COMPASS: Compressing Clinical Observations and Mitigating Premature Assumptions in Small-scale Diagnostic Agents

Anonymous ACL submission

Abstract

Interactive differential diagnosis requires models to sequentially gather clinical evidence, maintain evolving patient states, and update disease rankings under incomplete information. Existing multi-agent DDx systems mainly rely on large language models, making them costly and difficult to deploy locally in privacy-sensitive primary-care settings with limited GPU resources. When applied to smaller models, they often suffer from context overload, large candidate spaces, and premature closure bias. We propose **C**ompressing clinical **O**bservations and **M**itigating **P**remature **A**ssumptions in **S**mall-scale diagnostic agent**S**, a blackboard-style modular framework tailored for 8B-scale models. COMPASS decomposes interactive diagnosis into three roles: an Actor for inquiry, retrieval, and diagnosis; a Compressor for converting multi-turn dialogues into structured clinical belief states; and a Critic for evidence-level review of preliminary diagnoses to reduce premature closure. With role-specific tuning and GRPO-optimized Actor policy, COMPASS achieves strong performance on DDxPlus, iCraft-MD, and RareBench, outperforming 70B prompt-based multi-agent baselines on GTPA@1 while approaching GPT-4o performance at a lower inference cost.

1 Introduction

Differential Diagnosis (DDx) is central to clinical decision-making, where physicians progressively narrow a broad space of plausible diseases under incomplete and evolving evidence (Rhoads and Penick, 2017). This process is intrinsically interactive: clinicians elicit additional symptoms, verify or rule out competing hypotheses, incorporate medical knowledge, and continuously revise their diagnostic beliefs. Although Large Language Models (LLMs) have shown strong performance on static, single-turn medical question answering, real clinical consultations require sustained evidence acquisition and stateful reasoning across multiple turns. This has motivated a growing shift from static diagnosis toward Interactive DDx, where a model must decide what to ask, when to retrieve external knowledge, and when the available evidence is sufficient for diagnosis (Feng et al., 2026).

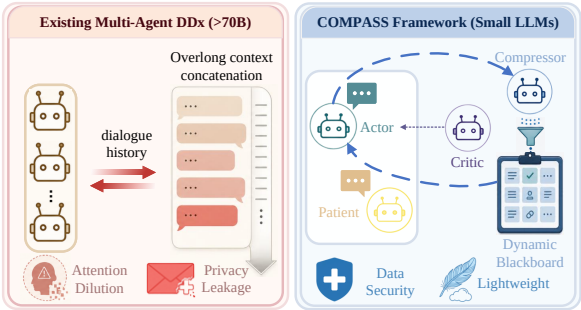

Figure 1: Overview of COMPASS for lightweight interactive differential diagnosis.

Building LLM-based consultation agents for primary care remains difficult, especially when deployment must rely on small or locally hosted models. Existing systems face three coupled challenges. First, many interactive DDx agents depend on large foundation models exceeding 70B parameters or closed-source APIs, leading to high inference costs and limited suitability for privacy-sensitive, resource-constrained clinical environments (Thirunavukarasu et al., 2023). Second, multi-turn consultation naturally produces long dialogue histories. Directly concatenating these histories into the context window introduces substantial redundancy and attention dilution, which is particularly harmful for small models as the interaction length grows (Liu et al., 2024). Third, diagnostic agents are vulnerable to premature closure and confirmation bias. Under a large and uncertain dis-

Code: https://anonymous.4open.science/r/COMPASS-FE26.

ease space, small models tend to anchor on early, common hypotheses, ask repetitive confirmatory questions, and fail to preserve rare or atypical alternatives (Liu et al., 2025). These issues indicate that simply scaling down existing multi-agent designs is insufficient for reliable interactive diagnosis.

To address these limitations, we propose COMPASS, a lightweight modular framework for small-scale diagnostic agents. As shown in Figure 1, COMPASS replaces raw dialogue-history concatenation with a dynamic blackboard state, which serves as a compact and structured clinical memory. A Compressor Agent incrementally transforms multi-turn interactions into high-density clinical belief states, including confirmed findings, ruled-out evidence, unknown-but-asked information, medical history, and exposure factors. Based on this structured state rather than an ever-growing transcript, the Actor Agent makes turn-level decisions, thereby reducing context redundancy and improving evidence-tracking stability in small models.

To mitigate premature closure and improve long-horizon consultation behaviour, COMPASS further integrates adversarial evidence review with actor-level strategy optimisation. When the Actor produces a preliminary DDx ranking, the Adversarial Critic reviews the blackboard state, candidate diseases, and retrieved knowledge to identify insufficient exclusionary evidence, overlooked alternatives, or unsupported ranking shifts. Its feedback is written back to the blackboard, guiding the Actor to continue targeted inquiry or retrieval before final diagnosis. Beyond this inference-time correction, the Actor is optimised with Group Relative Policy Optimization (GRPO), whose trajectory-level rewards jointly measure diagnostic ranking quality, interaction efficiency, and behavioural validity. This enables an 8B-scale model to learn when to enquire, retrieve, or diagnose under incomplete evidence, while avoiding the value-network overhead of PPO-style reinforcement learning.

In summary, our contributions are threefold:

- We propose **COMPASS**, a lightweight interactive DDx framework tailored for small-scale language models. By combining dynamic blackboard compression with adversarial evidence review, COMPASS addresses context bloat and premature closure in multi-turn diagnostic reasoning.
- We introduce a GRPO-based strategy alignment method for interactive medical agents, enabling the Actor to optimize inquiry, retrieval, and diagnosis decisions under multi-turn trajectory-level rewards.
- We evaluate **COMPASS** on three interactive DDx benchmarks: DDxPlus, iCraft-MD, and RareBench. Results show that it substantially improves Qwen3-8B, reduces token and memory overhead, and outperforms 70B multi-agent baselines such as MEDDxAgent on key diagnostic metrics including GTPA@1, demonstrating the feasibility of accurate, locally deployable clinical consultation agents.

2 Related Work

2.1 From Static Reasoning to Interactive DDx: The Monolithic Bottleneck

DDx provides the reasoning foundation for clinical decision-making, where clinicians compare diseases with overlapping symptoms and progressively rule out less plausible hypotheses using patient complaints, electronic health records, and other clinical evidence (Bowen, 2006). Recent advances in LLMs have substantially improved automated DDx reasoning, with frontier models such as GPT-4o, Gemini 2.5, and DeepSeek-R1 demonstrating strong clinical reasoning ability in complex diagnostic settings, including evaluations involving comorbidities and EHR-derived cohorts (Singhal et al., 2023; Tu et al., 2024; Thio et al., 2026). However, most medical LLM studies still follow a static, single-turn reasoning paradigm: they receive isolated clinical fragments or predefined question sets and directly produce diagnostic conclusions (Hsu et al., 2026), which does not match the iterative nature of real clinical consultations. Early-stage diagnosis is often characterized by incomplete, ambiguous, or evolving evidence, requiring systems to actively acquire missing information, update disease hypotheses, and determine when the evidence is sufficient for diagnosis (Zhou et al., 2026; Wang et al., 2026). Monolithic LLMs remain limited in evaluating information gain, planning inquiry trajectories, and maintaining consistent diagnostic states across turns, thereby restricting their applicability in realistic incomplete-information DDx scenarios (Guo et al., 2026).

2.2 Interactive Medical MAS and Lightweight Reinforcement Learning Alignment

To support dynamic inquiry, recent work has introduced Multi-Agent Systems (MAS) that decom-

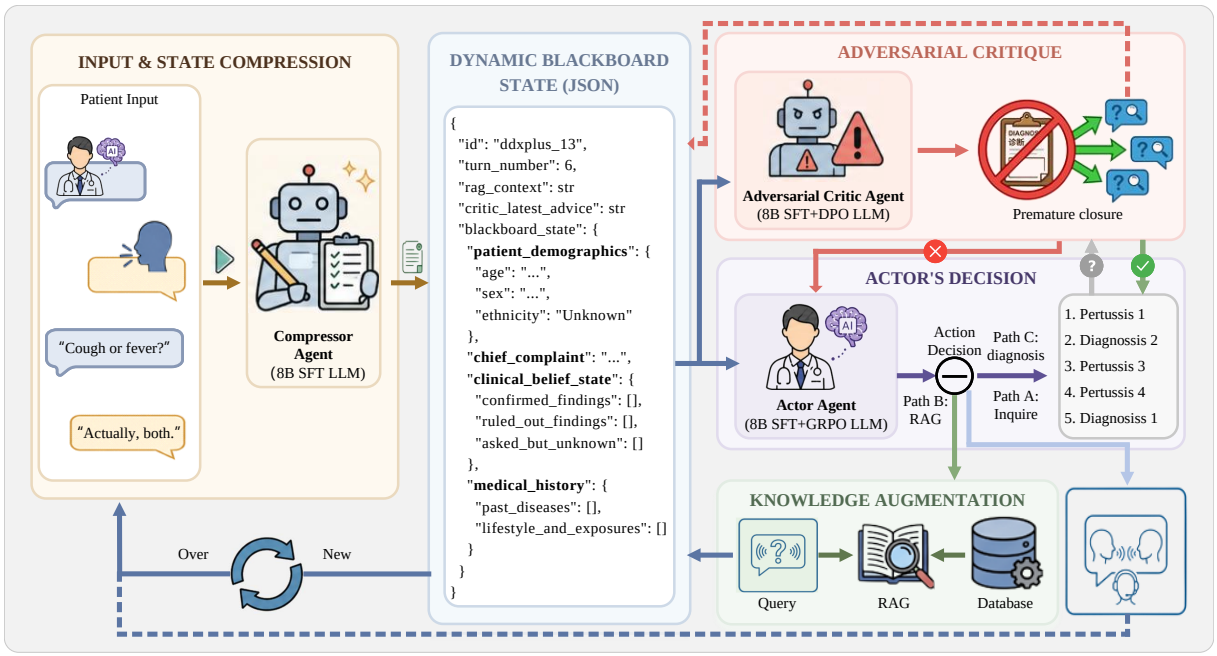

Figure 2: Architecture of COMPASS. The Actor routes each turn to inquiry, retrieval, or diagnosis, while the Compressor updates the dynamic blackboard and the Critic reviews preliminary rankings.

pose clinical consultation into specialized roles (Vatsal et al., 2026). For example, MEDDxAgent coordinates history-taking and knowledge agents through a centralized DDxDriver (Rose et al., 2025), while MedAgents aggregates multiple domain-oriented specialists for diagnostic consensus (Tang et al., 2024). Although such role decomposition improves medical reasoning, existing medical MAS typically rely on large models, centralized orchestration, and frequent inter-agent communication. As consultation length increases, the central controller must jointly handle patient complaints, dialogue tracking, agent coordination, and hypothesis aggregation, which can cause context saturation, attention shifts, and unstable reasoning chains (Liu et al., 2024). Their dependence on closed-source cloud APIs also raises inference-cost, latency, and privacy concerns, limiting deployment in local and resource-constrained clinical environments (Thirunavukarasu et al., 2023; Clusmann et al., 2023). Meanwhile, training interactive agents requires optimizing long-horizon consultation behavior beyond static imitation. SFT can teach format compliance and basic dialogue patterns, but offers limited incentives for active inquiry and delayed diagnostic correction. PPO-style RLHF is costly due to value-model training, while DPO mainly relies on offline preference pairs and is less suitable for multi-turn decision processes (Ouyang et al., 2022; Tian et al., 2024; Rafailov et al., 2023; Chen et al., 2023; Bao et al., 2023). GRPO estimates advantages from relative rewards within sampled output groups, eliminating the value network and making it suitable for lightweight policy optimization (Shao et al., 2024). Building on this property, we apply GRPO to optimize the Actor's inquiry, retrieval, and diagnosis policy in interactive DDx, using trajectory-level rewards that jointly capture diagnostic accuracy, consultation efficiency, and behavioral validity.

3 Methodology

LLMs face context overload, excessively large candidate disease spaces, and premature closure bias in interactive DDx. To address these issues, we propose COMPASS, a modular multi-agent framework designed for 8B-scale models (Figure 2). By adopting a blackboard architecture, the system utilizes a JSON-structured, globally shared dynamic state as clinical memory to maintain patient information, compressed clinical evidence, and collaborative states in real time.

Based on the blackboard state, the Actor Agent dynamically executes three action routes:

- Path A (Inquire): The Actor initiates follow-up questions, triggering the Compressor Agent to extract high-fidelity structured entities from the dialogue and update the blackboard.

- Path B (Retrieve): For knowledge blind spots or insufficient candidate coverage, the Actor uses retrieval to access external medical knowledge and, when needed, expand the candidate working set from the full normalized diagnosis option set. The retrieved evidence and candidate updates are summarized into the blackboard.
- Path C (Diagnose): The Actor submits a candidate disease ranking for adversarial review by the Critic, which either mandates further inquiry if evidence is insufficient or outputs the final diagnosis.

3.1 Clinical Reasoning Actor Agent

As the central decision-maker in COMPASS, the Actor Agent performs sequential multi-task reasoning under medical uncertainty. At each interaction turn t, reading the global blackboard state S_t, it dynamically selects an action from {Inquire, Retrieve, Diagnose}: the Actor invokes Inquire when evidence is insufficient to distinguish high-ranking diseases or when responding to Critic warnings. It asks the simulated patient one or two targeted questions while avoiding redundant queries. Retrieve is triggered when diagnostic criteria are ambiguous or rare diseases are suspected. The Actor queries PubMed with one to three diseases from the full candidate list, such as "Goldenhar syndrome," and condenses the retrieved knowledge into the blackboard to supplement long-tail medical knowledge. Diagnose produces the current differential diagnosis ranking, usually containing the top ten most likely diseases.

To ensure the 8B model masters fundamental clinical workflows and strictly outputs standardized JSON, we perform supervised fine-tuning on the Actor using "Context-Decision Pairs" derived from real-world consultation trajectories. This training internalizes execution rules, such as triggering a Retrieve action for knowledge blind spots or an Inquire action upon Critic rejection—thereby guaranteeing absolute stability in instruction formatting.

3.2 Adversarial Critic Agent

LLMs are highly susceptible to Premature Closure during multi-turn interactions, often finalizing diagnoses without crucial exclusionary evidence. To address this, we introduce an Adversarial Critic agent.

Upon the Actor executing a Diagnose action, control flow is mandatorily redirected to the Critic. Acting as a rigorous, label-free reviewer, the Critic evaluates the blackboard state, candidate lists, the Actor's Top-K diagnosis, and RAG contexts to identify ignored evidence, omitted candidates, premature diagnoses, or unreasonable generalizations. If a logical flaw is detected, such as failing to exclude pertussis-specific spasmodic coughing in a febrile patient, the Critic rejects the diagnosis and posts specific adversarial advice to the blackboard, explicitly forcing the Actor to revise its judgment and resume inquiry. When the current candidate pool fails to cover diseases that can plausibly explain the accumulated blackboard evidence, the Critic rejects the diagnosis and triggers a retrieval-based candidate expansion rather than forcing a final decision within the initial pool.

3.3 Blackboard State Compression

In multi-turn medical dialogues, natural language inputs rapidly saturate the context windows of small models, leading to critical information loss and attention dilution. To address this, we introduce the Compressor Agent to manage system memory. Acting as a background listener, it triggers upon new clinical inputs (e.g., doctor-patient dialogues, critic feedback, or RAG results), utilising the previous state and latest interactions to execute append-centric incremental updates.

To prevent invalid or erroneous information from accumulating, the Compressor continuously performs deduplication and conflict resolution. It converts lengthy colloquial conversations into structured medical entities and updates five core blackboard slots: confirmed findings, ruled-out findings, asked-but-unknown information, past diseases, and lifestyle or exposure history.

Specifically, the asked_but_unknown slot logs medical topics the system has enquired about but remain unconfirmed because the patient was unaware or the information was unavailable. This mechanism effectively prevents redundant questioning during subsequent interactions. Through dynamic compression and conflict normalisation, the structured blackboard state eliminates natural language redundancy typical of multi-turn interactions, minimising the token footprint required for the Actor's decision-making.

Finally, we fine-tune the Compressor on paired dialogue snippets and structured JSON updates to ensure high-fidelity medical entity linking, filter colloquial noise, and strictly prevent information

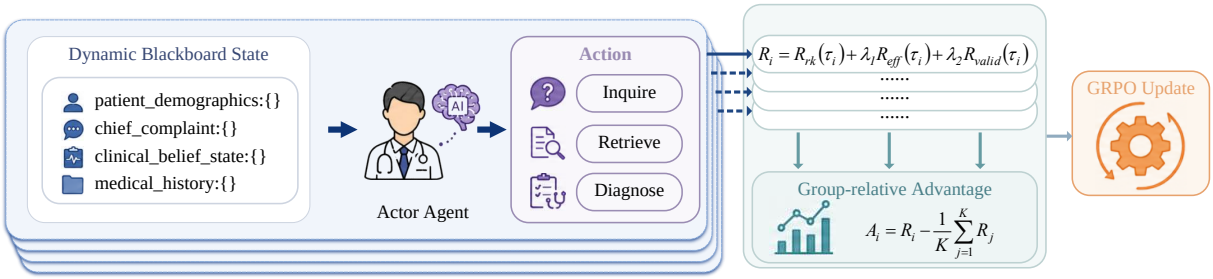

Figure 3: GRPO training of the Actor. Multiple trajectories are sampled from the same initial blackboard state, while only the Actor policy is updated.

loss or semantic hallucinations during state transitions.

3.4 Alignment GRPO Optimization

Supervised Fine-Tuning (SFT) provides the Actor with basic instruction-following, action formatting, and consultation patterns, but it only imitates static trajectories and does not directly optimize long-horizon diagnostic outcomes. We therefore apply GRPO to align the Actor's inquiry, retrieval, and diagnosis policy after SFT. At each turn t, the Actor observes only the current dynamic blackboard state S_t and samples an action

$$a_t \sim \pi_\theta(a_t \mid S_t), \tag{1}$$

where $a_t \in \{\text{INQUIRE}, \text{RETRIEVE}, \text{DIAGNOSE}\}$. Each patient case is treated as an episode. Given the initial blackboard state S_0, the frozen Patient Simulator, Compressor, RAG module, and Critic define the environment transition, while only the Actor policy π_θ is updated.

For each initial blackboard state, GRPO samples a group of K complete consultation trajectories, denoted as $\{\tau_i\}_{i=1}^{K}$. After the final diagnosis, each trajectory receives a scalar reward:

$$R(\tau_i) = R_{\text{rank}}(\tau_i) + \lambda_1 R_{\text{eff}}(\tau_i) + \lambda_2 R_{\text{valid}}(\tau_i). \tag{2}$$

Here, R_{rank} evaluates whether the ground-truth disease appears in the final Top-10 diagnosis list and assigns larger rewards to higher ranks. R_{eff} penalizes unnecessary inquiries, ineffective retrievals, and repeated Critic rejections to encourage concise consultations. R_{valid} penalizes invalid JSON, unnormalized disease names, repeated questions, and ineffective retrievals, thereby constraining the action space of the small model.

Unlike PPO, GRPO does not require an additional value network. Instead, it estimates the advantage of each trajectory by normalizing rewards within the sampled group:

$$\hat{A}_i = \frac{R(\tau_i) - \text{mean}(\{R(\tau_j)\}_{j=1}^{K})}{\text{std}(\{R(\tau_j)\}_{j=1}^{K}) + \epsilon}. \tag{3}$$

This group-relative advantage provides a trajectory-level learning signal without fitting a separate critic model, which is particularly suitable for 8B-scale policy optimization. The Actor is optimized with a clipped GRPO objective regularized by the KL divergence from a reference policy:

$$\mathcal{J}(\theta) = E\left[\frac{1}{K}\sum_{i=1}^{K} \min\Big(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right]. \tag{4}$$

In implementation, this ratio is computed over the Actor-generated action tokens, while the Patient Simulator, Compressor, and Critic remain frozen and gradient-free. The KL term stabilizes policy updates by constraining the optimized Actor from drifting excessively from the SFT-initialized reference policy. Freezing non-actor modules reduces credit-assignment noise, allowing rewards to be attributed to the Actor's decisions on when to inquire, retrieve, and diagnose.

Driven by GRPO's joint optimization of diagnostic accuracy, interaction costs, and behavioral validity, the Actor autonomously converges on a standardized, physician-like diagnostic workflow. This trajectory typically proceeds in three stages: Initial Diagnosis: The Actor conducts multi-turn interactions, proactively halting to output an initial Top-10 candidate list once sufficient information is gathered. Targeted Retrieval: Based on these preliminary results, the Actor initiates RAG for low-confidence diseases to generate a revised Top-10 list. Adversarial Review: The Critic evaluates this revised list for premature closure or missed diagnoses. Consequently, the Actor outputs the final diagnosis if the review passes.

4 Experiments

4.1 Research Questions and Evaluation Protocol

Following the MEDDxAgent cross-domain DDx benchmark setting, we evaluate COMPASS on DDxPlus, iCraft-MD, and RareBench, with detailed dataset statistics and representative examples provided in Appendix B.

Our evaluation is organized around two questions: whether an 8B-scale backbone has sufficient diagnostic capacity for medical DDx, and whether COMPASS can convert this capacity into an efficient and controllable interactive consultation system under incomplete information. To answer them, we adopt a hierarchical protocol that first examines Qwen3-8B in complete, static patient-profile settings, then evaluates the reliability of the dynamic blackboard as a compact clinical memory, and finally measures the end-to-end performance of COMPASS in interactive DDx scenarios. This design allows us to separate backbone capability, state-maintenance reliability, and system-level gains from retrieval, Critic review, and role-specific optimization.

4.2 Compared Systems

We compare three groups of systems. First, we evaluate monolithic Qwen3-8B under FullProfile and candidate-control settings to measure the backbone's diagnostic capacity without interaction. Second, we use MEDDxAgent as the external multi-agent baseline, including its Llama3.1-8B, Llama3.1-70B, and GPT-4o variants under the reported optimal fixed-iteration setting, since it shares the same DDxPlus, iCraft-MD, and RareBench evaluation environment. Since the original MEDDxAgent paper does not report GTPA@5, we obtain this metric from our reproduction under the same evaluation setting. Third, we evaluate COMPASS in two stages: COMPASS-Prompt, which uses frozen Qwen3-8B modules controlled only by prompting, and COMPASS-GRPO, which applies role-specific SFT followed by GRPO optimization of the Actor policy. This comparison isolates the gains from interaction design, role-specific training, and policy optimization under an 8B backbone.

4.3 Evaluation Metrics

We utilize GTPA@k ($k \in \{1, 5, 10\}$) as the primary metric to indicate whether the ground-truth disease appears within the top-k candidates. For detailed comparisons, we report AvgRank—the true disease's average position among the top-10 outputs (recorded as 11 if absent).

In RAG or multi-stage setups, ΔProgress quantifies the ground-truth's ranking improvement, with positive values denoting advancement.To evaluate candidate space control, Candidate Recall@50 and Candidate AvgRank respectively measure the true disease's presence and average position within the retrieved top-50 pool, verifying whether reduction strategies retain the correct diagnosis while alleviating the ranking burden.For interactive systems, consultation and inference overheads are quantified via AvgTurns, Tokens per Case, and RAG Calls.

Furthermore, blackboard quality is systematically assessed across three dimensions: (1) Information Completeness, the proportion of clinical facts covered by the blackboard state; (2) Positive Claim Accuracy, the factual support for documented findings, past diseases, and exposures; and (3) Negative Evidence Accuracy, the valid representation of explicitly denied facts.

4.4 Implementation Details

All COMPASS variants use Qwen3-8B as the backbone and constrain the final output to a Top-10 diagnosis list. Interactive consultations are capped at six turns; after a pre-RAG Top-10 ranking is produced, the Actor retrieves knowledge for one to three uncertain candidates, and the Critic performs at most two review rounds. This pool is not treated as a closed answer space: if the Critic detects insufficient candidate coverage, Retrieve expands the pool from the full normalized diagnosis option set using the current blackboard state. Detailed training prompts, hyperparameters, and ablations are provided in the Appendix C.

5 Results and Analysis

5.1 Main Interactive DDx Performance

Table 1 and Table 2 evaluate whether COMPASS can turn Qwen3-8B into an effective interactive DDx system. The results show that Qwen3-8B already possesses non-trivial diagnostic knowledge under complete case inputs, but this capability is difficult to elicit reliably without structural support: open-ended generation and large candidate spaces introduce disease-name drift, ranking noise, and unstable prioritization. COMPASS addresses this limitation by decomposing diagnosis into evidence ac-

System	Backbone	Avg			Avg Turns	Tokens / Case
		GTPA@1	GTPA@5	Avg Rank		
FullProfile + All Candidates	Qwen3-8B	0.337	0.663	4.757	5.83	3.9k
FullProfile + Candidate Control	Qwen3-8B	0.460	0.740	4.653	5.82	3.1k
MEDDxAgent	Llama3.1-8B	0.257	0.317	6.910	15.0	152.2K
MEDDxAgent	Llama3.1-70B	0.527	0.647	4.303	15.0	127.1K
MEDDxAgent	GPT-4o	**0.633**	**0.830**	3.393	15.0	125.8K
COMPASS-Prompt	Qwen3-8B	0.577	0.807	3.530	5.80	42.4K
COMPASS-GRPO	Qwen3-8B	0.627	**0.830**	**3.090**	**5.15**	25.3k

Table 1: Main results for interactive differential diagnosis. MEDDxAgent metrics reflect its optimal setting. Crucially, COMPASS relies solely on the dynamic Blackboard Text for candidate control—avoiding hidden full profiles—to simulate real-world interactive diagnosis.

System	Backbone	DDxPlus			iCraft-MD			RareBench		
		GTPA@1	GTPA@5	Avg Rank	GTPA@1	GTPA@5	Avg Rank	GTPA@1	GTPA@5	Avg Rank
FullProfile + All Candidates	Qwen3-8B	0.46	0.87	2.91	0.32	0.54	5.81	0.23	0.58	5.55
FullProfile + Candidate Control	Qwen3-8B	0.61	0.89	2.68	0.38	0.56	6.25	0.39	0.77	4.91
MEDDxAgent	Llama3.1-8B	0.58	0.60	3.10	0.12	0.17	9.07	0.07	0.18	8.56
MEDDxAgent	Llama3.1-70B	0.68	0.69	2.30	0.42	0.44	6.31	0.48	0.81	4.30
MEDDxAgent	GPT-4o	**0.86**	**0.95**	**1.29**	**0.54**	**0.72**	**4.80**	0.50	0.82	4.09
COMPASS-Prompt	Qwen3-8B	0.76	0.93	1.56	0.43	0.62	5.01	0.54	0.87	4.02
COMPASS-GRPO	Qwen3-8B	0.82	0.95	1.37	0.45	0.63	4.96	**0.61**	**0.91**	**2.94**

Table 2: Dataset-specific performance for interactive differential diagnosis. Details the ranking accuracy of baseline models and COMPASS across different disease cohorts.

quisition, blackboard-based state compression, targeted retrieval, and critic-guided revision, thereby converting latent diagnostic knowledge into a more controllable interactive process. This design yields consistent gains over monolithic Qwen3-8B baselines and surpasses the 70B MEDDxAgent baseline while approaching GPT-4o-level performance with substantially fewer consultation turns and lower token usage per case.

Dataset-level results further suggest that COMPASS is especially effective when diagnostic reasoning depends on accumulating and organizing dispersed clinical evidence, as in DDxPlus and RareBench. The relatively smaller gains on iCraft-MD highlight a remaining limitation of text-only blackboard representations: fine-grained dermatological distinctions often depend on subtle visual or morphological cues that are difficult to preserve through purely textual compression. Overall, these findings indicate that the main benefit of COMPASS does not come from simply adding more inference steps, but from imposing a structured interaction protocol that reduces memory burden, constrains candidate reasoning, and mitigates premature diagnostic commitment in small-scale diagnostic agents.

5.2 Backbone Diagnostic Capacity and Candidate-Space Control

To examine the backbone's diagnostic capacity and the role of candidate-space control, we evaluate Qwen3-8B in a static full-profile setting. This setting probes the model's upper-bound capability when complete evidence is available, rather than simulating interactive deployment. The results in Table 3 show that Qwen3-8B contains useful medical knowledge, but open-ended diagnosis and large candidate spaces make this knowledge difficult to elicit reliably. Without constraints, the model is vulnerable to disease-name drift, synonym variation, and unstable ranking among many similar diseases. Candidate-space control reduces this burden by routing the model's attention to a high-recall working set of clinically plausible diseases, rather than requiring the small model to rank the full disease ontology in a single pass. Importantly, this working set is dynamically constructed from the blackboard and can be expanded when Critic review indicates insufficient coverage.

The comparison among construction strategies shows that candidate reduction must be carefully designed. Case-driven retrieval captures benchmark-specific symptom patterns but may fail

Setting	DDxPlus		iCraft-MD		RareBench		Macro	
	GTPA@1	GTPA@5	GTPA@1	GTPA@5	GTPA@1	GTPA@5	GTPA@1	GTPA@5
NoOptions	0.29	0.46	0.27	0.34	0.28	0.57	0.280	0.457
AllOptions	0.46	0.87	0.32	0.54	0.23	0.58	0.337	0.663
CSC1	**0.70**	**0.92**	0.11	0.16	**0.47**	**0.83**	0.427	0.636
CSC2	0.54	0.86	**0.43**	**0.62**	0.25	0.68	0.407	0.72
CSC3	0.61	0.89	0.38	0.56	0.39	0.77	**0.460**	**0.74**

Table 3: Full-case diagnostic performance across different candidate spaces. Compares zero-option (NoOptions) and full-list (AllOptions) baselines against three retrieval strategies: CSC1 (case-driven TF-IDF retrieval), CSC2 (KB-driven dense semantic retrieval via PubMed), and CSC3 (a dual-source fusion of both methods).

in domains requiring fine-grained specialty knowledge, whereas knowledge-driven retrieval offers broader semantic grounding but weaker cohort alignment. The hybrid strategy combines these complementary signals, yielding more robust candidate boundaries across heterogeneous datasets. COMPASS performs this control using dynamically updated blackboard states rather than hidden full profiles, preserving the incomplete-information nature of interactive DDx. These findings suggest that candidate-space control is a key mechanism for transforming the latent diagnostic knowledge of small models into stable, controllable rankings.

5.3 Why COMPASS Works

We further analyze why COMPASS works by examining candidate recall, blackboard quality, retrieval, Critic review, and role-specific training.

Mechanism	Comparison	Key Result	Interpretation
Blackboard retrieval query	Full Profile vs. Real Blackboard, CSC1	R@50 $0.7522 \rightarrow 0.7373$	The compressed blackboard can serve as a compact retrieval query, retaining most disease-level evidence from the full profile.
	Full Profile vs. Real Blackboard, CSC2	R@50 $0.8716 \rightarrow 0.8597$	
	Full Profile vs. Real Blackboard, CSC3	R@50 $0.8985 \rightarrow 0.8588$	
Compressor-SFT	Prompt vs. SFT Compressor	Completeness $0.7468 \rightarrow 0.8512$ Negative Acc $0.4121 \rightarrow 0.6844$	SFT improves both factual coverage and evidence boundary control, especially by separating unknown information from explicit negative findings.
RAG	Pre vs. Post-RAG COMPASS-Prompt	GTPA@1 $0.513 \rightarrow 0.535$	Disease knowledge retrieval provides consistent Pre-to-Post diagnostic gains under both prompt-only and fine-tuned settings.
	Pre vs. Post-RAG COMPASS-GRPO	GTPA@1 $0.568 \rightarrow 0.586$	
Critic	Post-RAG vs. FinalCritic COMPASS-Prompt	GTPA@1 $0.535 \rightarrow 0.577$	Critic review further improves the final ranking by correcting premature or incomplete diagnoses after RAG.
	Post-RAG vs. FinalCritic COMPASS-GRPO	GTPA@1 $0.586 \rightarrow 0.627$	

Table 4: Component-level analysis of blackboard retrieval, Compressor SFT, RAG, and Critic review.

As summarized in Table 4, the dynamic blackboard functions not merely as a shorter dialogue history, but as a structured clinical memory that preserves diagnosis-relevant semantics while reducing context burden. Its effectiveness as a retrieval query indicates that most evidence needed for candidate construction can be retained in compressed form, enabling COMPASS to avoid relying on hidden full profiles during interaction. Role-specific SFT further strengthens this mechanism by turning the Compressor from a prompt-driven extractor into a calibrated state updater, especially in distinguishing confirmed findings, explicitly denied evidence, and information that remains unknown. This distinction is crucial for DDx, since confusing unknown evidence with negative evidence can prematurely eliminate valid hypotheses. RAG and Critic review provide complementary safeguards at the decision stage: retrieval supplements the small backbone with disease-specific knowledge, while the Critic uses the accumulated evidence to challenge premature or weakly supported rankings. The fact that these gains persist under both prompt-only and trained settings suggests that COMPASS's improvement comes from the interaction between structured memory, controlled knowledge access, and adversarial evidence review, rather than from any single module alone.

6 Conclusion

We propose COMPASS, a lightweight interactive DDx framework that enables 8B-scale language models to conduct multi-turn diagnostic reasoning under incomplete information, supporting local deployment in privacy-sensitive and GPU-constrained clinical settings. By integrating dynamic blackboard compression, retrieval-based candidate-space control, critic-guided evidence review, and Actor-level GRPO optimization, COMPASS addresses key limitations of small diagnostic agents, including context bloat, large candidate spaces, and premature diagnostic closure. Experiments on DDxPlus, iCraft-MD, and RareBench show that COMPASS substantially improves Qwen3-8B, outperforms a 70B interactive DDx baseline, and approaches GPT-4o performance with fewer interaction turns.

Limitations

COMPASS has several limitations. First, although role-specific SFT improves format control, state

writing, and interaction behavior, the medical knowledge of an 8B backbone remains limited. The Critic can mitigate premature closure, but it cannot always recover diagnoses that require knowledge absent from the model or retrieved documents. Stronger medical pretraining or knowledge-oriented fine-tuning may further improve performance. Second, COMPASS currently relies only on textual clinical evidence and cannot directly process multimodal inputs, such as MRI/CT scans, dermatological images, pathological slides, or longitudinal laboratory signals. This limits its applicability in specialties where visual or temporal evidence is essential. Third, while the dynamic blackboard reduces context bloat, it may compress away subtle clinical details, including fine-grained symptoms, uncertainty expressions, or negative-evidence boundaries. Final, candidate-space control also introduces a recall-dependent boundary: if the true disease is not retrieved into the working set, downstream ranking becomes difficult. COMPASS mitigates this issue by reporting Candidate Recall@50 and allowing Critic-triggered candidate expansion, but improving high-recall candidate construction remains an important direction for future work.

References

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

Judith L. Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

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.

Kevin H. Guo, Chao Yan, Avinash Baidya, Katherine Brown, Xiang Gao, Juming Xiong, Zhijun Yin, and Bradley A. Malin. 2026. Stop listening to me! how multi-turn conversations can degrade llm diagnostic reasoning. *Preprint*, arXiv:2603.11394.

Hsin-Ling Hsu, Zizheng Wang, Donghua Zhang, Nai-Chia Chen, Jerry Wang, Jun-En Ding, Chia-Hsuan Hsu, Guoan Wang, Feng Liu, Fang-Ming Hung, Chenwei Wu, and Liyue Shen. 2026. Medaction: Towards active multi-turn clinical diagnostic llms. *Preprint*, arXiv:2605.07305.

Benjamin Liu, Dillon Mehta, Rishi Malhotra, Adam Zobian, Yong Ying Tan, Samir Chopra, Daniella Rand, Natalie Pang, Abhiram Gudimella, Raghav Thallapragada, Derek Jiu, and Kevin Zhu. 2025. Examining the vulnerability of multi-agent medical systems to human interventions for clinical reasoning. In *NeurIPS 2025 Workshop on Regulatable ML*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Jacqueline Rhoads and Julie C Penick. 2017. *Formulating a differential diagnosis for the advanced practice provider*. Springer Publishing Company.

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, Vienna, Austria. Association for Computational Linguistics.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Thio, Matthew Lewis, Spiros Denaxas, and Richard J. B. Dobson. 2026. Unlocking electronic health records: a hybrid graph rag approach to safe clinical ai for patient qa. *Frontiers in Digital Health*, Volume 8 - 2026.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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. *Preprint*, arXiv:2311.06025.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Shubham Vatsal, Harsh Dubey, and Aditi Singh. 2026. Agentic ai in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of llm-based agents. *IEEE Access*, 14:4840–4863.

Qipeng Wang, Rui Sheng, Yafei Li, Huamin Qu, Yushi Sun, and Min Zhu. 2026. Medkgi: Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring. *Preprint*, arXiv:2512.24181.

Yang Zhou, Zhenting Sheng, Mingrui Tan, Yuting Song, Jun Zhou, Yu Heng Kwan, Lian Leng Low, Yang Bai, and Yong Liu. 2026. Note2chat: Improving llms for multi-turn clinical history taking using medical notes. *Preprint*, arXiv:2601.21551.

A Additional Ablation Results

This appendix provides detailed results for the component analyses discussed in Section 5.3. We report dataset-level results for three supporting experiments: retrieval robustness under compressed blackboard states, blackboard quality before and after Compressor SFT, and stage-wise gains from RAG and Critic review. These experiments are intended to complement the main results by isolating how each module contributes to the final COMPASS system.

A.1 Robustness of Blackboard Retrieval Queries

Table 5 compares candidate retrieval using full patient profiles and real blackboard states. The full-profile setting uses the complete patient case as the retrieval query, whereas the real-blackboard setting uses only the compressed blackboard state produced during interactive inquiry. This setting directly evaluates whether the blackboard can serve as a compact but reliable query for candidate-space control.

Across 300 benchmark cases, real-blackboard retrieval preserves most of the candidate recall achieved by full-profile retrieval. CSC2 shows the smallest overall degradation, dropping only from 87.16% to 85.97% R@50, suggesting that the blackboard retains sufficient semantic evidence for disease-knowledge retrieval. CSC3 achieves the highest full-profile recall, but it is more sensitive to compression, mainly because iCraft-MD requires fine-grained dermatological morphology that is harder to preserve in a compact textual state. These results support the blackboard as an effective retrieval query while also identifying fine-grained visual diagnosis as a remaining challenge.

A.2 Effect of Compressor SFT on Blackboard Quality

We further evaluate whether role-specific SFT improves the Compressor as a structured clinical memory writer. Tables 6 and 7 report blackboard quality before and after Compressor SFT. We measure information completeness against the gold patient profile, positive claim accuracy for confirmed clinical facts, and negative evidence accuracy for ruled-out findings.

Compressor SFT substantially improves both coverage and factual reliability. Overall information completeness increases from 74.68% to

Method	Dataset	n	R@50 (%)			AvgRank		
			Full Profile	Blackboard	ΔR@50	Full	Blackboard	ΔRank
B2a Disease Prototype	DDxPlus	100	100.00	100.00	+0.00	1.02	8.88	+7.86
	iCraft-MD	100	18.00	15.00	−3.00	46.82	47.44	+0.62
	RareBench	100	99.00	98.00	−1.00	2.85	8.53	+5.67
	ALL	300	72.33	71.00	−1.33	15.43	20.25	+4.82
B2b Disease Knowledge	DDxPlus	100	100.00	100.00	+0.00	6.94	9.29	+2.35
	iCraft-MD	100	72.00	65.00	−7.00	20.64	26.21	+5.57
	RareBench	100	89.00	91.00	+2.00	14.14	14.77	+0.63
	ALL	300	87.00	85.33	−1.67	13.93	16.55	+2.62
B2c Prototype–Knowledge Fusion	DDxPlus	100	100.00	100.00	+0.00	1.38	6.21	+4.83
	iCraft-MD	100	67.00	50.00	−17.00	26.32	30.55	+4.23
	RareBench	100	99.00	97.00	−2.00	5.28	7.87	+2.59
	ALL	300	88.67	82.33	−6.33	10.40	14.15	+3.75

Table 5: Candidate recall comparison between full patient profiles and real blackboard states. R@50 denotes ground-truth candidate recall within the top-50 retrieved diagnosis options. AvgRank denotes the average rank of the ground-truth diagnosis in the retrieved candidate list; lower is better. The real blackboard results are reported from two identical repeated runs, yielding the same mean value.

Dataset	Info. Completeness (%)	Positive Claim Acc. (%)	Negative Evidence Acc. (%)	Avg. Positive Claims	Avg. Negative Claims	Avg. turns
DDxPlus	61.63	95.89	41.00	5.27	2.11	5.58
iCraft-MD	73.77	97.19	62.07	3.74	3.55	5.83
RareBench	85.03	93.16	25.93	4.43	1.97	5.95
ALL	74.68	95.18	41.21	4.47	2.48	5.80

Table 6: Blackboard quality before Compressor SFT. Information Completeness measures the percentage of gold patient-profile facts covered by the final blackboard. Positive Claim Accuracy measures whether confirmed findings, history, and exposure claims are supported by the gold profile. Negative Evidence Accuracy measures whether ruled-out findings are explicitly supported rather than inferred from missing or uncertain information.

Dataset	Information Completeness (%)	Positive Claim Accuracy (%)	Negative Evidence Accuracy (%)	Avg. Positive Claims	Avg. Negative Claims	Avg. Turns
DDxPlus	77.50	96.30	66.00	6.20	2.10	5.60
iCraft-MD	84.80	97.80	74.00	4.35	3.25	5.85
RareBench	91.00	94.70	66.13	5.05	1.95	5.97
ALL	85.12	96.10	68.44	5.18	2.38	5.82

Table 7: Blackboard quality after Compressor SFT. Metrics follow Table 6. The comparison isolates the effect of role-specific SFT on the Compressor while keeping the interaction setting unchanged.

85.12%, while positive claim accuracy remains high. The largest improvement appears in negative evidence accuracy, which rises from 41.21% to 68.44%. This suggests that SFT helps the Compressor distinguish explicit negation from uncertainty or unmentioned information, reducing over-inference in ruled-out findings.

A.3 Stage-wise Effects of RAG and Critic Review

Table 8 reports the stage-wise performance of COMPASS before retrieval, after RAG, and after final Critic-guided revision. This analysis separates the contribution of retrieval-based knowledge augmentation from the final adversarial review stage.

Both COMPASS-Prompt and COMPASS-GRPO benefit from the two-stage refinement process. RAG provides a moderate but consistent improvement by supplying external evidence for uncertain candidate diseases. Critic-guided revision further improves final diagnostic ranking, especially for COMPASS-GRPO, where Avg GTPA@1 increases from 0.586 after RAG to 0.627 after final review. These results indicate that retrieval and review are complementary: RAG expands disease-specific evidence, while Critic review helps reduce premature diagnostic closure before the final Top-10 output.

B Dataset Details and Normalization

This appendix provides additional details about the benchmark datasets used in our experiments. We follow the cross-domain DDx evaluation setting introduced by MEDDxAgent and evaluate COMPASS on three complementary datasets: DDxPlus, iCraft-MD, and RareBench. These datasets cover common diseases, fine-grained specialist diagnoses, and long-tail rare diseases, respectively.

B.1 Dataset Statistics

Table 9 reports the dataset-level statistics used in our evaluation. For each dataset, we report the clin-

Stage	System	Overall				DDxPlus			iCraft-MD			RareBench		
		@1	@5	Rank	Turns	@1	@5	Rank	@1	@5	Rank	@1	@5	Rank
Pre-RAG	COMPASS-Prompt	0.513	0.773	3.840	5.700	0.700	0.910	1.860	0.370	0.570	5.490	0.470	0.840	4.170
	COMPASS-GRPO	0.568	0.800	3.450	5.050	0.790	0.930	1.620	0.410	0.610	5.180	0.500	0.860	3.550
Post-RAG	COMPASS-Prompt	0.535	0.787	3.690	5.700	0.720	0.920	1.720	0.390	0.590	5.330	0.500	0.850	4.020
	COMPASS-GRPO	0.586	0.813	3.310	5.050	0.800	0.940	1.520	0.430	0.620	5.050	0.530	0.880	3.360
Final-Critic	COMPASS-Prompt	0.577	0.807	3.530	5.800	0.760	0.930	1.560	0.430	0.620	5.010	0.540	0.870	4.020
	COMPASS-GRPO	**0.627**	**0.830**	**3.090**	5.150	**0.820**	**0.950**	**1.370**	**0.450**	**0.630**	**4.960**	**0.610**	**0.910**	**2.940**

Table 8: Stage-wise diagnostic performance of COMPASS-Prompt and COMPASS-GRPO. GTPA@1 and GTPA@5 measure whether the ground-truth diagnosis appears within the top-1 and top-5 predictions, respectively. Avg Rank denotes the average rank of the ground-truth diagnosis, where lower is better. Bold values indicate the best diagnostic performance in each metric column.

ical domain, number of evaluated cases, number of normalized diagnosis options, initial information available in the interactive setting, and the primary diagnostic challenge.

Dataset	Domain	Cases Used	Diagnosis Options	Input in Interactive Setting	Main Challenge
DDxPlus	Respiratory DDx	100	49	Age, sex, chief complaint	Common respiratory differential diagnosis
iCraft-MD	Dermatology DDx	100	387	Initial case description	Fine-grained dermatology and disease-name drift
RareBench	Rare disease DDx	100	102	Partial phenotype set	Long-tail phenotype-to-disease matching
All	Mixed	300	–	–	Cross-domain robustness

Table 9: Datasets and evaluation settings for the interactive medical diagnosis benchmark. This table outlines the domain, case count, diagnosis options, initial inputs, and main clinical challenges across three individual datasets (DDxPlus, iCraft-MD, RareBench) and a mixed evaluation set.

B.2 Dataset Characteristics

DDxPlus mainly focuses on common respiratory differential diagnosis. Since its diagnosis space is relatively small, it is useful for evaluating whether a system can perform basic symptom combination reasoning and avoid premature diagnostic closure in common disease scenarios.

iCraft-MD contains fine-grained dermatological diagnoses with a substantially larger candidate space. Many candidate diseases have similar surface names or depend on subtle morphological descriptions. This dataset is therefore used to evaluate whether candidate-space control can reduce disease-name drift and whether the blackboard can preserve morphology-related evidence during multi-turn interaction.

RareBench focuses on long-tail rare disease diagnosis. Its cases are represented by partial phenotype descriptions, making the task sensitive to phenotype-to-disease matching and external knowledge retrieval. This dataset is used to evaluate whether COMPASS can remain robust under sparse and long-tail diagnostic evidence.

C Training Data and Optimization Details

COMPASS-GRPO uses role-specific tuning only for the Actor and Compressor. The Critic is not fine-tuned in our current implementation; instead, it is used as a frozen prompt-based reviewer during inference and GRPO rollouts. The Patient Simulator, RAG module, Compressor, and Critic are also kept frozen during GRPO, so that policy updates are attributed only to the Actor's decisions on when to inquire, retrieve, or diagnose.

For Actor SFT, we construct context-decision pairs from benchmark-derived consultation trajectories. Each input contains the current dynamic blackboard state, candidate diseases, retrieved knowledge when available, the latest Critic advice, and stage-control information. The target output is a standardized JSON action in {INQUIRE, RETRIEVE, DIAGNOSE}, together with the corresponding question, retrieval target, or diagnosis list. This stage teaches the 8B model basic consultation patterns, action formatting, and execution rules, such as asking targeted follow-up questions under insufficient evidence, retrieving knowledge for uncertain or rare candidates, and revising decisions after Critic rejection. In total, the Acotor dataset contains 1,500 examples.

For Compressor SFT, we build paired examples of dialogue snippets and structured blackboard updates. The training data combines three sources: real-flow teacher-labeled examples, rule-based synthetic examples, and hard-template examples. The real-flow examples are generated from benchmark-excluded training cases across DDx-

Plus, iCraft-MD, and RareBench: the local Actor and Patient Simulator produce consultation turns, while a teacher model annotates the corresponding blackboard JSON updates. We further apply validation rules to filter noisy labels, especially cases where unknown information is incorrectly written as negative evidence or where unsupported positive findings are added from uncertain patient answers. In total, the Compressor dataset contains 1,600 examples, including 1,000 rule-based synthetic examples and 600 hard-template examples, with 10% held out for validation.

We fine-tune the Actor and Compressor from Qwen3-8B using LLaMA-Factory with LoRA. Unless otherwise specified, both modules use the same SFT recipe: LoRA is applied to all target modules with rank 32; the per-device batch size is 1 with 8 gradient accumulation steps; the learning rate is 5×10^{-7} with a cosine scheduler; training runs for 20 epochs using bf16 precision; the cutoff length is 2048; checkpoints are saved every 100 steps and logs are recorded every 10 steps. These settings are chosen to improve instruction stability and JSON-format compliance while avoiding excessive deviation from the base medical and reasoning capabilities of Qwen3-8B.

After SFT, we apply GRPO only to the Actor. For each initial blackboard state, the Actor samples a group of complete consultation trajectories, while the Patient Simulator, Compressor, RAG module, and Critic define the frozen environment. The trajectory reward follows and combines diagnostic ranking quality, interaction efficiency, and behavioral validity. In our implementation, we use a group size of $K = 4$, clipping coefficient $\epsilon = 0.2$, KL coefficient $\beta = 0.01$, rollout temperature 0.7, top-$p = 0.95$, maximum prompt length 2048, and maximum response length 1024. The learning rate is set to 1×10^{-6}, and the SFT Actor is used as the reference policy for KL regularization. Each consultation is capped at six turns; after a preliminary Top-10 diagnosis is produced, the Actor retrieves knowledge for one to three uncertain candidates, and the Critic performs at most two review rounds before the final diagnosis is required.

D Fine-grained Error Analysis

Although COMPASS achieves strong overall performance, its remaining errors are not randomly distributed. We find that most failures can be traced to a small set of interpretable bottlenecks: candidate recall limits, information loss in compressed blackboard states, biased retrieval target selection, and unreliable negative evidence boundaries. This appendix provides a compact taxonomy of these failure modes and representative case trajectories.

D.1 Error Taxonomy

Candidate Space Control (CSC) creates the most direct failure ceiling: Under a fixed Top-50 working set, candidate-space control creates a recall-dependent failure mode: if the ground-truth diagnosis is not retrieved, downstream ranking becomes difficult. This motivates the use of high-recall candidate construction and Critic-triggered expansion, especially in fine-grained specialty settings. This issue is most visible in fine-grained specialty settings, where disease names and clinical descriptions may not share stable lexical or semantic surface forms. In addition, blackboard compression can remove details that are crucial for diagnosis, such as lesion color, border, distribution, pathology, or treatment history. This explains why iCraft-MD remains more challenging than DDxPlus despite candidate-space control.

Retrieval introduces a different failure mode. RAG is helpful when the Actor selects uncertain but relevant diseases, but it can amplify early ranking errors when the selected targets are already biased toward an incorrect differential. Finally, negative evidence remains a clinically important boundary problem: prompt-only compression sometimes treats unknown or unmentioned findings as explicit absence. Compressor SFT reduces this behavior, but does not eliminate it.

D.2 Remaining Bottlenecks

Overall, the remaining failures suggest that COMPASS has moved the main bottleneck away from the backbone model's raw diagnostic capacity toward multi-stage decision reliability. Once Qwen3-8B is supported with candidate-space control, structured memory, retrieval, and review, most residual errors arise from whether the correct disease is preserved in the candidate pool, whether the blackboard retains fine-grained evidence, and whether retrieval targets are anchored to the right uncertainty.

These findings also point to direct future improvements. Candidate recall can be improved with more contrastive disease representations, especially for dermatology and rare diseases. Blackboard compression should preserve specialty-specific at-

Error Type	Description	Main Cause	Typical Effect
Candidate recall limit	The ground-truth disease is absent from the Top-50 candidate pool.	Candidate Space Control imposes a hard recall ceiling.	The Actor cannot recover the correct disease downstream.
Fine-grained compression loss	Specialty-specific details are compressed into coarse findings.	Blackboard state loses morphology, temporal, or pathology-level cues.	Retrieval and final ranking degrade, especially in iCraft-MD.
RAG target selection bias	The Actor retrieves knowledge for plausible but incorrect diseases.	Pre-RAG ranking anchors retrieval toward wrong hypotheses.	RAG reinforces an incorrect differential and may lower the ground-truth rank.
Negative evidence boundary error	Unknown or unmentioned information is written as ruled-out evidence.	Compressor confuses uncertainty with explicit negation.	The Actor may prematurely discard valid diagnostic clues.

Table 10: Taxonomy of residual COMPASS errors. The table summarizes the major failure modes observed in case-level trajectories. We focus on errors that affect the final Top-10 diagnosis list, including ground-truth misses, post-RAG ranking drops, and unsupported blackboard claims.

tributes rather than only general symptoms. RAG should retrieve contrastive evidence between competing candidates, rather than independent descriptions of selected diseases. Finally, negative evidence should be treated as a separate clinical state, with stricter distinctions among confirmed absence, unknown information, and unasked findings.

E Agent Prompt Templates

This appendix presents the core prompt templates used by the four prompted agents in COMPASS. Figure 4 shows the Actor prompt, which defines the stage-aware action space for inquiry, retrieval, and diagnosis. Figure 5 shows the Patient Simulator prompt, which constrains patient responses to the hidden gold profile and prevents unsupported symptom generation. Figure 6 shows the Compressor prompt, which incrementally updates the structured blackboard while distinguishing confirmed findings, ruled-out findings, and unknown information. Figure 7 shows the Critic prompt, which reviews the final Top-10 diagnosis list using only visible evidence, candidate options, RAG context, and the compressed blackboard.

Actor Agent Prompt

You are the Actor Agent in an interactive differential diagnosis system.

You can only observe:

- the initial patient information,
- the compressed blackboard state,
- local RAG results,
- Critic feedback,
- candidate_diseases,
- current turn number,
- diagnosis-stage control fields.

You must not assume access to the hidden full patient profile or the raw complete doctor-patient dialogue.

The blackboard is your only explicit memory:

- `clinical_belief_state.confirmed_findings`: confirmed patient facts.
- `clinical_belief_state.ruled_out_findings`: findings explicitly denied by the patient.
- `clinical_belief_state.asked_but_unknown`: topics that were asked but remain unknown.
- `medical_history.past_diseases` / `lifestyle_and_exposures`: past diseases and exposure facts.

Candidate disease rules:

1. `candidate_diseases` is the allowed diagnosis pool.
2. If `candidate_diseases` is non-empty, `final_diagnoses` must use exact disease names from this pool.
3. Do not translate, rewrite, or invent disease names.
4. Retrieve can select only 1 to 3 exact diseases from `candidate_diseases`.
5. If `candidate_diseases` is empty, you may generate disease names based on medical knowledge.

Diagnosis stages:

- `pre_rag_diagnosed`: a pre-RAG Top-10 has been produced; select 1-3 uncertain diseases for retrieval.
- `rag_done`: RAG evidence has returned, produce a post-RAG Top-10.
- `post_rag_diagnosed`: wait for Critic review.
- `critic_revision`: revise the Top-10 according to Critic feedback.
- `final_revision`: output the final Top-10 diagnosis list.

Action policy:

- If information is insufficient, ask a focused patient-facing question.
- If information is sufficient, produce a Top-10 diagnosis list.
- If `force_diagnose` is true or `max_turns` is reached, you must Diagnose.
- Do not stop because the blackboard remains unchanged.
- Do not treat `candidate_diseases` as a checklist for asking exclusion questions.
- Before diagnosing, prioritize uncovered clinical slots: chief complaint details, lesion morphology, associated symptoms, relevant negative symptoms, exposures, treatment history, test results, past diseases, and family history.
- Avoid repeating topics already answered or recorded as unknown.
- For dermatology cases, prioritize lesion morphology, distribution, course, symptoms, treatment history, tests, and exposures.
- For rare disease cases, cover developmental, neurological, metabolic, laboratory, hepatic, hematologic, age-of-onset, and family-history clues.
- For DDxPlus cases, avoid repeatedly asking about the same chief complaint; move to associated symptoms, risk factors, history, medications, and tests.

Stage-specific rules:

- `pre_rag_diagnosed`: you must Retrieve 1-3 uncertain or easily confused diseases.
- `rag_done`: you must Diagnose using the blackboard and RAG context.
- `critic_revision`: you must Diagnose according to Critic advice.
- `final_revision`: you must output the final Top-10.

Output JSON only.

Inquire:

```
{
    "action": "Inquire",
    "action_content": "a specific patient-facing question",
    "target_diseases": [],
    "final_diagnoses": []
}
```

Retrieve:

```
{
    "action": "Retrieve",
    "action_content": "brief retrieval goal",
    "target_diseases": ["exact candidate disease 1",
                        "exact candidate disease 2"],
    "final_diagnoses": []
}
```

Diagnose:

```
{
    "action": "Diagnose",
    "action_content": "",
    "target_diseases": [],
    "final_diagnoses": [
       "ranked diagnosis 1",
       "ranked diagnosis 2",
       "... up to 10 diagnoses"
    ]
}
```

Figure 4: Prompt template for the Actor agent. The Actor receives only visible COMPASS state, including the compressed blackboard, candidate diseases, RAG context, Critic feedback, and diagnosis-stage control fields. It outputs a structured JSON action among `Inquire`, `Retrieve`, and `Diagnose`.

Patient Simulator Prompt

You are simulating a patient during a medical consultation.

The following is your complete true patient profile:

<patient_profile>
{patient_profile}
</patient_profile>

When the doctor asks a question, follow these rules strictly:

1. If the asked symptom or history is explicitly present in the patient profile, answer "Yes" and provide a brief natural description.

2. Only answer "No" when the patient profile explicitly states the absence of that symptom or history.

3. If the asked information is not mentioned in the patient profile, even if it is medically common or plausible, answer:
   "I do not know", "I did not notice", or
   "The record does not mention this."

4. Never fabricate, infer, or hallucinate any symptom, history, exposure, test result, or treatment not present in the profile.

5. Keep the response natural and concise.
   Do not explain that you are reading the patient profile.
   Do not output template text such as "Doctor asked" or "According to the rules".

6. If the doctor asks about a broad symptom category, such as neurological problems, hearing problems, skin changes, or skeletal abnormalities, check whether the profile contains specific findings under that category.
   If yes, list up to five relevant findings.
   If explicitly denied, answer "No".
   If not mentioned, answer that you do not know.

7. Use semantic matching rather than exact string matching, but never add information that is absent from the profile.

8. If the doctor asks an open-ended slot question, such as symptoms, history, tests, exposures, disease course, or treatment, provide all relevant information from the profile, including positive facts and explicit negatives.

Doctor question:
{doctor_question}

Figure 5: Prompt template for the Patient Simulator. The simulator is the only module with access to the hidden full patient profile. It must answer strictly according to the profile, distinguish explicit absence from unknown information, and avoid unsupported symptom generation.

Compressor Prompt

You are a professional Medical Information Extractor.

Your task is to update the patient's clinical state JSON incrementally based on the latest doctor-patient interaction.

Core rules:

1. Preserve all previous information.
   Do not delete or omit facts already present in the old state.

2. **Classify information precisely:**
   - confirmed symptoms or signs -> confirmed_findings
   - explicitly denied symptoms or signs -> ruled_out_findings
   - asked topics that remain unknown -> asked_but_unknown
   - past medical history -> past_diseases
   - lifestyle, occupational, or environmental exposure -> lifestyle_and_exposures

3. **Do not hallucinate.**
   Never add symptoms or history that were not mentioned in the dialogue.

4. **Extract facts only from the Patient answer.**
   The Doctor question may help interpret the topic, but symptoms, diseases, tests, or exposures mentioned only in the question must not be written as patient facts.

5. **Be extremely conservative with ruled_out_findings.**
   Only record a negative finding when the patient explicitly denies a specific symptom, sign, or history.
   Generic answers such as "no other symptoms" must not be expanded into many specific negative findings.
   Answers such as "I do not know", "I did not notice", or "not mentioned in the record" must not be written into ruled_out_findings.
   If uncertainty and negation occur near the same topic, treat the topic as unknown.

6. asked_but_unknown should contain only short medical topics that were asked but remain unknown.
   Do not copy full questions.
   Do not write phrases such as "I do not know" or "not noticed".

7. If the patient only says that they do not know, and the doctor question is broad, it is acceptable to add no new field.

8. Treatment history, medication, test results, treatment response, or negative tests may be written into confirmed_findings.
   Exercise, water exposure, occupation, sexual contact, travel, sunlight exposure, and similar items should be written into lifestyle_and_exposures.
   Relevant family history may be written into past_diseases.

9. Output valid JSON only. Do not include explanations.

Output schema:

```
{
  "clinical_belief_state": {
    "confirmed_findings": [],
    "ruled_out_findings": [],
    "asked_but_unknown": []
  },
  "medical_history": {
    "past_diseases": [],
    "lifestyle_and_exposures": []
  }
}
```

Old clinical state:

{old_state_json}

Latest interaction:

{latest_interaction}

Return the updated complete JSON object.

Figure 6: Prompt template for the Compressor agent. The Compressor updates the blackboard from the latest doctor-patient interaction and writes structured fields for confirmed findings, ruled-out findings, asked-but-unknown topics, medical history, and exposures.

Critic Prompt

You are the Critic agent in an interactive differential diagnosis system.

Your job is to evaluate only the Actor's final Top-10 diagnosis list.

You do not see the hidden full patient profile. You may use:
- visible patient state,
- compressed blackboard,
- local RAG context,
- candidate_diseases,
- Actor final_diagnoses.

Rules:

1. candidate_diseases is the allowed disease pool.
   Prefer exact disease terminology.

2. Reject only for material diagnostic problems:
   - final_diagnoses is missing, too short, duplicated, or outside candidate_diseases;
   - the ranking ignores important revealed findings;
   - local RAG or the compressed blackboard clearly supports a better candidate;
   - the list is too broad or contains non-disease-like outputs.

3. Do not reject only to ask for more information when max turns have been reached
   or the case is reasonably diagnosable.

4. Keep feedback concise and actionable.
   Mention which candidates should move up or down and what evidence matters.

Return JSON only:
```
{
  "approve_diagnosis": true | false,
  "critic_latest_advice": "short advice if rejected, otherwise empty string"
}
```

Current visible COMPASS state:
{compass_state}

Evaluate the final diagnosis list and return JSON only.

Figure 7: Prompt template for the Critic agent. The Critic evaluates the final Top-10 diagnosis list using only visible evidence and returns a JSON approval decision with concise revision advice when necessary.