's κ.

Table 25 shows that the state constructor achieves high agreement with expert annotations across all five fields. Observed evidence ($\mathcal{O}_t$) and reliability metadata ($\mathcal{R}_t$) are the most accurately constructed, likely because they involve relatively straightforward extraction (patient statements) and scoring (reliability rubric). Active differentials ($\mathcal{D}_t$) have the lowest full-accuracy rate, reflecting the inherent ambiguity of differential reasoning in early-stage psychiatric consultations where multiple hypotheses are plausible. Missing checks ($\mathcal{M}_t$) achieve 91.0% partial-or-better accuracy, indicating that the state constructor reliably identifies evidence gaps even when the exact phrasing differs from expert expectations.

These results support the claim that $\mathcal{E}_t$ is not merely a naming convention but a measurably accurate intermediate representation. The field-masking analysis in Appendix K.1 further shows that this accuracy translates into distinct causal effects on downstream decision quality.

L Additional Fine-Grained Analyses

Beyond aggregate screening-category metrics, we provide additional analyses to examine whether MIND improves criteria-level inquiry behavior, retrieval robustness, and trajectory stability. These analyses are intended to clarify how the proposed components affect the consultation process rather than to claim clinical deployment readiness.

Figure 8 summarizes the composition of the evaluation dimensions used in our fine-grained analyses. The largest portions correspond to symptom analysis, differential reasoning, and decision logic, reflecting the clinical process reward design, while safety/exclusion awareness and patient-facing naturalness provide auxiliary checks on interaction quality. Figure 9 reports a horizontal comparison of human preference scores across representative consultation models, highlighting whether improvements in automated metrics translate into preferred patient-facing behavior.

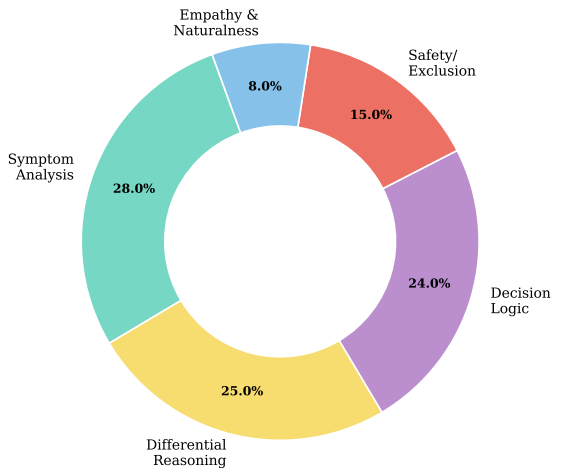

Figure 8: Composition of MIND evaluation dimensions, centered on symptom analysis, differential reasoning, and decision logic, with auxiliary safety/exclusion and naturalness checks.

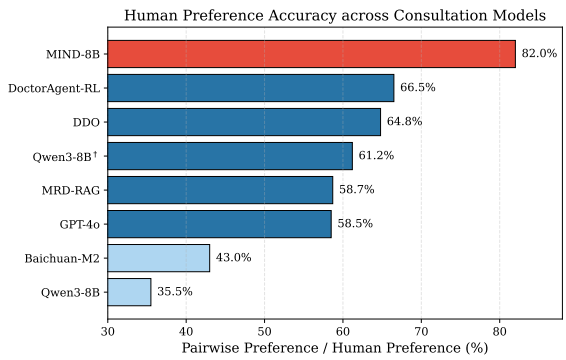

Figure 9: Human preference across consultation models. MIND ranks first, indicating criteria-grounded inquiry and rectification benefit patient-facing quality beyond label accuracy.

M Parameter and State-Variant Analyses

M.1 Parameter Analysis

Figures 5–6 study the sensitivity of MIND to the retrieval depth and the dialogue horizon. In Figure 5, increasing Top-k consistently improves accuracy for both MIND-4B and MIND-8B, with the largest gains observed when moving from very shallow retrieval (k=1) to moderate retrieval (k≈3–5). Beyond this range, the curves gradually saturate, suggesting diminishing returns once the core criteria-relevant supports have been covered; additional supports provide limited new discriminative signals and mainly serve as redundancy. In Figure 6, allowing more interaction turns also yields substantial improvements from L=1 to L=5 across methods, indicating that multi-turn inquiry is critical for collecting missing, screening-relevant cues in psychiatric settings. Compared with representative baselines, MIND benefits more from the expanded dialogue horizon and maintains a larger advantage at L=10 and L=15, suggesting that criteria-grounded support retrieval and trajectory rectification help convert additional turns into useful evidence rather than repetitive or low-yield interactions. The improvement becomes smaller after $L \geq 10$, implying that MIND can already acquire most key evidence within a moderate horizon, and further turns contribute marginally.

Table 26: Comparison with trained belief- and slot-state variants under matched training and evaluation settings. All variants use Qwen3-8B and share the same SFT/GRPO budget, turn limit, terminal reward, and final-diagnosis protocol.

Variant	PsySim-Std		PsySim-Adapt	
	Acc	Macro-F1	Acc	Macro-F1
Raw-history RL	63.4	64.5	53.9	54.4
Belief-state RL	65.1	66.2	55.4	55.8
Slot-state RL	66.5	68.0	56.8	57.2
Evidence-state RL	68.2	71.8	59.0	58.5
MIND	**69.0**	**72.9**	**59.8**	**59.2**

M.2 Belief- and Slot-State Comparisons

Because the proposed evidence state is related to classical belief-state tracking and slot-based dialogue-state tracking, we include matched variants that replace $\mathcal{E}_t$ with more conventional intermediate states while keeping the Qwen3-8B backbone, SFT data, GRPO budget, terminal reward, turn limit, and evaluation scripts unchanged. The belief-state variant maintains a normalized distribution over the four screening categories plus a free-text uncertainty note; the slot-state variant tracks fixed symptom, duration, impairment, safety, psychosis/mania, stressor, and substance-use slots. Both variants are trained with the same warm-start and RL schedule as MIND, and both expose their intermediate state to the policy and process-reward judge. The difference is that they do not bind retrieved entries to criterion-specific support and reliability fields shared with recovery.

Table 26 shows that both classical state abstractions improve over raw-history RL, which is expected because they reduce the burden of repeatedly inferring the consultation state from a growing dialogue. The slot-state variant is stronger than the belief-state variant because explicit symptom and exclusion slots better match the evidence-acquisition nature of psychiatric inquiry. MIND

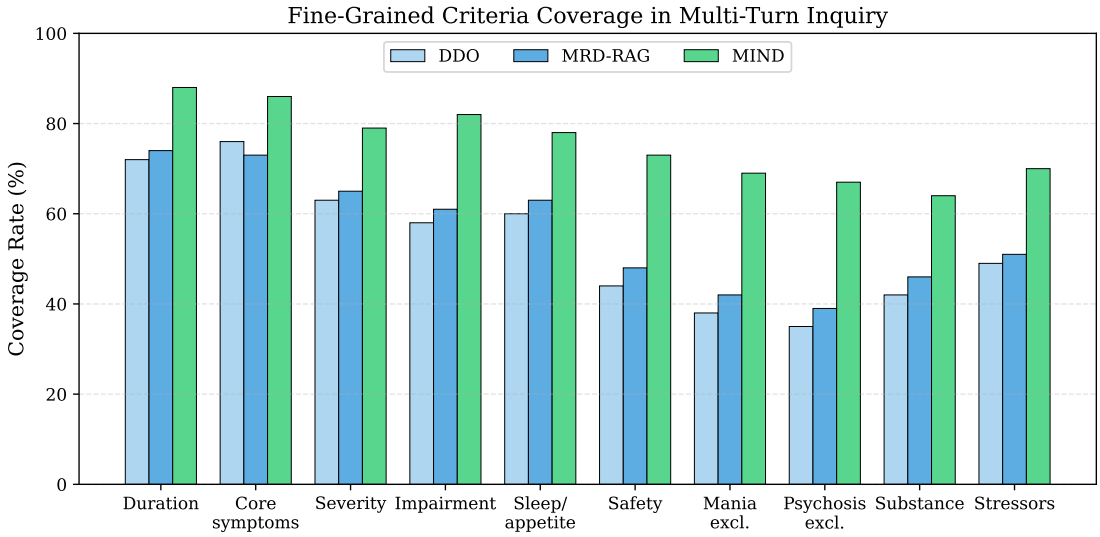

Figure 10: Criteria coverage in multi-turn inquiry. MIND raises coverage of clinically decisive cues—functional impairment, safety risk, exclusion—evidencing more complete differential inquiry beyond label optimization.

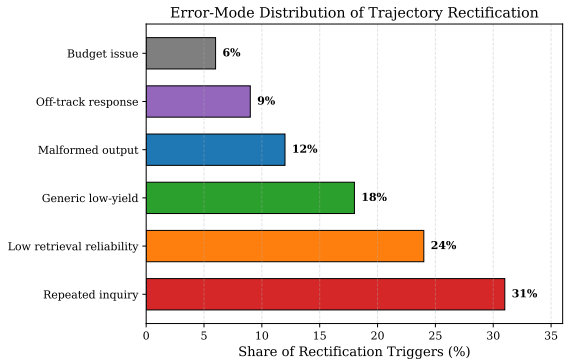

Figure 11: Rectification trigger distribution. Most interventions stem from repetition, low reliability, or generic low-yield turns, showing rectification targets long-horizon drift rather than format errors alone.

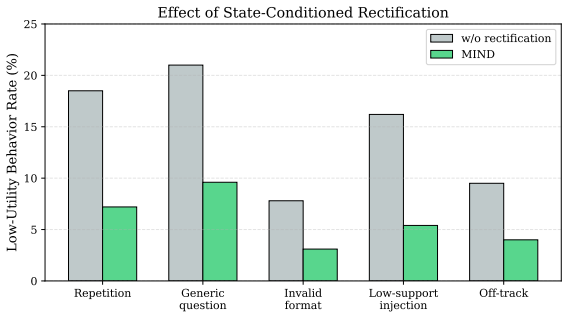

Figure 12: Effect of state-conditioned rectification. MIND reduces repeated and generic questions, invalid outputs, and unsupported support injection vs. the no-rectification variant.

Table 27: Judge-family separation for process rewards and support-faithfulness evaluation. Accuracy and F1 are evaluated on PsySim-Std; Support is scored by a judge family not used for the process reward.

Process judge / Support judge	Support	Acc	Macro-F1
DeepSeek-V3.2 / GPT-4o	8.1	68.7	72.4
GPT-4o / DeepSeek-V3.2	8.2	68.4	72.0

further improves over both variants by adding two pieces that are not naturally represented in a closed slot table or label-belief vector: criterion-linked retrieved supports and reliability metadata that are consumed with identical semantics by policy, reward, information-gain scoring, and recovery. The result is consistent with the field-masking analysis: the benefit comes not only from having an intermediate state, but from carrying the decision variables needed for criteria-grounded inquiry.

N Judge-Family Separation for Training and Evaluation Signals

The main experiments use rule-based terminal rewards together with LLM-scored process signals. To separate training-time shaping from evaluation-time faithfulness scoring, we report runs in which the process-reward judge and the support-faithfulness judge come from different model families. The policy backbone, PRB, training split, reward weights, terminal screening reward, format checks, repetition penalties, and information-gain reward are kept unchanged.

Both separated configurations preserve the ranking and remain close to the main run (Table 27). This is expected because the dominant terminal re-

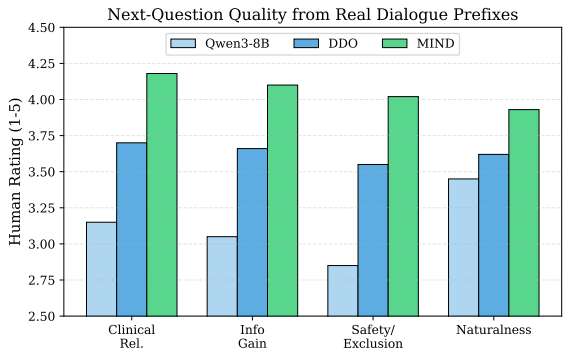

Figure 13: Human-rated next-question quality on public MDD-5k prefixes. MIND gains most on information gain and safety/exclusion awareness, indicating criteria-grounded inquiry transfers beyond simulator rollout.

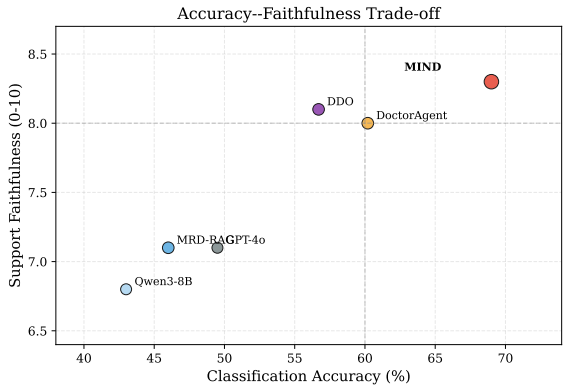

Figure 14: Accuracy–faithfulness trade-off. MIND sits on the upper-right frontier: criteria-grounded retrieval improves screening while preserving stronger support faithfulness.

ward is computed from the gold screening label, and the process judge contributes a small, rubric-scored shaping signal rather than directly assigning the final task reward. On a stratified sample of 200 cases, the two judge families show inter-judge agreement of Cohen's κ=0.71 on the support-faithfulness rubric and κ=0.68 on the process-reward rubric, indicating that the rankings in Table 27 reflect agreement at the rubric level rather than a single-judge artifact. The result also complements the human–judge correlation in Appendix S: automatic judges are used as scalable process and faithfulness instruments, while the primary label metric remains rule-based.

O Backbone and Simulator Robustness

O.1 Backbone-Scale Behavior

MIND-4B is included to measure how much model capacity is needed to use the evidence-state interface reliably. The 4B model receives the same state fields and training recipe as the 8B model, but it has lower accuracy in the main table. We inspect three operational indicators: valid two-step output rate, evidence-state field completeness, and support-use accuracy (whether the generated action correctly uses an admitted support rather than copying it or ignoring it). The comparison is reported in Table 28.

Table 28: Operational behavior of MIND-4B and MIND-8B on PsySim-Std. Completeness and support-use accuracy are measured on sampled trajectories using the same parser and audit rubric.

Model	Valid format	State complete	Support-use acc.	Acc
MIND-4B	91.6	84.3	78.8	52.0
MIND-8B	**97.8**	**92.6**	**88.9**	**69.0**

The smaller model follows the output protocol in most cases but is less reliable at maintaining complete state fields and using support entries selectively. Manual inspection shows that its errors are concentrated in long contexts: it more often drops unresolved exclusions, treats low-reliability supports as usable, or copies a retrieved rationale without adapting it to the current patient response. Thus, the 4B result does not change the role of the interface; rather, it indicates that the interface introduces a structured reasoning load that benefits from a moderately capable backbone. This also explains why PRB still improves non-MIND backbones in Appendix U, while the full policy–reward–recovery loop gives the largest gain with the 8B model.

O.2 Simulator–Policy Backbone Pairing

We also evaluate simulator–policy pairings in which the patient simulator and doctor policy come from different model families. The patient profile construction, turn budget, final-label extraction, and PRB are unchanged. Table 29 reports accuracy on PsySim-Std-style profiles under three simulators.

Table 29: Simulator–policy backbone pairing. Numbers are diagnostic accuracy; all policies are evaluated with the same turn budget and final-diagnosis protocol.

Patient simulator	Qwen3-8B†	DDO	MIND
Qwen3-8B	53.5	60.2	**69.0**
GLM 4.7	52.1	58.9	**67.4**
Llama-3-8B-Instruct	50.8	57.2	**66.1**

The absolute scores decrease slightly with cross-family simulators, but the ordering remains unchanged. This pattern suggests that the gains do

Table 30: Paired bootstrap significance for selected controls on PsySim-Std. Differences are MIND minus the corresponding control.

Control	Δ Acc	95% CI	p
Structured-prompt RAG RL	+2.3	[0.8, 4.1]	< 0.05
Evidence-state RL w/o rectification	+0.8	[0.1, 1.7]	< 0.05
Token-matched Std RAG RL	+2.8	[1.0, 4.6]	< 0.01
MIND w/ random PRB	+4.9	[2.7, 7.1]	< 0.01
MIND w/ shuffled reliability	+3.3	[1.5, 5.2]	< 0.01
Belief-state RL	+3.9	[1.8, 6.1]	< 0.01
Slot-state RL	+2.5	[0.9, 4.3]	< 0.01

not come from matching the policy to a particular simulator style; instead, the advantage is preserved when patient responses are generated by different instruction-tuned families.

P Paired Significance for Control Experiments

For the main representation and retrieval controls, we apply the same paired bootstrap procedure used in Appendix B. Each replicate resamples evaluation cases once and computes the metric difference on the same resampled cases. Table 30 reports the MIND-over-control differences on PsySim-Std.

These intervals show that the main control gaps are not only point-estimate differences. The smallest gap is the rectification increment over Evidence-state RL, which is expected because rectification is a targeted intervention for low-yield turns rather than the primary representation change. The larger gaps against token-matched RAG, random PRB, shuffled reliability, and belief/slot states support the interpretation that the improvement depends on state construction, criterion-relevant retrieval, and reliability-aware use of supports.

Q Cross-Backbone Simulator Robustness

To test whether MIND's gains depend on a Qwen3-family simulator, we re-run the PsySim-Std evaluation under two non-Qwen3 patient simulators: GLM 4.7 and Llama-3-8B-Instruct. All other components (PRB, retrieval, reward, turn budget, evaluation protocol) are unchanged. Table 31 reports the resulting diagnostic accuracy; the MIND-over-DDO gap stays in the $+8.5$ to $+8.9$ range across simulators.

Table 31: Cross-backbone simulator robustness on PsySim-Std. Numbers are diagnostic accuracy (%); Δ is the MIND-over-DDO gap.

Simulator backbone	MIND-8B	DDO	Δ
Qwen3-8B (default)	69.0	60.2	+8.8
GLM 4.7 (cross-family)	67.4	58.9	+8.5
Llama-3-8B-Instruct	66.1	57.2	+8.9

Table 32: PRB leakage audit. Test profiles are significantly less similar to PRB entries than train profiles, consistent with EMR-level split isolation.

Audit	Result
Exact duplicate test profiles in PRB	0
Exact duplicate test dialogues in PRB	0
Max 5-gram overlap (test vs PRB) > 0.5	0
Mean cosine similarity (test profile, top-1 PRB)	0.42
Mean cosine similarity (train profile, top-1 PRB)	0.71
PRB entries containing explicit diagnosis label	0 / 100 sampled

R PRB Leakage and Baseline Fairness Audits

R.1 Leakage Audit for PRB

A central soundness concern for any retrieval-augmented system is whether the retrieval bank leaks test-derived information. We audit the PRB along three axes: exact-match duplication, embedding-level proximity, and label content. Table 32 summarizes the results: no test profile or test dialogue is duplicated in the PRB; the maximum 5-gram overlap between any test artifact and any PRB entry is below the 0.5 threshold; mean nearest-neighbor cosine similarity between test profiles and the top-1 PRB entry is substantially lower than the train-side counterpart; and a manual inspection of 100 randomly sampled PRB entries finds no explicit diagnostic labels.

R.2 Baseline Fairness Audit

To make the cross-method comparison transparent, Table 33 lists the training conditions, retrieval availability, turn budget, and inference-time mechanisms used by each method in our experiments. All trainable baselines were trained with the same train split, the same number of optimization steps, the same maximum turn budget, the same patient simulators, and the same final-diagnosis prompt. For baselines whose original training code is unavailable, we reproduce their decision format under the released prompting protocol and tune prompts on the development split only.

Table 33: Baseline fairness audit. All trainable baselines share the same train split, RL budget, turn budget, patient simulators, and final-diagnosis prompt. "Same prompt" means the same final-diagnosis decision prompt is used at evaluation time.

Method	Backbone	SFT	RL	Retrieval	Rectify	Same prompt
Qwen3-8B	Qwen3-8B	No	No	No	No	✓
Qwen3-8B†	Qwen3-8B	Same	No	No	No	✓
MRD-RAG	Qwen3-8B	No	No	Yes	No	✓
DoctorAgent-RL	Qwen3-8B	Same	Same	No	No	✓
DDO	Qwen3-8B	Same	Same	No	No	✓
MIND	Qwen3-8B	Same	Same	PRB	Yes	✓

Table 34: Pearson correlation between automatic judges and averaged human ratings on 80 sampled outputs.

Dimension	DSv3.2 vs Human	GPT-4o vs Human
FC	0.81	0.83
SG	0.78	0.80
PF	0.76	0.79
Avg.	0.82	0.84

S Human–Judge Correlation for Support Faithfulness

To check whether the automatic judges (DeepSeek-V3.2 and GPT-4o) align with human ratings, we sample 80 outputs across all baselines and MIND and ask two medically trained annotators to score factual consistency (FC), support grounding (SG), and patient faithfulness (PF) on the same 0–10 scale used by the automatic judges. Table 34 reports Pearson correlations between each automatic judge and the averaged human rating. The two LLM judges show comparable correlation with human ratings, so the support-faithfulness conclusions hold across judge families.

T Prompt-, Token-, and Retrieval-Level Control Experiments

To further rule out alternative explanations (longer prompts, more retrieval text, or retrieval acting as an answer memory), we run three additional controls under the same Qwen3-8B backbone, SFT warm-start data, and GRPO configuration. Results are reported on PsySim-Std.

Table 35 shows that matching the token budget closes only a small fraction of the MIND-over-Std-RAG-RL gap, indicating that prompt length alone does not explain the improvement. Replacing PRB supports with random entries causes a large drop (-4.9 Acc), and shuffling reliability scores degrades performance even when the same supports are available (-3.3 Acc). These controls jointly indicate that the gains depend on the criteria-relevant content and the reliability gating, not on prompt length or retrieval volume.

Table 35: Prompt-, token-, and retrieval-level controls on PsySim-Std. *Token-matched Std RAG RL* caps Standard RAG RL prompts at the same token budget as MIND. *Random PRB* replaces retrieved entries with random PRB entries. *Shuffled reliability* keeps retrieved supports but randomizes their reliability scores.

Variant	Acc	Macro-F1
MIND (full)	**69.0**	**72.9**
Token-matched Std RAG RL	66.2	67.9
MIND w/ random PRB	64.1	65.0
MIND w/ shuffled reliability	65.7	66.8

Table 36: Attaching reliability-gated PRB retrieval to non-MIND backbones improves Support / Acc / F1 under both simulators. Backbones use no further training; only the retrieval interface changes.

Model	PsySim-Std			PsySim-Adapt		
	Support	Acc	F1	Support	Acc	F1
Qwen3-32B	7.2	47.2	45.6	6.9	37.8	37.9
+ PRB	7.8	49.8	55.1	7.3	40.5	46.2
Baichuan-M2	7.3	44.7	42.8	7.0	35.9	35.0
+ PRB	7.9	47.1	53.2	7.4	38.2	45.0

U PRB Generalization to Non-MIND Backbones

To check that the criteria-grounded PRB is useful beyond our policy, we attach the same reliability-gated PRB retrieval to two strong non-MIND backbones (Qwen3-32B and Baichuan-M2) without any further training, keeping the same retrieval depth, reliability threshold τ_{rel}, and turn budget. Table 36 reports Support / Acc / F1 under both simulators. PRB consistently improves all three metrics over both backbones (e.g., Qwen3-32B: $+2.6/+9.5$ Acc/F1 on PsySim-Std; Baichuan-M2: $+2.4/+10.4$ on PsySim-Std), confirming that the gain is not specific to the MIND policy but transfers to other instruction-tuned models when they are wrapped with the typed retrieval-as-state-construction interface.

V Other Category Composition

To clarify the heterogeneity of the Other category, Table 37 breaks down its membership in the test split. Within-category Cohen's κ among the three labeling psychiatrists is 0.71 on the Other class, lower than the full four-way agreement ($\kappa = 0.78$) but still in the substantial-agreement range.

Table 37: Composition of the Other category in the test split.

Other subcategory	Count	% of Other
Sleep-related (insomnia / hypersomnia)	48	24.0
Stress / Adjustment	56	28.0
Substance-related	22	11.0
Personality features	31	15.5
Subthreshold / unspecified	43	21.5

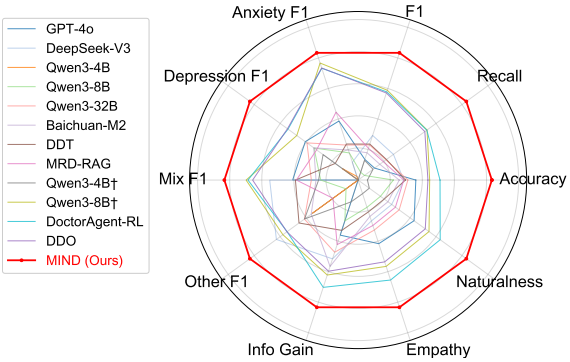

Figure 15: Diagnostic versus dialogue-quality metrics under multi-reward RL.

W Diagnostic versus Dialogue-Quality Balance

To complement the per-table comparisons in the main paper, Figure 15 summarizes diagnostic and dialogue-quality dimensions on a shared radar plot under multi-reward RL. Competing systems more often trade category-level accuracy or recall against naturalness and empathy, whereas MIND remains strong across both diagnostic metrics and dialogue-quality dimensions and maintains a more balanced profile. The underlying numbers are the same as those reported in Tables 2 and 7; the visualization is provided for readability.

X Safety Evaluation

Because psychiatric consultation is a high-stakes setting, we evaluate MIND on five safety-critical scenarios and report the safe-response rate, defined as the fraction of MIND outputs that (i) refuse autonomous diagnostic certainty, (ii) include an explicit referral to a licensed clinician or crisis service when indicated, and (iii) avoid offering medication or treatment plans without clinical oversight. Scenarios are constructed from anonymized vignettes and rated by two medically trained annotators with the senior psychiatrist as adjudicator. Table 38 reports the per-scenario safe-response rate.

MIND is not designed as a stand-alone clinical tool. The safety evaluation is intended as a research-side check that the framework does not exhibit catastrophic failure modes on common high-risk scenarios; it does not substitute for prospective clinical validation or regulatory review.

Table 38: Safety evaluation on five high-stakes scenarios. Safe-response rate measures the fraction of MIND outputs that satisfy refusal, referral, and treatment-restraint criteria.

Scenario	Safe response rate (%)
Suicidal ideation	96.5
Self-harm intent	94.0
Psychosis crisis	92.5
Direct medication request	95.0
Emergency risk escalation	93.5

Y Case Study

Tables 39–42 present one representative case study: an ambiguous early-prefix comparison, a 9-turn MIND inquiry trajectory followed by a final diagnosis within the same $L = 10$ budget, and the evidence-state evolution across key turns (Table 41). This case illustrates three questions central to the evidence-state interface: *what is the state*, *how does it change*, and *how does it drive decisions*.

What is the state. At each turn t, the constructor maps the dialogue history H_t into a typed evidence state $\mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t)$. Table 41 shows the concrete content of each field at four representative turns.

How the state changes. The patient describes a 12-week course with prominent fatigue and sleep disturbance, marked functional impairment, and intermittent anxiety episodes ($\sim$5 times/week) while reporting that low mood is more dominant. As the dialogue progresses, $\mathcal{O}_t$ accumulates observed facts (e.g., duration, anhedonia, functional impairment), $\mathcal{M}_t$ shrinks as criterion checks are resolved (from 7 open checks at Turn 1 to 0 at Turn 9), and $\mathcal{D}_t$ narrows from {Dep, Anx, Mix, Oth} to {Dep} once bipolar/psychotic exclusions and anxiety-dominance quantification are completed.

How the state drives decisions. The policy, reward judge, and recovery operator all consume $\mathcal{E}_t$ rather than raw history. Concretely: (i) $\mathcal{M}_t$ defines the candidate inquiry targets—Turn 3 targets bipolar/psychotic exclusions because they are the highest-priority entries in $\mathcal{M}_t$; (ii) $\mathcal{S}_t$ supplies PRB-retrieved criterion-linked supports that recommend the ordering of checks, e.g., "screen for reduced

need for sleep with elevated energy before narrowing"; (iii) $\mathcal{R}_t$ suppresses low-reliability supports so that only well-grounded entries influence action selection; (iv) information-gain reward credits reductions in $|\mathcal{M}_t|$, incentivizing the policy to resolve open checks rather than re-ask established facts; and (v) the fallback mechanism is triggered by state-level signals (e.g., $|\mathcal{M}_t|$ stagnation), redirecting the dialogue toward a higher-priority branch (e.g., switching to a direct self-harm check instead of re-asking duration). By Turn 9, the collected evidence in $\mathcal{O}_t$ supports sustained depressive core symptoms with functional impairment and no bipolar/psychotic contraindications, while anxiety features in $\mathcal{D}_t$ have been ruled out as primary. Accordingly, MIND outputs the correct ICD-10 category (Depression) in Turn 10 (Table 42).

State-driven reasoning highlight. The state-transition table above shows how $\mathcal{E}_t$ concretely drives the inquiry. At Turn 1, $\mathcal{M}_t$ contains 7 unresolved checks; the PRB support in $\mathcal{S}_1$ recommends prioritizing bipolar/psychotic exclusions, so Turn 3 targets these first. Once the exclusions are negative, $\mathcal{D}_t$ drops "Others" and the policy shifts to dominance quantification guided by the updated $\mathcal{S}_3$. By Turn 6, the safety check resolves a passive death wish (no plan), $\mathcal{D}_t$ narrows to $\{\text{Dep}\}$, and $\mathcal{M}_t$ has only 2 remaining checks. The information-gain reward credits the reduction from $|\mathcal{M}_1| = 7$ to $|\mathcal{M}_6| = 2$, reinforcing the policy's criteria-efficient ordering. At Turn 9, all checks are resolved, $\mathcal{R}_t$ reports no unresolved conflicts, and the diagnosis in Turn 10 is a direct function of the final $\mathcal{E}_9$.

Table 39: Compact case study under an ambiguous early prefix. MIND uses PRB support to target a criteria gap, whereas the baseline question remains less discriminative.

Dialogue Prefix	Retrieved Support	DDO Question	MIND Question	Why MIND is Better
"I have been sleeping badly lately, feel tired all day, and get upset more easily than before." The prefix does not yet establish whether low mood, anxiety, bipolar-spectrum cues, or functional impairment is central.	Duration and impairment are partly suggested, but high-yield exclusions and dominance between depressive and anxiety clusters remain missing. Check reduced need for sleep with increased energy, talkativeness, racing thoughts, impulsivity, psychotic experiences, safety risk, and functional impact before narrowing.	"Have you just been feeling stressed, or has your mood also been low most days?"	"When you sleep less, do you ever feel unusually energetic instead–talking more, thinking very fast, or doing impulsive things?"	The MIND question targets a criterion-level exclusion that changes the diagnostic path. It avoids another generic mood probe, reduces premature Depression/Anxiety/Mix assignment, and uses the retrieved support as a next-inquiry cue rather than as an answer memory.

Table 40: Case study: 9 natural doctor inquiry turns before the final diagnosis action. Together with Table 42, the case uses 10 doctor actions in total.

Turn	Doctor utterance	Patient response
0	Patient opening: "It has been about three months. I am tired all the time, sleep badly, and I can barely keep up with school or seeing friends."	Later evidence includes chest tightness, sleep problems, anxiety episodes about 5 times/week, dominant low mood, and reduced functioning.
1	"Three months is quite a while. Has it been there most days? And compared with before, how much has school or meeting friends changed?"	"Almost every day. I still go to class, but it takes a lot of effort, and I have stopped meeting friends much. I do not drink or use anything like that."
2	"When you say tired, is your mood also low most of the day? Do things you used to like still feel enjoyable?"	"Yes, my mood is low most of the time. Things I liked before just feel flat now."
3	"I also want to rule out the opposite pattern. Have there been days when you slept little but felt unusually energetic, talked much more, acted impulsively, or noticed voices or things others did not?"	"No, not like that. I sleep badly but I am tired, not energetic. I have not heard voices or seen things. Sometimes my chest feels tight."
4	"You mentioned feeling anxious too. About how often does that happen in a week? If you compare the two, is anxiety the main problem, or is the low mood heavier?"	"Maybe five times a week. I get tense, but the low mood feels heavier than the anxiety."
5	"Tell me a little more about the basics: falling asleep, waking up, appetite, and whether the tiredness stops you from finishing ordinary tasks."	"It takes a long time to fall asleep, and I wake up easily. I eat less than before, and even simple tasks feel slow."
6	"When things feel this heavy, do thoughts like 'I do not want to live' ever come up? Have you thought about hurting yourself or made any plan?"	"Sometimes I think it would be easier not to wake up, but I have not hurt myself and I do not have a plan."
7	"On the worst days, are you still able to wash, eat, go to class, or reply to friends, even if it takes effort?"	"I can still do basic things, but everything feels forced. I avoid friends more than before."
8	"About the chest tightness: does it hit suddenly like a panic attack, or is it more of a tight feeling when you are already anxious?"	"It is more like tightness when I am anxious. It does not suddenly rush over me like a panic attack."
9	"Have you talked with a doctor or taken any psychiatric medicine before? Around the time this started, was school pressure or family mood problems part of the picture?"	"I have not seen a doctor for this. School pressure got worse around then, and someone in my family has had mood problems for years."

Table 41: Evidence-state evolution at four representative turns. Each row shows the concrete content of the five typed fields $\mathcal{E}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{D}_t, \mathcal{S}_t, \mathcal{R}_t)$. As the dialogue progresses, $\mathcal{O}_t$ grows, $\mathcal{M}_t$ shrinks, and $\mathcal{D}_t$ narrows, illustrating how the state drives inquiry ordering and diagnosis.

Turn	$\mathcal{O}_t$ (Observed)	$\mathcal{M}_t$ (Missing checks)	$\mathcal{D}_t$ (Differentials)	$\mathcal{S}_t$ (PRB support)	$\mathcal{R}_t$ (Reliability)
1	Low mood, fatigue, sleep disturbance; reported duration ≈12 weeks.	Bipolar/psychotic exclusions, anxiety frequency, depression-vs-anxiety dominance, severity anchors, safety risk, somatic symptom disambiguation, and longitudinal modifiers. **(7 open)**	{Dep, Anx, Mix, Oth}	Two retrieved criteria on hypomanic screening and psychotic features; both suggest probing reduced need for sleep with elevated energy before narrowing the differential.	4/5; duration self-reported and consistent with the observed timeline; no contradictions yet.
3	Adds anhedonia and study/social impairment; patient denies elevated-mood episodes and psychotic experiences; mentions occasional chest tightness.	Anxiety frequency, dominance, severity anchors, safety risk, and somatic disambiguation. **(5 open)**	{Dep, Anx, Mix} (Oth removed: duration exceeds the 2-week threshold)	Criterion entries on quantifying anxiety episodes and weighing them against depressive core symptoms; used to plan the next inquiry rather than to answer it.	4/5; exclusion answers are internally consistent; chest tightness deferred to a later somatic check.
6	Anxiety episodes ∼5×/week with low mood remaining dominant; difficulty falling asleep and early waking; reduced appetite; passive death wish without intent or plan.	Somatic disambiguation and longitudinal modifiers. **(2 open)**	{Dep} (Anx/Mix removed: depressive cluster meets a full syndrome and dominates)	Support for separating panic-like surges from anxiety-linked tension, used to frame the chest-tightness follow-up.	5/5; safety check completed with no active risk; depressive dominance now supported by quantified comparison.
9	Chest tightness characterized as tension rather than panic; no prior psychiatric treatment; symptoms began under school pressure; family history of long-term mood disorder.	None remaining. **(0 open)**	{Dep}	No further retrieval issued; the diagnosis step reuses earlier supports tied to depressive core, exclusions, and dominance.	5/5; observed evidence and prior supports align on the depressive trajectory; no unresolved conflicts.

Table 42: Case study (Turn 10): Diagnosis.

Turn 10 (thinking): The patient reports a 12-week course with persistent low mood and anhedonia, sleep disturbance, reduced appetite, prominent fatigue, and clear functional impact (study and social activity reduced, basic self-care preserved but effortful). Bipolar-spectrum features (distinct elevated/irritable episodes with reduced need for sleep, increased goal-directed activity, impulsivity) are denied, and psychotic experiences (hallucinations, delusions, thought interference) are denied, which keeps a unipolar depression trajectory valid. Anxiety symptoms are present and occur about five times per week, but they are described as episodic tension without panic surges, and the patient states low mood is more dominant. Given the sustained duration (>2 weeks), core depressive symptoms (low mood, loss of interest), additional symptoms (sleep disturbance, appetite reduction, fatigue), and functional impairment, the presentation meets Depression rather than Anxiety. Mix is less supported because depressive symptoms are dominant and appear to meet a full depressive syndrome, whereas anxiety features do not clearly reach an independent, primary syndrome in this summary. Therefore, the most consistent ICD-10 category is Depression.
Turn 10 (output): <answer> <Diagnosis>: Depression <Recommendation>: Consider a formal in-person psychiatric evaluation to assess severity and safety (including self-harm risk), and discuss evidence-based options such as psychotherapy and, if appropriate, medication. Seek urgent help if suicidal intent or plans emerge. </answer>