MedDDC-Eval: Diagnosis-Decoupled Evaluation of Multi-Turn Medical Consultation Agents

Guofeng Zhang*,†, Yizeng Quan*, Huaiyi Fang, Jianwei Lv,
Jinyao Liu, Xunxu Duan, Lening An, Yu Ouyang, Junfeng Wang

Baidu, Inc.

Abstract

Multi-turn medical consultation agents must decide what to ask, how to adapt to patient responses, and when the collected evidence is sufficient. Yet coupled evaluation conflates the value of the policy-elicited history with policy-specific terminal diagnosis generation: strong diagnosis generation can compensate for a thin history, while weaker generation can obscure a rich one. We introduce **MedDDC-Eval**, a diagnosis-decoupled evaluation testbed for multi-turn medical consultation agents. It treats the policy-elicited history as the primary comparison object and holds the history-to-diagnosis mapping constant through a shared frozen reader. A grounded interface over two held-out sources and an auditable diagnosis–trajectory–efficiency (D/T/E) harness measure downstream diagnostic usefulness, information acquisition, and efficiency. Directional semantic coverage followed by deterministic one-to-one assignment yields coherent precision–recall counts for open-ended items while limiting each prediction and reference to one credited match.

Holding histories fixed, changing only the diagnostic reader shifts diagnosis F1 by 2.2–19.0 points and reverses 18% and 36% of pairwise policy orderings on the Record and Dialogue splits. We then test whether the measurements can guide model development by applying standard Group Relative Policy Optimization (GRPO) over interactive multi-turn rollouts to post-train Qwen3-32B with diagnosis-result and trajectory feedback. On the 100-case Record and 70-case Dialogue splits, the trained policy improves over its initialization by 9.7 and 4.6 total-score points; removing either primary signal lowers held-out joint performance. This same-family comparison shows that MedDDC-Eval's measurement signals can guide evidence-acquisition policy development. MedDDC-Eval thus supports controlled attribution, interpretable elicited-history measurement, and evaluation-guided policy optimization.

Introduction

Static medical QA and diagnosis benchmarks evaluate models on evidence already supplied in a case description [Singhal et al. 2023, Wang et al. 2024, Ding et al. 2025, Bedi et al. 2025]. Multi-turn medical consultation agents face a different task: through their questions, they help determine which evidence becomes available for diagnosis [Qiao et al. 2026a, Sanghvi et al. 2026, Lai et al. 2026]. They must adapt to patient responses, screen for red flags, and judge when the collected evidence is sufficient within a bounded interaction. Evaluation must therefore ask not only whether an agent reaches a plausible diagnosis, but also whether it acquires the evidence needed to support that diagnosis.

Recent benchmarks have made the consultation process explicit through diagnostic dialogue, physician-authored criteria, consultation rubrics, and real patient question threads [Tu et al. 2025, Arora et al. 2025, Gong et al. 2026, Munnangi and Savage 2026]. Complementary work targets information acquisition, challenging patient behavior, and inquiry-oriented training [Qiao et al. 2026a, Sanghvi et al. 2026, Lai et al. 2026]. Together, these studies establish question selection and evidence acquisition as part of the evaluated capability rather than incidental properties of the dialogue. They do not, however, determine how much of an agent's diagnosis score is attributable to the evidence it acquired.

These advances leave a central attribution confound unresolved. In representative consultation systems, the same evaluated system both elicits the patient history and generates the terminal diagnosis [Tu et al. 2025, Lai et al. 2026]. An end-to-end diagnosis score is therefore determined jointly by two components: the evidence contained in the policy-elicited history and the history-to-diagnosis mapping applied to it. Strong diagnosis generation can compensate for a thin history, while weaker diagnosis generation can obscure a rich one. Under coupled evaluation, a policy-bound diagnosis score cannot isolate history usefulness and may misrank policies, obscure whether apparent gains reflect better history taking or stronger terminal generation, and misdirect benchmark conclusions and reward design.

Figure 1 isolates this confound by holding the pre-diagnosis history, diagnostic prompt, and decoding settings fixed while replacing only reader identity. The resulting case-level reference-match change and aggregate policy reordering show that reader choice materially affects the diagnosis score assigned to fixed histories. This attribution problem motivates our work.

To address this attribution confound, we introduce

*These authors contributed equally.
†Corresponding author: zhangguofeng@baidu.com

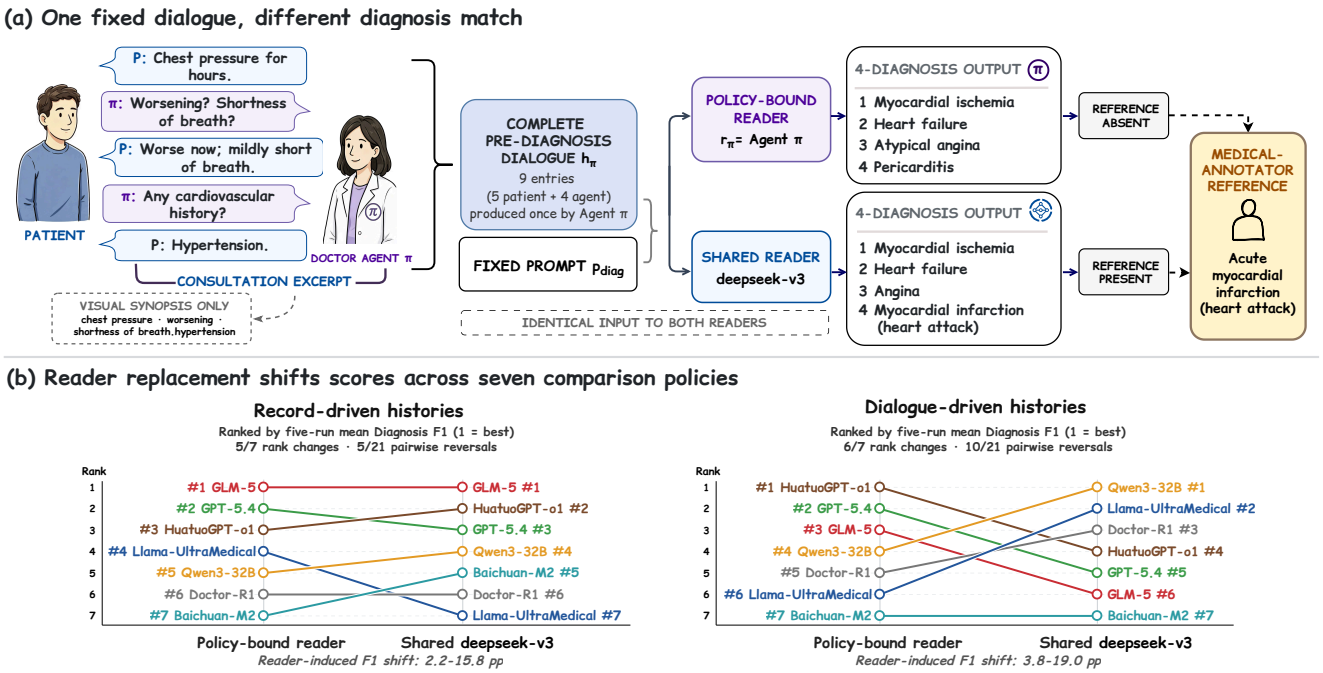

Figure 1: MedDDC-Eval fixed-history diagnostic-reader intervention at case and aggregate levels. (a) A case-derived pre-diagnosis history, diagnostic prompt, and decoding settings are held fixed while only reader identity changes, altering binary reference coverage. (b) The seven-policy motivation view deliberately excludes the trained policy and shows 5/21 Record and 10/21 Dialogue pairwise reversals under reader replacement. The shared reader provides comparison control, and scores remain conditional on reader choice; exact endpoints and the complete eight-policy audit (5/28 and 10/28 reversals) appear in Supplementary Table S7.

MedDDC-Eval, a diagnosis-decoupled evaluation testbed for multi-turn medical consultation agents. MedDDC-Eval makes the policy-elicited history the primary comparison object and maps every bounded history to a diagnosis through the same frozen DeepSeek-v3 (DS-V3) reader, prompt, and decoding interface. This shared mapping controls policy-specific diagnosis generation as a between-agent source of variation, making diagnosis differences more directly attributable to elicited histories while complementing end-to-end inquiry–diagnosis evaluation.

No single score fully characterizes an elicited history. We report diagnosis F1 for downstream diagnostic usefulness and trajectory F1 for coverage of expert-normalized information needs as related but non-substitutable axes, together with acquisition efficiency as an auxiliary measure of how early recognized evidence is obtained. Directional semantic coverage relations followed by deterministic one-to-one assignment prevent repeated semantic credit and yield coherent precision–recall counts.

We next test whether these measurements can guide model development. We apply standard Group Relative Policy Optimization (GRPO) [Shao et al. 2024] over interactive multi-turn doctor–patient rollouts, computing diagnosis-result and trajectory feedback after complete trajectories with trajectory-level credit assignment. Under the shared evaluation reader, the trained Qwen3-32B policy improves over its initialization by 9.7 and 4.6 total-score points on the two held-out sources. Reward removals trace the contribution of both feedback signals, while repeated runs and sensitivity analyses characterize the stability of the measured gains.

Together, MedDDC-Eval connects diagnosis-decoupled comparison with auditable elicited-history measurement and evaluation-guided policy development. Our contributions are:

1. **MedDDC-Eval.** We introduce a diagnosis-decoupled evaluation testbed that compares policy-elicited histories under a shared frozen diagnostic reader; a fixed-history intervention shows that policy-bound diagnosis generation changes scores and rankings.

2. **Elicited-history measurement protocol.** Diagnosis and trajectory provide related but non-substitutable views of an elicited history, while acquisition efficiency supplies an auxiliary timing measure; directional coverage and one-to-one assignment limit each prediction and reference to one credited match.

3. **Evaluation-guided policy optimization.** We instantiate the diagnosis and trajectory signals as feedback for standard GRPO over interactive multi-turn rollouts. Relative to the same Qwen3-32B initialization, the trained policy improves by 9.7 and 4.6 total-score points on the two held-out sources; reward removals demonstrate the utility of both feedback signals.

Related Work

Static Evidence and Clinical Agents

Static medical QA and diagnosis benchmarks evaluate reasoning from evidence already provided to the model [Singhal et al. 2023, Wang et al. 2024, Ding et al. 2025, Bedi et al. 2025]. Clinical-agent benchmarks extend evaluation to tool use, workflow execution, and sequential decisions [Schmidgall et al. 2024, Jiang et al. 2025, Liu et al. 2026, Lee et al. 2025, Qiao et al. 2026b]. Structured differential-diagnosis simulators such as DDXPlus further study evidence acquisition in controlled interactive settings [Tchango et al. 2022]. The distinction matters for our setting: static tasks condition on a fixed case description, whereas a multi-turn consultation agent helps determine which clinical facts enter its own diagnostic context.

Medical Dialogue and Inquiry Evaluation

Recent benchmarks make the consultation process increasingly explicit. AMIE, HealthBench, MedDialogRubrics, MedConsultBench, and ThReadMed-QA evaluate conversational diagnosis, physician-authored criteria, consultation rubrics, information acquisition, or real patient question threads [Tu et al. 2025, Arora et al. 2025, Gong et al. 2026, Qiao et al. 2026a, Munnangi and Savage 2026]. Other work studies challenging patient behaviors and targeted questioning [Li et al. 2026b, Sanghvi et al. 2026], while Doctor-R1 trains a clinical inquiry agent with both process and terminal rewards [Lai et al. 2026]. These studies show that the questions an agent asks are part of the capability being evaluated, rather than merely a path to a final answer.

Our work addresses a remaining attribution problem specific to multi-turn consultation agents. When the evaluated agent both elicits the dialogue evidence and produces the terminal diagnosis, diagnosis scores jointly reflect history usefulness and agent-specific diagnosis generation. More generally, consultation systems can differ in both the elicited history and how the diagnosis is generated. Process metrics reveal aspects of inquiry quality, but do not by themselves control this second source of variation. We therefore introduce MedDDC-Eval to compare agent-elicited histories under a shared frozen reader. This protocol adds controlled evidence-acquisition comparison alongside evaluation of complete inquiry–diagnosis pipelines.

Multi-Turn RL for LLM Agents

Group Relative Policy Optimization (GRPO) is a standard critic-free grouped-reward method introduced in DeepSeekMath [Shao et al. 2024]. Related LLM reinforcement learning (RL) work studies reproducibility, prompt and length effects, scalable agent training, and finer-grained credit assignment [Yu et al. 2025, Liu et al. 2025, Zhang et al. 2025, Wei et al. 2025, Feng et al. 2025, Li et al. 2026a, Djuhera et al. 2026]; medical inquiry work also combines process and terminal rewards [Lai et al. 2026]. Our training study uses standard GRPO for evaluation-guided reward design under a fixed diagnostic interface. Interaction and reward computation span multi-turn rollouts, while grouped normalization and credit assignment remain trajectory-level.

MedDDC-Eval: Diagnosis-Decoupled Evaluation

MedDDC-Eval comprises three tightly coupled components: a grounded multi-turn consultation interface over two held-out source strata, a shared-reader control that holds terminal diagnosis generation constant, and an auditable D/T/E harness combining semantic candidate generation with deterministic one-to-one assignment. Table 1 separates the clinical sample, frozen reference inventory, and repeated evaluation workload. Together, these components turn free-form policy-elicited histories into controlled and traceable evaluation objects.

Source Strata and Held-Out Splits

MedDDC-Eval begins from two Chinese-language source strata: expert-labeled hospital records and online multi-turn consultations. The record pipeline screens an initial pool of 3,616 records spanning 153 source department labels into 2,904 annotated candidates across 139 normalized secondary departments, from which 912 cases are curated for RL training. Each training case is instantiated under 3–4 and 5–7 turn prompt buckets, yielding 1,824 prompt rows. This companion training pool supports the controlled application study and is not part of the MedDDC-Eval held-out evaluation set. The held-out evaluation set is case-disjoint from RL training and contains the 100-case **Record split**, drawn from the same annotated candidate pool with departmental stratification, and the 70-case **Dialogue split**, constructed from online-consultation sources with frozen diagnosis-label and trajectory-target inventories. This split provides an online-consultation-derived source shift rather than hospital-certified record evidence. The supplementary document reports the complete funnel and frozen reference inventory.

Case Schema and Reference Targets

Both strata are normalized into a case $c = (s, x, F, G, Q, b)$, where s identifies the source stratum, x is the initial complaint or dialogue context, F is the de-identified fact store available to the grounded patient simulator, G is the reference diagnosis set, Q is the set of trajectory targets, and b is the turn budget. Each trajectory target $q \in Q$ is a normalized clinical information need pairing a natural-language reference question with the associated clinical condition or fact. During construction, annotators rate how strongly each drafted target bears on the diagnosis or treatment plan on a three-level importance scale, and only targets rated at the highest level enter the frozen Q; the held-out evaluator counts these frozen targets equally in TP/FP/FN. Q is not a canonical wording or a unique question sequence: related subquestions may remain bundled as one clinically natural inquiry unit, and semantically compatible questions can satisfy a target. Whether a target has already been volunteered is derived from the realized dialogue and removed from that dialogue's scoring denominator.

Component	Role	Inventory / scope	Boundary
Grounded consultation interface	Produces bounded policy-elicited histories over two source strata.	170 unique held-out cases: 100 Record and 70 Dialogue; grounded simulator; first-summary boundary.	Offline simulated-consultation setting with 170 held-out clinical cases.
Shared-reader control	Holds terminal diagnosis generation constant across policies.	One frozen DS-V3 prompt/decoding interface applied to every completed history.	Between-agent attribution control; scores remain reader-conditional.
Auditable D/T/E harness	Measures diagnostic usefulness, reviewed information acquisition, and timing.	287 diagnosis labels; 1,451 trajectory targets; semantic candidate edges plus deterministic one-to-one assignment.	D/T/E are primary axes; Total is an aggregate summary.
Repeated evaluation workload	Characterizes rollout variability and supports audit.	Main-8 $\times$ 5 runs; 6,800 prespecified case–system–run evaluation slots.	The clinical sample remains the 170 held-out cases.

Table 1: MedDDC-Eval component and inventory map. The held-out clinical sample, frozen reference inventory, and repeated evaluation workload are different units and are reported separately.

Annotation, Governance, and Release Boundary

Fifteen medical annotators construct G and Q across four batches. They annotate only targets they judge to bear on the diagnosis or treatment plan: they first draft case-grounded candidate targets, then normalize their medical meaning and rate their three-level importance, retaining only the most important targets, and finally review clinical relevance, redundancy, and semantic consistency; disagreements are adjudicated before targets are frozen. References are thus finalized through staged drafting, medical review, and adjudication. Raw records and original consultations are used only to create de-identified case objects and remain restricted. Case-level identifiers are deduplicated before split assignment, and the training and evaluation sets are identifier-disjoint. Patient-derived text, linkable case identifiers, and fine-grained records are excluded from unconditional public artifacts; regardless of source-data authorization, the planned minimum release contains the schema, reviewed prompt templates, scoring code, aggregate results, and fully synthetic examples. When permitted by source authorization, de-identification requirements, and institutional review constraints, we additionally plan to release approved subsets of the test data. The supplementary Dataset Construction and Validation section organizes the construction statistics, annotation inventory, release tiers, and validity evidence.

Patient Simulation

The simulator constructs a new interactive doctor–patient dialogue for both source strata. In record mode, the patient starts from the structured chief complaint and answers with record-grounded facts; in dialogue mode, the simulator uses an extracted queue of reference facts from the corresponding source consultation. The evaluated doctor receives only the current simulated interaction, not the retained source consultation. The doctor asks questions or issues a structured diagnosis action, and the rollout stops at the diagnosis action or the maximum turn budget; the main training setup uses at most seven consultation turns.

Holding Terminal Diagnosis Generation Constant

When the doctor policy stops, its complete dialogue history is passed to a frozen prompt-based DeepSeek-v3 diagnostic reader. Applied with the same prompt and decoding interface to every policy, this shared reader produces the diagnostic output used in the main comparison. A policy-bound reader—the rollout policy itself applied to the fixed history under the same diagnostic prompt—is used only in the fixed-history intervention. Formally, the primary object of comparison is the policy-elicited history h_π, while the history-to-diagnosis mapping $r_{\mathrm{shared}}(h_\pi)$ is held constant. Diagnosis F1 therefore tests the downstream usefulness of each history under a common reader rather than under a reader that varies with the rollout policy. Section characterizes the importance of this control through a fixed-history intervention that changes only reader identity.

Measurement Contract and One-to-One Assignment

The evaluator operates on the complete dialogue history h, shared-reader diagnostic output S, and case references (G, Q). It extracts diagnostic predictions and clinically meaningful doctor-question units and removes trajectory targets already volunteered by the patient. As Figure 2 summarizes, the LLM-assisted judge only proposes directional, protocol-admissible coverage edges; a deterministic program selects the maximum-cardinality one-to-one assignment, derives TP/FP/FN counts, and computes diagnostic usefulness D, trajectory coverage T, and acquisition efficiency E.

Table 2 states the self-contained measurement contract: D/T/E are the primary axes, whereas Total is a prespecified aggregate summary rather than a fourth clinical construct.

For either diagnosis or trajectory, let $P = \{p_i\}_{i=1}^{m}$ be the extracted units and $R = \{r_j\}_{j=1}^{n}$ the corresponding reference units. An LLM-assisted semantic judge proposes candidate edges $C \subseteq P \times R$, where (p_i, r_j) means that prediction p_i covers reference r_j under prespecified dimension and subdimension rules. This relation is directional rather than literal or bidirectional equivalence: a trajectory prediction may cover one clinically valid component of a bundled reference question, and diagnosis matching permits protocol-specified granularity relations. We then compute a deterministic maximum-cardinality bipartite matching $M \subseteq C$. Each predicted and reference unit can occur in at most one selected edge, giving

$$\mathrm{TP} = |M|, \quad \mathrm{FP} = m - |M|, \quad \mathrm{FN} = n - |M|, \qquad (1)$$

M	Question answered	Evaluation unit	Aggregation	Main boundary
D	Does the elicited history support reference diagnoses under the shared reader?	Diagnosis item	Run-level micro F1	Conditional on the frozen reader and proposed semantic edges.
T	Does the consultation elicit reviewed clinical information needs?	Doctor question–target unit	Run-level micro F1 before first summary	Bundled-target resolution; volunteered facts are filtered.
E	How early is recognized evidence acquired?	Matched question turn	Deterministic timing score	Coupled to evidence recognized by T; auxiliary axis.
Total	What is the prespecified joint ordering?	System–run	$0.5D + 0.4T + 0.1E$	Prespecified aggregate summary for system comparison.

Table 2: Self-contained D/T/E measurement contract. Semantic models propose protocol-admissible candidate edges; deterministic one-to-one assignment selects and counts them.

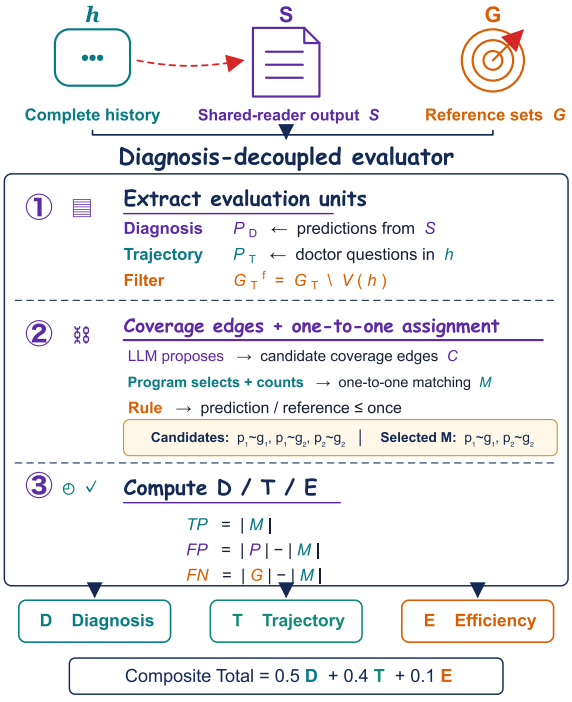

Figure 2: MedDDC-Eval diagnosis-decoupled evaluation protocol. One fixed shared reader maps each stored history h to output S; G contains diagnosis and trajectory references. An LLM proposes directional, protocol-admissible coverage edges, while a deterministic program selects the maximum-cardinality one-to-one assignment and derives TP/FP/FN counts. T/E use only doctor questions before the first summary action. $D/T/E$ are the three reported axes; the prespecified weighted Total is used only for aggregate ordering and comparisons, not as a fourth axis.

followed by standard precision, recall, and F1. Candidate generation handles open-ended paraphrases and allowed inclusion relations; the programmatic assignment only constrains counting and does not certify edge validity or recover an admissible edge the judge failed to propose. Repeated candidate edges cannot increase TP beyond the number of distinct prediction or reference nodes. The evaluator records candidate, selected, excluded, and unmapped edges for audit.

Diagnostic Usefulness

The evaluator extracts diagnostic predictions from the shared-reader output S: from its structured tool call when available, and otherwise from its free-form text. These predictions are assigned one-to-one against the reference diagnosis set. Run-level diagnosis F1 is computed from micro-aggregated TP/FP/FN counts. Under the MedDDC-Eval protocol, this score asks whether the agent-elicited history supports the reference diagnosis through a common downstream mapping.

Trajectory Coverage

The evaluator extracts doctor-question units before the first summary action and matches them to the expert-annotated reference trajectory with the same coverage-and-assignment protocol. Trajectory targets deliberately retain clinically natural bundled questions: physicians commonly ask related aspects of one information need in a single turn, whereas fully atomic decomposition would reward checklist-style splitting and change the estimand to slot filling. A prediction may therefore cover one valid component of a bundle under the prespecified inquiry-type constraints; once selected, that prediction and bundle each consume one matching capacity. If the patient voluntarily provides a fact before being asked, the corresponding trajectory target is filtered. Run-level trajectory F1 is computed from micro-aggregated counts.

Efficiency and Aggregate Total

The efficiency score rewards early collection of key information and penalizes ineffective turns. Let t_i be the turn index where the i-th key question is asked:

$$\bar{t} = \frac{\sum_i t_i \exp(-0.15(t_i - 1))}{\sum_i \exp(-0.15(t_i - 1))}. \tag{2}$$

The weighted position is converted to a timing score and multiplied by an ineffective-question penalty; the resulting efficiency lies in $[0.1, 1.0]$. The supplementary document gives the full mapping. We report the 100 and 70 held-out cases as statistical units; operation counts are a reproducibility audit, not extra clinical samples.

The three reported axes are combined into a prespecified composite Total for aggregate ordering and comparisons:

$$\text{Total} = 0.5D + 0.4T + 0.1E. \tag{3}$$

We freeze this protocol—the two held-out splits, patient simulator, shared reader, judge prompts, and D/T/E scoring with the composite weights—as **MedDDC-Eval v1.0**; all reported comparisons use this frozen version.

Evaluation-Guided Policy Optimization

Having defined MedDDC-Eval as a common interface for comparing evidence-acquisition policies, we ask whether its diagnosis-result and trajectory signals can guide policy development. We test this question by post-training Qwen3-32B on cases disjoint from held-out evaluation, using standard GRPO with feedback derived from the proposed measurement axes.

Interactive Multi-Turn Rollouts

We apply standard GRPO to interactive multi-turn doctor–patient rollouts. We refer to this training setup as multi-turn GRPO: the multi-turn structure lies in rollout generation and trajectory-level reward computation, while grouped normalization and policy optimization follow standard GRPO. Each GRPO sample is a complete interactive consultation trajectory: the doctor policy alternates with the grounded patient simulator until a structured diagnosis action or the seven-turn limit, after which trajectory-level rewards are computed and normalized within the prompt group. The supplementary material tabulates the rollout, reward, credit, and train–evaluation boundaries.

Process-Aware Reward Design

The global reward combines two primary medical signals with auxiliary behavioral constraints. The primary signals are diagnosis-result feedback from the frozen diagnostic tool and trajectory matching against the annotated consultation targets. The auxiliary constraints encourage usable rollouts by controlling turn budget, output format, repetition, question count, and prefix similarity:

$$R = R_{\text{diag}} + R_{\text{traj}} + \lambda_{\text{turn}} R_{\text{turn}} + \lambda_{\text{format}} R_{\text{format}} + \lambda_{\text{rep}} R_{\text{rep}} + R_{\text{qcount}} + \lambda_{\text{prefix}} R_{\text{prefix}}. \quad (4)$$

This training reward is distinct from the held-out total-score formula in the evaluation protocol; held-out trajectory scores are computed by the evaluation pipeline as run-level micro-aggregated matching scores before being combined with diagnosis F1 and efficiency. Reward-side terminal generation, LLM-assisted extraction, and semantic judging use DeepSeek-v3.2 during training, whereas the shared diagnostic reader and offline extractor/judge use DeepSeek-v3 during held-out evaluation. The active scalar coefficients are $\lambda_{\text{turn}} = 0.8$, $\lambda_{\text{format}} = 4.0$, $\lambda_{\text{rep}} = 0.5$, and $\lambda_{\text{prefix}} = 0.5$; the diagnosis, trajectory, and question-count components use their implementation-level scales without an additional outer multiplier. Full implementation variables are listed in the supplementary document.

GRPO Update and Credit Scope

For each prompt, GRPO compares eight complete sampled trajectories and normalizes their rewards within the group. We implement standard GRPO with the SLIME post-training framework [Zhu et al. 2025]. The normalized sequence reward is broadcast across the policy-generated response tokens, yielding trajectory-level credit assignment over multi-turn rollouts. Process awareness is encoded by the diagnosis-result, trajectory, and behavioral reward signals. Held-out evaluation keeps MedDDC-Eval v1.0 fixed and uses DeepSeek-v3; reward-side terminal generation, extraction, and judging use DeepSeek-v3.2.

Ablation Controls

The reward-component ablations measure the contribution of outcome and process supervision. Both ablations keep the rollout environment, frozen diagnostic tool, and held-out diagnosis–trajectory–efficiency scoring protocol fixed. One removes the diagnosis-result reward, leaving trajectory and auxiliary behavioral rewards; the other removes the trajectory-F1 reward, leaving diagnosis-result and auxiliary behavioral rewards.

Experiments

Experimental Comparison

The main comparison contains our GRPO-trained Qwen3-32B policy, its Qwen3-32B initialization, and six external consultation agents: GPT-5.4, GLM-5, Doctor-R1, Baichuan-M2-32B [Baichuan-M2 Team et al. 2025], Llama-3-70B-UltraMedical [Zhang et al. 2024], and HuatuoGPT-o1-70B [Chen et al. 2024]. Tables denote the trained policy as Qwen3-32B + GRPO; prose refers to it as the trained policy thereafter. All agents use MedDDC-Eval v1.0 with the same Chinese-language simulator, turn budget, shared DS-V3 reader, D/T/E evaluator, and five-run reporting procedure. The trained–base comparison is the controlled training contrast; external agents contextualize the resulting capability.

Reader Choice Changes Fixed-History Rankings

A controlled reader intervention exposes a hidden source of ranking instability in coupled consultation evaluation. Figure 1 holds each completed consultation history, diagnostic prompt, and decoding settings fixed, and changes only the reader from a policy-bound reader to shared DS-V3. The figure uses a seven-policy motivation view; Supplementary Table S7 reports the complete eight-policy audit. Reader replacement changes diagnosis F1 by 2.2–19.0 points and reverses 5/28 (18%) Record and 10/28 (36%) Dialogue pairwise orderings, while questions, patient answers, trajectory, and efficiency remain fixed. The heterogeneous shifts and reversals isolate reader identity as a material source of between-policy diagnosis-score variation and motivate a common history-to-diagnosis mapping.

Diagnosis and trajectory align at the system level yet resolve different case-level behavior. Across the eight system means, their Pearson association is positive on Record ($r = 0.569$) and Dialogue ($r = 0.303$), whereas across case–model means it is much weaker ($r = 0.147$ and 0.101). Both high-trajectory/low-diagnosis and low-trajectory/high-

Agent / policy	MedDDC-Eval Record (100 cases)				MedDDC-Eval Dialogue (70 cases)				Δ Total vs. Qwen3-32B
	D	T	E	Total	D	T	E	Total	Record / Dialogue
Controlled same-family comparison									
Qwen3-32B + GRPO	**44.6**	57.3	**83.0**	**53.5**	**41.8**	**52.6**	**82.4**	**50.2**	**+9.7 / +4.6**
Qwen3-32B	33.1	48.1	80.2	43.8	39.8	44.7	78.0	45.6	–
External reference agents									
GPT-5.4	36.3	**58.8**	79.7	49.7	35.5	52.4	79.1	46.6	+5.9 / +1.0
GLM-5	38.0	54.4	66.4	47.4	34.3	44.7	74.9	42.5	+3.6 / -3.1
HuatuoGPT-o1-70B	37.7	46.2	77.9	45.1	37.2	40.8	75.9	42.5	+1.3 / -3.1
Llama-3-70B-UltraMedical	31.4	50.5	81.2	44.0	39.5	46.9	79.4	46.4	+0.2 / +0.8
Baichuan-M2-32B	32.3	46.3	81.8	42.9	33.5	44.2	80.8	42.5	-0.9 / -3.1
Doctor-R1	31.5	40.3	75.0	39.3	37.4	41.2	78.6	43.1	-4.5 / -2.5

Table 3: Evidence-acquisition profiles under MedDDC-Eval v1.0. Values are percentage means over five fixed-checkpoint rollout repeats; D and T use run-level micro one-to-one counts and E uses the first summary round. Qwen3-32B + GRPO versus Qwen3-32B is the controlled training contrast; external agents provide capability context. Total is $0.5D + 0.4T + 0.1E$, and Δ reports display-level Total differences from Qwen3-32B for Record / Dialogue. Component profiles show how each aggregate is obtained; sample SDs appear in the supplementary material.

diagnosis cases remain common. This scale-dependent pattern supports reporting D and T separately: systems can improve both objectives on average while individual consultations still require outcome and process localization.

Evaluation-Guided Training Improves Both Measurement Axes

Evaluation-guided training improves the same Qwen3-32B policy on both held-out sources and on both primary measurement axes. Relative to its initialization, the trained policy improves Record diagnosis, trajectory, efficiency, and Total by 11.5, 9.2, 2.8, and 9.7 points; the Dialogue gains are 2.0, 7.9, 4.4, and 4.6 points. With terminal diagnosis generation held fixed, the D and T gains show that the trained policy elicits histories that are more diagnostically useful and cover more reference information needs under the reported protocol. The component composition differs across the two source strata, but the experiment does not identify the case or source properties responsible for that difference.

Similar Totals Conceal Different Capability Profiles

The aggregate ordering conceals distinct evidence-acquisition profiles. The trained policy has the highest Total in this eight-system comparison: 53.5 on Record and 50.2 on Dialogue, compared with 49.7 and 46.6 for the strongest external Total reference, GPT-5.4. Yet GPT-5.4 is 1.5 points higher in Record trajectory F1, while the trained policy is 8.3 points higher in diagnosis F1 and 3.3 points higher in efficiency; on Dialogue the trained policy is higher on all three components. Total identifies the leading joint score, while D/T/E reveals the capability composition behind it. The external comparison supplies capability context; the same-family contrast supplies the controlled training evidence.

The aggregate gains persist under paired case resampling. After averaging repeated rollouts per case, the trained–base Total difference is 10.6 points on Record (95% CI [8.4, 12.9]) and 5.2 points on Dialogue ([2.7, 7.6]). The paired differences from GPT-5.4 are 3.9 ([1.4, 6.5]) and 3.3 ([0.5, 6.2]) points. All four intervals remain above zero, supporting positive mean differences under held-out case resampling for the evaluated checkpoints.

Average Gains Decompose into Distinct Case-Level Modes

Figure 3 decomposes the average trained–base gain into distinct, partially overlapping case-level modes across 169 complete pairs. Sixty-two cases (36.7%) meet the predefined stable trajectory-improvement criterion; the bucket-center abdominal-pain case changes by $+22.5$ trajectory and -0.7 diagnosis points. Seven cases (4.1%) meet the diagnosis-gain criterion while retaining sparse trajectory coverage; the jaundice-like case changes by $+43.3$ diagnosis and -1.2 trajectory points. These modes show why the D/T decomposition is useful for development: a higher aggregate can arise from improved information acquisition in some cases and improved downstream diagnostic usefulness in others. The marker sets localize recurring patterns under the stated criteria and are non-exhaustive.

Outcome and Trajectory Feedback Make Complementary Contributions

The two feedback signals make complementary, non-metric-specific contributions in the reported training configuration. Removing diagnosis-result feedback lowers Total by 5.0 points on Record and 3.6 points on Dialogue; diagnosis and trajectory also decline even as nominal efficiency rises by 7.8 and 6.6 points. Removing trajectory feedback lowers Total by 4.1 and 4.2 points and reduces all three components on both sources. The first pattern shows that higher nominal efficiency alone does not imply better joint consultation quality, while the two removal contrasts jointly show that outcome and process feedback each contribute to the full policy's measured balance. Paired case confidence intervals for all four Total differences remain above zero. The cross-component responses are consistent with interacting supervision signals; the removals establish their contribution within this Qwen3-32B

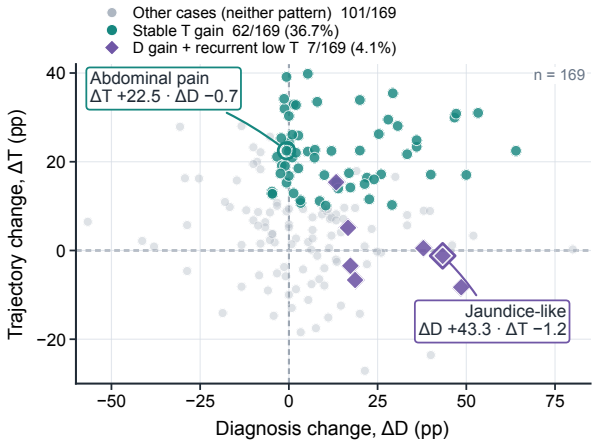

Figure 3: Case-level decomposition of trained–base changes: each point is the five-run mean $(\Delta D, \Delta T)$ for one complete case under the shared-reader one-to-one evaluator ($n = 169$). Colors identify two predefined multi-criterion patterns, rather than all cases with large mean changes; S denotes common-language superiority over 25 cross-run trained–base pairs. The predefined teal stable trajectory-gain pattern requires $\Delta T \geq 10$, $S_T \geq 0.70$, and $\Delta D \geq -5$ points (62/169). The predefined violet diagnosis-gain pattern requires $\Delta D \geq 10$, $S_D \geq 0.70$, and trained-policy membership in split-specific bottom-T/top-D tertiles in at least 3/5 runs (7/169); low T is an absolute trained-policy level, not $\Delta T < 0$. The patterns are non-exhaustive and may overlap; outlined callouts identify two bucket-center, paraphrased examples, and runs are fixed-checkpoint rollout repeats.

Setting	D	T	E	Total
Record-driven				
Full reward	**44.6**	**57.3**	**83.0**	**53.5**
w/o diagnosis reward	36.4	53.2	90.8	48.5
w/o trajectory reward	38.8	55.4	78.1	49.4
Dialogue-driven				
Full reward	**41.8**	**52.6**	**82.4**	**50.2**
w/o diagnosis reward	35.3	50.1	89.0	46.6
w/o trajectory reward	36.3	50.3	77.0	46.0

Table 4: Reward-component ablations for Qwen3-32B policies trained with standard GRPO. Values are percentage means over five fixed-checkpoint rollout repeats; the full-reward rows match Table 3. Each ablation changes only one primary training signal; the rollout environment, auxiliary constraints, held-out MedDDC-Eval v1.0 evaluator, and $0.5D + 0.4T + 0.1E$ aggregate remain fixed.

configuration, rather than a universal internal mechanism.

Meta-Evaluation Separates Stability from Human Alignment

The main ordering is stable across the tested rollout, weight, and reader perturbations, although one alternative-reader comparison remains close. Across the 16 model–split summaries, the median Total standard deviation over five rollout runs is 1.41 points. Seven local weight perturbations keep the trained policy first on both sources, with top-to-second margins of 3.25–4.34 points. Replacing DS-V3 on fixed histories with Qwen3-32B or GPT-5.4 also preserves its highest mean Total on both sources. The GPT-5.4-reader Dialogue setting narrows the margin to 0.44 points, compared with 1.92–2.56 points in the other reader–source settings, localizing reader choice as the closest tested sensitivity condition.

The audits separate reproducibility of semantic judgments from coverage and human alignment. Across 1,448 prediction–reference pair labels, two model auditors attain 0.960 raw agreement and Cohen's $\kappa = 0.873$, indicating high consistency under the audited prompts. Their candidate-edge recall is 0.430 and 0.455, however, so missed admissible edges remain a distinct measurement risk even though recall does not differ significantly across the eight systems ($p = 0.53$ and 0.59). In the separate 80-output human audit, automatic professional-quality ratings exactly match adjudicated ratings from medically qualified reviewers on 62.5% of outputs, with quadratic-weighted Cohen's $\kappa = 0.415$. Thus, the protocol shows strong cross-auditor reproducibility, incomplete candidate coverage, and moderate physician-anchored ordinal alignment under the tested conditions; full audit results appear in the supplementary material.

Limitations

MedDDC-Eval compares evidence acquisition in simulation and does not evaluate clinical deployment. Grounded patient simulators do not reproduce every patient behavior, the Dialogue split reflects online-consultation sources rather than hospital-certified evidence, and valid consultations can differ from reference trajectories in ordering and phrasing. Directional coverage edges come from LLM-assisted judges. In a stratified two-model cross-audit (supplementary material), candidate-edge recall was 0.430 and 0.455; we detected no significant recall heterogeneity across the eight evaluated systems ($p = 0.53$ and 0.59), and the trained policy did not have higher audited recall than the external agents. Nevertheless, missed admissible edges remain the main observed measurement risk because one-to-one assignment cannot recover an edge the judge fails to propose; bundled targets are also coarser than atomic facts. The frozen reader is a between-agent control, not a clinical oracle, so measured scores are conditional on the chosen reader; the RQ4 reader replacements bound but do not remove this dependence. Efficiency reflects the timing of recognized evidence acquisition rather than an independent clinical-quality axis, and the fixed $0.5/0.4/0.1$ total supports a single ordering while component scores carry the interpretation. Repeated runs resample rollouts from fixed checkpoints, not training seeds, and cross-lingual transfer is unestablished.

Ethical Considerations

This work concerns medical dialogue, a high-stakes domain. The system is for offline evaluation, not patient-facing advice; no evaluated model is suitable for autonomous clinical use, and deployment would require clinical validation, safety review, privacy protection, and regulatory compliance. Privacy and licensing constraints prevent release of original records and raw consultations. Regardless of source-data authorization, the planned minimum release contains the evaluator schema, approved prompts, matching and scoring code, aggregate results, and a fully synthetic end-to-end fixture. When permitted by source authorization, de-identification requirements, and institutional review constraints, we plan to release approved test-data subsets; upon acceptance, we will prioritize completing these clearance steps for the held-out evaluation set and runnable source-derived instances, which remain conditional on authorization and de-identification review. LLM-based evaluators may encode biases, so the protocol is not safety certification.

Conclusion

MedDDC-Eval makes the policy-elicited history the primary comparison object and holds its history-to-diagnosis mapping constant with a shared reader. The fixed-history stress test shows why that control is needed; auditable one-to-one D/T/E measurement then supplies complementary views of diagnostic usefulness, information acquisition, and efficiency. A controlled study using standard GRPO over interactive multi-turn rollouts improves Qwen3-32B on both held-out splits, and either reward removal weakens joint performance in the reported configurations. Together, these results establish a measurement-to-optimization loop in which histories are compared under a common diagnostic interface, analyzed through complementary outcome and process measurements, and improved through evaluation-guided feedback.

References

R. K. Arora, J. Wei, R. Soskin Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpourlas, M. Sharman, M. Shah, A. Vallone, A. Beutel, J. Heidecke, and K. Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. doi: 10.48550/arXiv.2505.08775. URL https://arxiv.org/abs/2505.08775.

Baichuan-M2 Team, C. Dou, C. Liu, F. Yang, F. Li, J. Jia, M. Chen, Q. Ju, S. Wang, S. Dang, T. Li, X. Zeng, Y. Zhou, C. Zhu, D. Pan, F. Deng, G. Ai, G. Dong, H. Zhang, J. Tai, J. Hong, K. Lu, L. Sun, P. Guo, Q. Ma, R. Xin, S. Yang, S. Zhang, Y. Mo, Z. Liang, Z. Zhang, H. Cui, Z. Zhu, and X. Wang. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*, 2025. doi: 10.48550/arXiv.2509.02208. URL https://arxiv.org/abs/2509.02208.

S. Bedi, H. Cui, M. Fuentes, A. Unell, M. Wornow, J. M. Banda, N. Kotecha, T. Keyes, Y. Mai, M. Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025. doi: 10.48550/arXiv.2505.23802. URL https://arxiv.org/abs/2505.23802.

J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang. Huatuogpt-o1, towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024. doi: 10.48550/arXiv.2412.18925. URL https://arxiv.org/abs/2412.18925.

J. Ding, L. Lu, C. Ding, M. Bian, J. Chen, W. Pang, R. Chen, X. Peng, R. Lu, S. Ren, G. Zhu, X. Wu, Z. Liu, R. Zhang, L. Jiang, B. Han, Y. Wang, and J. Xu. Medbench v4: A robust and scalable benchmark for evaluating chinese medical language models, multimodal models, and intelligent agents. *arXiv preprint arXiv:2511.14439*, 2025. doi: 10.48550/arXiv.2511.14439. URL https://arxiv.org/abs/2511.14439.

A. Djuhera, S. R. Kadhe, F. Ahmed, and H. Boche. Tsr: Trajectory-search rollouts for multi-turn rl of llm agents. *arXiv preprint arXiv:2602.11767*, 2026. doi: 10.48550/arXiv.2602.11767. URL https://arxiv.org/abs/2602.11767.

L. Feng, Z. Xue, T. Liu, and B. An. Group-in-group policy optimization for llm agent training. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

L. Gong, W. Fang, T. Yang, D. Tao, C. Guo, P. Wei, B. Xie, J. Guan, Z. Chen, F. Shi, J. Gu, and J. Liu. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*, 2026. doi: 10.48550/arXiv.2601.03023. URL https://arxiv.org/abs/2601.03023.

Y. Jiang, K. C. Black, G. Geng, D. Park, J. Zou, A. Y. Ng, and J. H. Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents. *arXiv preprint arXiv:2501.14654*, 2025. doi: 10.48550/arXiv.2501.14654. URL https://arxiv.org/abs/2501.14654.

Y. Lai, K. Liu, Z. Wang, W. Ma, and Y. Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. *arXiv preprint arXiv:2510.04284*, 2026. doi: 10.48550/arXiv.2510.04284. URL https://arxiv.org/abs/2510.04284.

G. Lee, E. Bach, E. Yang, T. Pollard, A. Johnson, E. Choi, Y. Jia, and J. H. Lee. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. *arXiv preprint arXiv:2509.19319*, 2025. doi: 10.48550/arXiv.2509.19319. URL https://arxiv.org/abs/2509.19319.

J. Li, P. Zhou, R. Meng, M. P. Vadera, L. Li, and Y. Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms. *arXiv preprint arXiv:2512.17008*, 2026a. doi: 10.48550/arXiv.2512.17008. URL https://arxiv.org/abs/2512.17008.

Y. Li, X. Jie, W. Ruan, X. Zhang, H. Zhu, Y. Gao, C. Du, and R. Liu. Beyond idealized patients: Evaluating llms under challenging patient behaviors in medical consultations. *arXiv preprint arXiv:2603.29373*, 2026b. doi: 10.48550/arXiv.2603.29373. URL https://arxiv.org/abs/2603.29373.

R. Liu, I. Q. Mohiuddin, A. J. Schoeffler, K. Renduchintala, A. Nayak, P. L. Vemu, S. C. Vedak, K. C. Black, J. L. Havlik, I. Ogunmola, et al. Physicianbench: Evaluating llm agents in real-world ehr environments. *arXiv preprint arXiv:2605.02240*, 2026. doi: 10.48550/arXiv.2605.02240. URL https://arxiv.org/abs/2605.02240.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. doi: 10.48550/arXiv.2503.20783. URL https://arxiv.org/abs/2503.20783.

M. Munnangi and S. Savage. Threadmed-qa: A multi-turn medical dialogue benchmark from real patient questions. *arXiv preprint arXiv:2603.11281*, 2026. doi: 10.48550/arXiv.2603.11281. URL https://arxiv.org/abs/2603.11281.

C. Qiao, J. Huang, D. Zhao, Z. Liu, Y. Shen, B. Cheng, W. Lin, and K. Wu. Medconsultbench: A full-cycle, fine-grained, process-aware benchmark for medical consultation agents. *arXiv preprint arXiv:2601.12661*, 2026a. doi: 10.48550/arXiv.2601.12661. URL https://arxiv.org/abs/2601.12661.

Y. Qiao, L. Liu, Y. Shen, J. Wang, J. Gu, Z. Chu, and K. Ren. Ehr-complex: Benchmarking medical agents for complex clinical reasoning. *arXiv preprint arXiv:2606.23301*, 2026b. doi: 10.48550/arXiv.2606.23301. URL https://arxiv.org/abs/2606.23301.

A. Sanghvi, N. Akash, R. Imam, A. Sharma, and M. Jain. Medxagent: Multi-agent consultation for interactive medical diagnosis. *arXiv preprint arXiv:2606.03416*, 2026. doi: 10.48550/arXiv.2606.03416. URL https://arxiv.org/abs/2606.03416.

S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, and M. Moor. Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024. doi: 10.48550/arXiv.2405.07960. URL https://arxiv.org/abs/2405.07960.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

A. F. Tchango, R. Goel, Z. Wen, J. Martel, and J. Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/cae73a974390c0edd95ae7aeae09139c-Abstract-Datasets_and_Benchmarks.html.

T. Tu, M. Schaekermann, A. Palepu, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642: 442–450, 2025. doi: 10.1038/s41586-025-08866-7. URL https://www.nature.com/articles/s41586-025-08866-7.

X. Wang, G. Chen, D. Song, Z. Zhang, Z. Chen, Q. Xiao, J. Chen, F. Jiang, J. Li, X. Wan, B. Wang, and H. Li. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.343. URL https://aclanthology.org/2024.naacl-long.343/.

Q. Wei, S. Zeng, C. Li, W. Brown, O. Frunza, W. Deng, A. Schneider, Y. Nevmyvaka, Y. K. Zhao, A. Garcia, and M. Hong. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *arXiv preprint arXiv:2505.11821*, 2025. doi: 10.48550/arXiv.2505.11821. URL https://arxiv.org/abs/2505.11821.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. doi: 10.48550/arXiv.2503.14476. URL https://arxiv.org/abs/2503.14476.

H. Zhang, X. Liu, B. Lv, X. Sun, B. Jing, I. L. Iong, Z. Hou, Z. Qi, H. Lai, Y. Xu, R. Lu, H. Wang, J. Tang, and Y. Dong. Agentrl: Scaling agentic reinforcement learning with a multi-turn, multi-task framework. *arXiv preprint arXiv:2510.04206*, 2025. doi: 10.48550/arXiv.2510.04206. URL https://arxiv.org/abs/2510.04206.

K. Zhang, S. Zeng, E. Hua, N. Ding, Z.-R. Chen, Z. Ma, H. Li, G. Cui, B. Qi, X. Zhu, X. Lv, J.-F. Hu, Z. Liu, and B. Zhou. Ultramedical: Building specialized generalists in biomedicine. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2406.03949. URL https://arxiv.org/abs/2406.03949.

Z. Zhu, C. Xie, X. Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. https://github.com/THUDM/slime, 2025. GitHub repository.

Supplementary Material

Appendix A
Dataset Construction and Validation

A.1 Sources, Funnel, and Experimental Language

All source records, online consultation dialogues, simulator interactions, doctor outputs, diagnostic-reader inputs, and evaluator-facing outputs are Chinese. No translation is used in the experimental pipeline. The authoritative runtime prompts are Chinese; English translations are explanatory, non-executable reading aids.

Table S1 traces the record-source construction funnel and the two MedDDC-Eval held-out splits. The 912 training cases, which support the controlled application study rather than the MedDDC-Eval held-out evaluation set, are instantiated once in each of the 3–4 and 5–7 turn buckets. The 100-case Record split is drawn from the same annotated candidate pool with departmental stratification, and the 70-case Dialogue split is constructed from online-consultation sources with frozen diagnosis-label and reference-trajectory inventories. Training and held-out evaluation are separated by case identifier. Beyond identifier-level deduplication, we did not conduct a semantic near-duplicate audit across the heterogeneous source text.

A.2 Case Schema and Reference Construction

The common case object contains a source stratum, initial patient-facing context, a de-identified simulator fact store, reference diagnoses, trajectory targets, and a turn budget. Table S2 specifies how these fields enter simulation and evaluation. A trajectory target pairs a reference question with its associated clinical condition or fact; it denotes a normalized clinical information need rather than a literal gold wording or a unique path. Targets may retain related subquestions as clinically natural bundles, and directional semantic coverage separates information acquisition from wording imitation. Volunteered-target status is computed from each realized dialogue rather than treated as a static case label.

Fifteen medical annotators worked across four batches. They annotated only targets they judged to bear on the diagnosis or treatment plan. Construction followed three stages: case-grounded drafting; medical normalization with a three-level importance rating, of which only the highest-rated targets are retained; and review of clinical relevance, redundancy, and semantic consistency with adjudication before freezing. Because this workflow produces a reviewed consensus reference rather than a matrix of independent parallel labels, a construction-set inter-annotator coefficient is not applicable.

The staged review above provides construction validity: trajectory targets are reviewed normative information needs, not an assertion that one consultation sequence is uniquely correct. Scoring integrity and human alignment are evaluated separately below.

Appendix B
Evaluator Contract and One-to-One Assignment

B.1 Matching and Aggregation

The evaluator extracts diagnostic items and clinically meaningful doctor-question units, obtains directional, protocol-admissible coverage edges from the LLM-assisted judge, and applies deterministic maximum-cardinality bipartite matching. If m predictions and n references yield a matching of size k, then $\mathrm{TP} = k$, $\mathrm{FP} = m - k$, and $\mathrm{FN} = n - k$. Candidate edges after the first summary action are excluded from trajectory scoring, and reference items already volunteered by the patient are filtered before matching.

Candidate, selected, excluded, and unmapped edges are recorded. MedDDC-Eval v1.0 micro-aggregates observed counts within each run and reports the mean and sample SD across five fixed-checkpoint rollout repeats. One saved semantic-edge artifact was unavailable; the corresponding aggregate uses the observed counts without imputation. Held-out diagnosis generation, extraction, and candidate generation use DeepSeek-v3 through frozen role-specific gateways; training reward-side generation, extraction, and judging instead use DeepSeek-v3.2. The original Chinese prompts are authoritative runtime artifacts.

B.2 Efficiency

Let $\mathcal{T} = \{t_1, \ldots, t_m\}$ contain the first matched rounds before the first summary action. If no item is matched, $E = 0.1$. Otherwise, with $\lambda = 0.15$,

$$w_i = \exp[-\lambda(t_i - 1)], \qquad \bar{t} = \frac{\sum_i t_i w_i}{\sum_i w_i}. \tag{5}$$

The timing map is

$$\mathrm{timing}(\bar{t}) = \begin{cases} 1.0, & \bar{t} \le 2, \\ 0.95 - 0.05(\bar{t} - 2), & 2 < \bar{t} \le 3, \\ 0.90 - 0.15(\bar{t} - 3), & 3 < \bar{t} \le 5, \\ 0.60 - 0.10(\bar{t} - 5), & 5 < \bar{t} \le 8, \\ \max(0.1, 0.30 - 0.02(\bar{t} - 8)), & \bar{t} > 8. \end{cases} \tag{6}$$

With T the first summary round, or exported total rounds when no summary is found,

$$r_{\mathrm{ineff}} = \max\left(0, \frac{T - m}{T}\right), \tag{7}$$

$$E = \min\left(1, \max\left(0.1, \mathrm{timing}(\bar{t})\left(1 - 0.3 r_{\mathrm{ineff}}\right)\right)\right). \tag{8}$$

Appendix C
Evaluator Meta-Evaluation and Measurement Boundaries

(a) Record-source construction funnel and dataset roles

Construction output	Cases	Prompt rows	Scope	Role
Initial record pool	3,616	–	153 source department labels	Construction input
Eligible annotated candidates	2,904	–	139 normalized secondary departments	Candidate pool
Curated RL cases	912	1,824	3–4 and 5–7 turn buckets	Main RL training
Scaled RL candidates	2,904	5,808	Same two turn buckets	Exploratory scaling only
Record-driven evaluation	100	100	Held-out record-derived cases	Main evaluation
Dialogue-driven evaluation	70	70	Held-out dialogue-derived cases	Source-shift evaluation

(b) Frozen held-out reference inventory

Stratum	Cases	Diagnosis labels	Trajectory targets	Targets/case, median (range)
Record-driven	100	190	863	8 (5–18)
Dialogue-driven	70	97	588	8 (4–14)
Total	170	287	1,451	–

Table S1: Non-identifying dataset construction and held-out reference statistics. A training case yields two prompt rows because it is instantiated under two turn-budget buckets. Held-out labels and targets are counted once per unique case before consultation-specific volunteered-fact filtering. For the Dialogue split, online-consultation sources are converted into the same two frozen reference inventories: diagnosis labels and trajectory targets. Counts describe construction outputs rather than annotator-hours, population prevalence, or sampling rates. Raw records and original dialogues are not released.

The meta-evaluation combines fixed-checkpoint repeats, local aggregate-weight perturbations, alternative diagnostic readers, semantic candidate-edge cross-auditing, and physician-anchored professional-quality assessment. In the 1,448-pair stratified cross-audit, GPT-5.5 and Gemini 3.1 Pro yield candidate-edge precision/recall of $0.935/0.430$ and $0.877/0.455$; raw cross-auditor agreement is 0.960 with Cohen's $\kappa = 0.873$. Per-system recall spans 0.340–0.560 and 0.385–0.607, with no detected heterogeneity across the eight systems ($p = 0.53$ and 0.59); the trained policy's audited recall is not higher than the external agents'. These are model-based edge audits, not physician adjudication. The distinct 80-output human-alignment audit uses six medically qualified reviewers who independently scored anonymized outputs under blinded conditions, with disagreements resolved by a senior physician. It reports exact professional-quality agreement of 62.5% and quadratic-weighted $\kappa = 0.415$ between automatic ratings and the adjudicated human reference ratings. It supplies physician-anchored ordinal grounding, not D/T/E calibration, semantic-edge physician validation, or clinical certification.

Appendix D
Fixed-History Diagnostic-Reader Intervention

Supplementary Table S7 provides the complete eight-policy fixed-history audit underlying the main-paper reader-intervention figure. Each rollout history and the diagnostic prompt are held fixed; only the terminal diagnostic reader changes from the rollout policy to shared DS-V3. The table retains all exact diagnosis-F1 endpoints and rank changes, including the trained policy excluded from the figure's seven-policy aggregate braid.

Appendix E
Controlled Optimization Details

E.1 Rollout, Reward, and Ablation Controls

Table S9 summarizes the interaction unit, termination rule, grouped sampling, reward timing, credit scope, optimizer, and train–evaluation separation used in the controlled optimization study.

Each doctor output is either a question or a structured diagnosis action. The grounded patient simulator responds from case information, and rollouts stop at the diagnosis action or the seven-turn limit. GRPO normalizes grouped rollout rewards and broadcasts the normalized sequence reward across response tokens. The active reward is

$$R = R_{\mathrm{diag}} + R_{\mathrm{traj}} + \lambda_{\mathrm{turn}} R_{\mathrm{turn}} + \lambda_{\mathrm{format}} R_{\mathrm{format}} + \lambda_{\mathrm{rep}} R_{\mathrm{rep}} + R_{\mathrm{qcount}} + \lambda_{\mathrm{prefix}} R_{\mathrm{prefix}}. \tag{9}$$

This training signal is distinct from the held-out $0.5D + 0.4T + 0.1E$ aggregate Total formula. Reward ablations remove only diagnosis-result feedback or trajectory feedback while preserving the simulator, auxiliary constraints, shared reader, held-out evaluator, and score coefficients.

Appendix F
Main Result Uncertainty and Robustness

F.1 Paired Cases and Weight Sensitivity

The paired analysis averages repeated rollouts within each case, pairs the trained policy and a baseline by case, and bootstraps mean differences over cases with 10,000 resamples. It quantifies held-out case-sampling uncertainty for fixed trained checkpoints, not training-seed variability.

Case field	Construction meaning	Simulation, scoring, and release role
Source stratum s	Record-driven or dialogue-driven provenance.	Defines the evaluation stratum; only aggregate stratum labels are public.
Initial context x	Patient-facing complaint or retained dialogue context used to start the consultation.	Text is released only in synthetic or approved fully anonymized examples.
Grounded fact store F	De-identified case facts available to the patient simulator.	Constrains patient replies; patient-derived source text and linkages are not released.
Diagnosis set G	Reference diagnostic labels supported by the case.	Used by one-to-one diagnosis scoring; labels are public only in synthetic or approved controlled instances.
Trajectory target q	Normalized clinical information need the doctor should elicit, not a verbatim source span or unique gold wording.	Scored through directional semantic coverage under inquiry-type constraints.
Reference question	Case-grounded natural-language question associated with the trajectory target.	Provides the reference node for coverage without requiring the model to reproduce its wording.
Clinical condition/fact	Case fact or condition that the reference question is intended to elicit.	Grounds interpretation of the target while protected source text and linkages remain unreleased.
Importance rating	Three-level construction-time rating of how strongly a drafted target bears on the diagnosis or treatment plan.	Only highest-rated targets enter the frozen reference set; each frozen target counts once in one-to-one TP/FP/FN.
Volunteered status	Dialogue-specific indicator that the patient supplied the target before it was asked.	Derived after rollout and filtered from that dialogue's denominator; not a static gold label.
Turn budget b	Maximum consultation horizon for the instantiated prompt.	Controls rollout length; training cases use the 3–4 and 5–7 buckets.

Table S2: Common case and annotation schema. Reference trajectories are sets of reviewed trajectory targets, not a single literal sequence. Related subquestions may remain bundled as a natural inquiry unit; one-to-one assignment limits each predicted and reference unit to one credited edge.

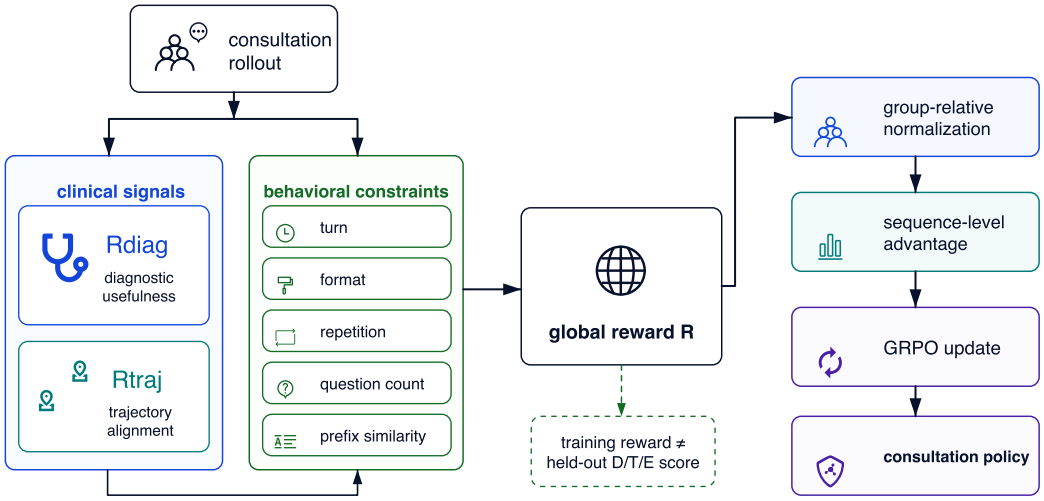

Figure 4: Process-aware reward and GRPO update. Diagnosis usefulness, trajectory alignment, and behavioral constraints form one trajectory-level training reward after a complete rollout. Group-relative normalization yields a sequence-level advantage for the standard GRPO update; this training reward is distinct from the held-out D/T/E score.

F.2 Alternative Readers

Table S17 replaces DS-V3 with Qwen3-32B or GPT-5.4 on the same stored consultation histories. Diagnosis is recomputed under the same one-to-one assignment; trajectory and efficiency remain fixed. The trained policy has the highest mean total for both tested readers and both sources, although the closest setting is not interpreted as a resolved pairwise advantage.

Evaluator step	LLM / parameters	Expected output	Fallback or validation
Medical item extraction	DeepSeek-v3; temperature 0.2, top-p 0.9, JSON mode when used.	Dimension-specific top-k lists for diagnosis, tests, treatment, medication, lifestyle advice, and medical guidance.	Structured tool-call content is used first when available; otherwise the evaluator extracts from the full dialogue. Invalid or missing fields are normalized to empty lists.
Dimension coverage judgment	Configured dimension-judgment prompt with the same default evaluator model.	Directional, protocol-admissible coverage edges between extracted predictions and references.	Parsed JSON is structure-validated; malformed items produce no candidate edge. A deterministic maximum-cardinality matcher selects the final one-to-one edges.
Question extraction	DeepSeek-v3 with the configured trajectory-question-extraction prompt; JSON object response.	List of doctor questions extracted from the dialogue.	If LLM extraction fails, the extracted list is empty and downstream trajectory matching receives no matched questions.
Question-to-turn alignment	String containment, then LLM semantic match, then keyword containment.	Round index for each extracted doctor question.	Semantic match uses the trajectory-matching prompt; failure falls back to keyword containment only when a sufficiently long question string is available.
Trajectory coverage matching	Configured trajectory-matching prompt; JSON object response.	Candidate question–target coverage edges with aligned rounds.	Bundled targets permit valid component coverage under type constraints; post-summary and invalid edges are discarded; deterministic one-to-one matching produces the TP/FP/FN counts.
Patient-volunteered filtering	DeepSeek-v3 with patient-coverage prompt; temperature 0.0, top-p 0.9, JSON mode.	Whether an unmatched gold target was already volunteered by the patient.	If model judgment is unavailable, string fallback checks whether the question or condition text appears in patient turns.
Efficiency computation	Deterministic after matched rounds are available.	Timing score from matched trajectory rounds and ineffective-question penalty.	No additional LLM call is used for the timing formula.

Table S3: LLM-assisted candidate generation followed by deterministic one-to-one assignment. The LLM proposes semantic edges; programmatic matching determines the final confusion counts.

Policy	Rec. Q3	Rec. GPT	Dial. Q3	Dial. GPT
Qwen3-32B + GRPO	**50.0±2.1**	**50.5±0.7**	**46.0±0.9**	**44.8±1.2**
Qwen3-32B	40.5±1.3	42.8±0.8	40.5±1.0	40.0±1.5
GPT-5.4	47.4±1.8	48.1±1.2	44.1±2.3	44.3±0.7
GLM-5	44.5±4.3	45.6±3.7	40.0±1.2	40.6±1.4
HuatuoGPT-o1-70B	41.9±0.7	42.3±1.0	38.7±3.4	39.3±2.4
Llama-3-70B-UltraMedical	43.9±0.7	45.1±1.0	41.1±1.4	42.8±1.3
Baichuan-M2-32B	38.3±2.1	38.8±1.8	37.6±0.8	39.1±0.4
Doctor-R1	37.8±1.5	37.2±1.0	37.7±2.4	40.3±2.4

Table S17: Alternative-reader total (%), mean ± sample SD over five runs. Q3 denotes the Qwen3-32B reader and GPT denotes the GPT-5.4 reader; only the reader changes.

Appendix G Diagnosis–Trajectory Complementarity

System means show a positive diagnosis–trajectory association, while case–model associations are weak. Split-specific 33rd/67th percentile buckets expose both off-diagonal directions. This supports measurements that are aligned at the model-objective level but non-interchangeable at the individual-output level.

Validation channel	Evaluation unit	Main statistic / result	Evidential role
Repeated evaluation	16 model–split summaries over five rollout repeats	Median Total SD: 1.41 points	Characterizes fixed-checkpoint rollout variability.
Score-weight sensitivity	Seven D/T/E weight settings on both held-out sources	Top ordering preserved; margins 3.25–4.34 points	Tests dependence on the prespecified aggregate weights.
Alternative readers	Two replacement readers on fixed consultation histories and both sources	Mean top ordering preserved	Tests dependence on terminal diagnostic-reader choice.
Candidate-edge cross-audit	1,448 pair labels independently audited by two models	Raw agreement 0.960; Cohen's $\kappa = 0.873$	Tests consistency of protocol-valid semantic-edge judgments.
Human professional-quality alignment	80 outputs; six blinded medical reviewers; senior-physician adjudication	Exact agreement 62.5%; QW $\kappa = 0.415$	Provides physician-anchored ordinal agreement with adjudicated human reference ratings.

Table S4: Complementary meta-evaluation channels for MedDDC-Eval. Repeated runs characterize rollout variability, reader and weight checks probe sensitivity, cross-auditing inspects semantic judgments, and the human-alignment audit supplies physician-anchored ordinal grounding.

(a) Directional coverage-edge cross-audit

Auditor	Scope	Pairs	Candidate	Protocol-pos.	TP/FP/FN	Precision	Recall
GPT-5.5	All	1,448	138	300	129/9/171	0.935	0.430
GPT-5.5	Diagnosis	113	17	48	17/0/31	1.000	0.354
GPT-5.5	Trajectory	1,335	121	252	112/9/140	0.926	0.444
Gemini 3.1 Pro	All	1,448	138	266	121/17/145	0.877	0.455
Gemini 3.1 Pro	Diagnosis	113	17	36	17/0/19	1.000	0.472
Gemini 3.1 Pro	Trajectory	1,335	121	230	104/17/126	0.860	0.452

(b) Extracted-unit cross-audit

Auditor	Items	Valid	Validity/precision	Atomic	Atomicity	Omission tasks	Recall
GPT-5.5	267	181	0.678	84	0.315	64/64	0.588
Gemini 3.1 Pro	267	206	0.772	126	0.472	0/64	–

Table S5: Aggregate LLM cross-audit of the automatic evaluator. A protocol-positive pair is a directional match under the prespecified dimension, subdimension, and bundled-target rules, rather than strict bidirectional equivalence. The item audit defines validity by source-span presence and clinical eligibility; atomicity is reported separately because bundled trajectory questions are valid evaluation units. Across all 1,448 pair labels, raw cross-auditor agreement is 0.960 and Cohen's κ is 0.873. The output-level sample was stratified by automatic-score strata, so precision and recall describe this audit sample and are not population-weighted estimates over all rollouts. Gemini's protocol-aware omission audit was not completed, and its extraction recall is therefore not reported. These are model-based cross-audit results, not physician adjudication or clinical validation.

Level	Split	n	r	ρ	HT/LD; LT/HD
Model means	Record	8	0.569	0.238	–
Model means	Dialogue	8	0.303	0.381	–
Case–model means	Record	800	0.147	0.139	70; 76
Case–model means	Dialogue	560	0.101	0.086	50; 63

Table S18: Diagnosis–trajectory association under protocol-admissible coverage scoring with deterministic one-to-one assignment. System means are positively associated, while case–model associations are weak and both off-diagonal directions remain common under split-specific 33rd/67th percentile thresholds. This supports treating diagnosis and trajectory as related but non-substitutable axes.

Appendix H
Interpretive and Deployment Boundaries

The semantic judge and frozen diagnostic reader remain model components. Directional coverage edges are protocol-admissible relations, not claims of literal equivalence. One-to-one assignment constrains counting but does not guarantee that every clinically valid coverage edge is proposed; the coverage-aware audit found higher candidate precision than recall. The public prompts, candidate-edge schema, matching code, and aggregate audits make this dependency inspectable. Reader swaps test terminal-generation dependence; they do not validate the reader as a clinical authority.

Trajectory targets are reviewed normative information needs expressed as clinically natural question units. Related

Aspect	Protocol / observation	Evidential interpretation
Sampling	20 cases $\times$ 4 systems, yielding 80 anonymized consultation outputs.	Output-level audit spanning multiple systems and cases.
Review	Six medically qualified reviewers independently scored anonymized outputs under blinded conditions; disagreements were resolved by a senior physician.	The adjudicated rating provides one human reference rating per output.
Agreement	Exact agreement $= 62.5\%$; quadratic-weighted Cohen's $\kappa = 0.415$.	Moderate ordinal agreement beyond chance; larger score discrepancies receive greater penalty.
Role	Physician-anchored agreement complements repeated-run, reader-swap, and semantic-audit evidence.	Ordinal grounding for professional-quality assessment; component calibration and clinical safety require separate validation.

Table S6: Physician-anchored human-alignment audit. Six medically qualified reviewers independently scored anonymized outputs under blinded conditions, with disagreements resolved by a senior physician. Exact agreement compares automatic professional-quality ratings with the adjudicated human reference ratings; quadratic-weighted Cohen's κ additionally corrects for chance agreement and disagreement severity.

Policy	Diagnosis F1: bound $\rightarrow$ shared	Rank change
Record-driven		
Qwen3-32B + GRPO	40.5 $\rightarrow$ 44.6 (+4.1)	1 $\rightarrow$ 1
Qwen3-32B	26.5 $\rightarrow$ 33.1 (+6.6)	6 $\rightarrow$ 5
GPT-5.4	33.2 $\rightarrow$ 36.3 (+3.1)	3 $\rightarrow$ 4
GLM-5	33.9 $\rightarrow$ 38.0 (+4.1)	2 $\rightarrow$ 2
HuatuoGPT-o1-70B	33.0 $\rightarrow$ 37.7 (+4.7)	4 $\rightarrow$ 3
Llama-3-70B-UltraMedical	29.2 $\rightarrow$ 31.4 (+2.2)	5 $\rightarrow$ 8
Baichuan-M2-32B	16.6 $\rightarrow$ 32.3 (+15.7)	8 $\rightarrow$ 6
Doctor-R1	24.9 $\rightarrow$ 31.5 (+6.6)	7 $\rightarrow$ 7
Dialogue-driven		
Qwen3-32B + GRPO	36.5 $\rightarrow$ 41.8 (+5.3)	1 $\rightarrow$ 1
Qwen3-32B	29.6 $\rightarrow$ 39.8 (+10.2)	5 $\rightarrow$ 2
GPT-5.4	30.9 $\rightarrow$ 35.5 (+4.6)	3 $\rightarrow$ 6
GLM-5	30.5 $\rightarrow$ 34.3 (+3.8)	4 $\rightarrow$ 7
HuatuoGPT-o1-70B	32.4 $\rightarrow$ 37.2 (+4.8)	2 $\rightarrow$ 5
Llama-3-70B-UltraMedical	20.8 $\rightarrow$ 39.5 (+18.7)	7 $\rightarrow$ 3
Baichuan-M2-32B	14.5 $\rightarrow$ 33.5 (+19.0)	8 $\rightarrow$ 8
Doctor-R1	27.8 $\rightarrow$ 37.4 (+9.6)	6 $\rightarrow$ 4

Table S7: Fixed-history diagnostic-reader intervention. The policy-bound condition applies each rollout policy as the diagnostic reader under a common diagnostic prompt; the shared condition applies DS-V3 to the identical history. Questions, patient answers, trajectory score, and efficiency are held fixed. The heterogeneous diagnosis-F1 shifts and rank changes show that reader identity materially affects diagnosis-based comparisons of elicited histories.

subquestions may remain bundled because physicians commonly ask them together; a prediction can receive coverage for one valid component under the prespecified inquiry-type constraints. This is deliberately different from atomic-fact recall and gives coarser within-bundle resolution. The targets allow semantically compatible questions and remove facts already volunteered by the simulated patient, but they do not enumerate every acceptable consultation strategy. Diagnosis, trajectory, and efficiency should therefore be read together. In particular, efficiency measures when matched information is collected and is mathematically coupled to trajectory matching.

The physician-anchored audit measures alignment of ordinal professional-quality ratings. Six medically qualified reviewers independently scored anonymized outputs under blinded conditions, with disagreements resolved by a senior physician. The 62.5% exact agreement and quadratic-weighted Cohen's $\kappa = 0.415$ compare automatic ratings with the adjudicated human reference ratings; they do not calibrate every D/T/E component or establish clinical safety. MedDDC-Eval remains an offline Chinese-language research setting and does not establish clinical safety or deployment readiness.

Configuration	Group Relative Policy Optimization (GRPO) setting
Base family	Qwen3-32B
Initialization	Qwen3-32B checkpoint
Rollout function	Custom multi-turn rollout
Reward function	Process-aware reward including diagnosis-result feedback and trajectory F1
Advantage estimator	GRPO
Maximum turns	7
Samples per prompt	8
Rollout batch size	32
Rollouts	300
Maximum response length	2,048 tokens
Sampling temperature	1.0
Global batch size	64
KL loss coefficient	0.005
Clipping	$\epsilon = 0.2$, high clip 0.28
Optimizer	Adam, learning rate 1×10^{-6}, weight decay 0.01
Parallelism	1 node, 8 actor GPUs, tensor parallelism 4

Table S8: Process-aware GRPO training configuration.

Field	Reported setting	Interpretive boundary
Training unit / interaction	Complete doctor–patient trajectory; policy alternates with the grounded simulator.	Multi-turn structure is introduced through environment interaction.
Termination	Structured diagnosis action or seven-turn limit.	Defines the trajectory receiving one global reward.
GRPO group	Eight sampled trajectories per prompt.	Rewards are normalized within the prompt group.
Reward timing and signals	Computed after the complete rollout from diagnosis-result and trajectory feedback, plus behavioral constraints.	One reward is assigned to each complete trajectory.
Advantage / credit scope	Group-normalized sequence reward broadcast across policy-generated response tokens.	Credit assignment operates at the trajectory level.
Optimizer	Standard GRPO implemented with SLIME.	Process awareness is encoded in the reward signals.
Train–evaluation separation	Disjoint cases; reward-side terminal generation, extraction, and judging use DS-3.2; held-out reader and evaluator use DS-V3.	Evaluator versions are stage-specific: DS-3.2 for training rewards and DS-V3 for held-out evaluation.

Table S9: Interactive multi-turn rollout and GRPO update contract. Multi-turn structure enters through environment interaction and trajectory-level reward computation; grouped normalization and policy optimization use standard GRPO.

Component	Implementation variable	Scale / coefficient	Computation summary
Diagnosis result	`accuracy_score`	no outer multiplier	Frozen diagnostic-tool output is compared with the reference diagnosis.
Trajectory match	`reward_trajectory_f1`	no outer multiplier	Weighted trajectory F1 from semantic matching of doctor questions to reference targets.
Turn budget	`reward_turn`	$\lambda_{\text{turn}} = 0.8$	Reward for stopping within the prompt-specific target turn bucket.
Format compliance	`reward_from_format`	$\lambda_{\text{format}} = 4.0$	XML/tool-call format score averaged across assistant turns.
Repetition control	`repetition_score`	$\lambda_{\text{rep}} = 0.5$	Repetition score clipped at 6 before scaling.
Question count	`question_num_score`	no outer multiplier	Per-turn question-count compliance with penalties for under- or over-questioning.
Prefix control	`prefix_penalty`	$\lambda_{\text{prefix}} = 0.5$	Penalty for overly similar response prefixes across turns.

Table S10: Active reward components in the reported process-aware training configuration. Style and session-level total-length terms are disabled, so they are not part of the active reward formula or reward-component ablations. The training reward is distinct from the held-out diagnosis–trajectory–efficiency total score.

System group	Control / status	Evidential role
Qwen3-32B	Starting checkpoint; same simulator, budget, tools, and metrics	Same-family reference for the controlled training contrast
Qwen3-32B + GRPO	Process-aware rewards; same held-out splits and score formula	Headline controlled training result
GPT-5.4 / GLM-5	Prompt-only external systems under the common protocol	Context against strong general systems
Doctor-R1 / HuatuoGPT-o1 / UltraMedical / Baichuan-M2	External medical or clinical-inquiry systems	Tests whether specialization alone resolves trajectory alignment

Table S11: Baseline protocol and interpretation. Trained same-family comparisons are separated from external references so closed-model comparisons are not treated as optimizer ablations.

Agent / policy	Diagnosis F1 (%)	Trajectory F1 (%)	Efficiency (%)	Total (%)
Qwen3-32B + GRPO	**44.6**$\pm$**3.0**	57.3$\pm$1.5	**83.0**$\pm$**0.6**	**53.5**$\pm$**1.3**
GPT-5.4	36.3$\pm$2.3	**58.8**$\pm$**0.7**	79.7$\pm$0.7	49.7$\pm$1.3
GLM-5	38.0$\pm$7.7	54.4$\pm$1.6	66.4$\pm$16.6	47.4$\pm$5.4
HuatuoGPT-o1-70B	37.7$\pm$1.5	46.2$\pm$2.0	77.9$\pm$0.5	45.1$\pm$1.1
Llama-3-70B-UltraMedical	31.4$\pm$2.4	50.5$\pm$1.7	81.2$\pm$1.0	44.0$\pm$1.0
Qwen3-32B	33.1$\pm$0.9	48.1$\pm$1.3	80.2$\pm$1.1	43.8$\pm$1.0
Baichuan-M2-32B	32.3$\pm$3.2	46.3$\pm$2.9	81.8$\pm$1.2	42.9$\pm$2.1
Doctor-R1	31.5$\pm$1.8	40.3$\pm$2.7	75.0$\pm$2.1	39.3$\pm$1.5

Table S12: Component results on 100 record-driven held-out cases, ordered by total score. Values are mean $\pm$ sample standard deviation over five repeated runs. Diagnosis and trajectory F1 use protocol-admissible coverage edges followed by deterministic one-to-one assignment within each run; efficiency uses the first summary round and a lower bound of 0.1 when no valid trajectory item is matched.

Agent / policy	Diagnosis F1 (%)	Trajectory F1 (%)	Efficiency (%)	Total (%)
Qwen3-32B + GRPO	**41.8**$\pm$**2.1**	**52.6**$\pm$**1.1**	**82.4**$\pm$**1.2**	**50.2**$\pm$**1.3**
GPT-5.4	35.5$\pm$2.2	52.4$\pm$0.7	79.1$\pm$0.3	46.6$\pm$1.0
Llama-3-70B-UltraMedical	39.5$\pm$3.6	46.9$\pm$2.2	79.4$\pm$1.5	46.4$\pm$2.5
Qwen3-32B	39.8$\pm$1.4	44.7$\pm$1.9	78.0$\pm$2.0	45.6$\pm$1.5
Doctor-R1	37.4$\pm$2.4	41.2$\pm$2.6	78.6$\pm$3.4	43.1$\pm$1.5
GLM-5	34.3$\pm$4.2	44.7$\pm$1.7	74.9$\pm$2.0	42.5$\pm$2.2
Baichuan-M2-32B	33.5$\pm$2.8	44.2$\pm$1.2	80.8$\pm$2.2	42.5$\pm$1.3
HuatuoGPT-o1-70B	37.2$\pm$2.8	40.8$\pm$2.6	75.9$\pm$2.3	42.5$\pm$2.2

Table S13: Component results on 70 dialogue-driven held-out cases, ordered by total score. Values are mean $\pm$ sample standard deviation over five repeated runs. Diagnosis and trajectory F1 use protocol-admissible coverage edges followed by deterministic one-to-one assignment within each run; efficiency uses the first summary round and a lower bound of 0.1 when no valid trajectory item is matched.

Split	Baseline	Cases	Mean diff.	95% CI	$P(\Delta > 0)$	W/T/L
Dialogue	Baichuan-M2-32B	70	11.5	[8.8, 14.4]	>0.9999	58/2/10
Dialogue	GLM-5	70	8.3	[5.8, 10.8]	>0.9999	57/4/9
Dialogue	Doctor-R1	70	7.0	[4.4, 9.5]	>0.9999	53/2/15
Dialogue	HuatuoGPT-o1-70B	70	6.8	[4.2, 9.3]	>0.9999	47/7/16
Dialogue	Qwen3-32B	70	5.2	[2.7, 7.6]	>0.9999	50/5/15
Dialogue	GPT-5.4	70	3.3	[0.5, 6.2]	0.9881	41/4/25
Dialogue	Llama-3-70B-UltraMedical	70	3.3	[0.8, 5.8]	0.9958	43/4/23
Record	Doctor-R1	100	15.3	[12.9, 17.8]	>0.9999	89/4/7
Record	Baichuan-M2-32B	100	12.0	[9.5, 14.4]	>0.9999	72/8/20
Record	GLM-5	100	11.3	[8.8, 14.0]	>0.9999	77/4/19
Record	Llama-3-70B-UltraMedical	100	11.1	[8.7, 13.6]	>0.9999	78/8/14
Record	Qwen3-32B	100	10.6	[8.4, 12.9]	>0.9999	77/6/17
Record	HuatuoGPT-o1-70B	100	8.5	[5.7, 11.3]	>0.9999	65/8/27
Record	GPT-5.4	100	3.9	[1.4, 6.5]	0.9989	54/7/39

Table S14: Case-level paired bootstrap for the trained Qwen3-32B + GRPO policy minus each baseline. Each case is first averaged over repeated runs; differences and confidence intervals are percentage points from 10,000 bootstrap resamples. W/T/L uses a one-point tie tolerance.

Split	Baseline	Component	Mean diff.	95% CI	$P(\Delta > 0)$
Dialogue	GPT-5.4	Diagnosis	5.7	[0.4, 11.0]	0.9835
Dialogue	GPT-5.4	Efficiency	3.3	[2.1, 4.4]	>0.9999
Dialogue	GPT-5.4	Trajectory	0.3	[-2.9, 3.5]	0.5664
Dialogue	Qwen3-32B	Diagnosis	2.4	[-2.0, 6.9]	0.8634
Dialogue	Qwen3-32B	Efficiency	4.4	[2.2, 7.0]	>0.9999
Dialogue	Qwen3-32B	Trajectory	8.8	[5.8, 11.6]	>0.9999
Record	GPT-5.4	Diagnosis	7.3	[2.8, 11.8]	0.9992
Record	GPT-5.4	Efficiency	3.4	[2.0, 4.7]	>0.9999
Record	GPT-5.4	Trajectory	-0.3	[-3.3, 2.7]	0.4259
Record	Qwen3-32B	Diagnosis	12.2	[8.7, 15.9]	>0.9999
Record	Qwen3-32B	Efficiency	2.8	[1.3, 4.4]	0.9999
Record	Qwen3-32B	Trajectory	10.6	[7.9, 13.2]	>0.9999

Table S15: Component-level paired bootstrap for the controlled Qwen3-32B comparison and strongest external total-score reference. Differences are the trained Qwen3-32B + GRPO policy minus baseline in percentage points.

Weights $(D/T/E)$	Record top	Margin	Dialogue top	Margin
0.50 / 0.40 / 0.10	Qwen3-32B + GRPO	3.85	Qwen3-32B + GRPO	3.56
0.55 / 0.35 / 0.10	Qwen3-32B + GRPO	4.34	Qwen3-32B + GRPO	3.58
0.45 / 0.45 / 0.10	Qwen3-32B + GRPO	3.36	Qwen3-32B + GRPO	3.25
0.45 / 0.40 / 0.15	Qwen3-32B + GRPO	3.61	Qwen3-32B + GRPO	3.41
0.50 / 0.35 / 0.15	Qwen3-32B + GRPO	4.10	Qwen3-32B + GRPO	3.61
0.55 / 0.40 / 0.05	Qwen3-32B + GRPO	4.10	Qwen3-32B + GRPO	3.71
0.50 / 0.45 / 0.05	Qwen3-32B + GRPO	3.61	Qwen3-32B + GRPO	3.41

Table S16: Local score-weight sensitivity. Margins are percentage-point differences between the top and second system. The trained Qwen3-32B + GRPO policy remains first for all tested weight vectors; coefficients sum to one.

Split	System	Runs	Cases	Diag.	Traj.	Total	HT/LD ≥ 3	LT/HD ≥ 3
Dialogue	Qwen3-32B + GRPO	5	70	44.2	53.3	51.7	4	3
Dialogue	Llama-3-70B-UltraMedical	5	70	42.6	47.9	48.4	1	7
Dialogue	GPT-5.4	5	70	38.6	53.0	48.4	3	8
Dialogue	Qwen3-32B	5	70	41.8	44.5	46.5	3	6
Dialogue	HuatuoGPT-o1-70B	5	70	41.2	41.7	44.8	7	5
Dialogue	Doctor-R1	5	70	40.5	41.5	44.7	9	3
Dialogue	GLM-5	5	70	35.9	44.6	43.3	7	5
Dialogue	Baichuan-M2-32B	5	70	29.1	43.7	40.1	4	4
Record	Qwen3-32B + GRPO	5	100	45.9	58.8	54.8	6	13
Record	GPT-5.4	5	100	38.6	59.1	50.9	7	9
Record	HuatuoGPT-o1-70B	5	100	39.8	46.5	46.3	6	6
Record	Qwen3-32B	5	100	33.7	48.2	44.2	4	4
Record	Llama-3-70B-UltraMedical	5	100	30.0	51.4	43.7	11	4
Record	GLM-5	5	100	37.3	45.5	43.5	3	2
Record	Baichuan-M2-32B	5	100	31.3	47.4	42.8	8	6
Record	Doctor-R1	5	100	32.0	40.1	39.5	5	4

Table S19: Equal-weighted case-level decomposition. Scores are percentages after averaging repeated runs per case. HT/LD and LT/HD count cases assigned to the corresponding model-specific tertile bucket in at least three of five runs. This case-unit view complements, rather than reproduces, the run-level micro aggregate used for the main comparison.

Pattern	Bucket frequency	Case-derived setting	Evidence summary	Five-run change
Stable process improvement	62/169 (36.7%)	Acute abdominal pain	The trained policy more consistently covers pain course, associated gastrointestinal symptoms, triggers/history, prior tests, and instability signs; the starting policy is more variable and sometimes repeats narrow symptom checks.	Trajectory $+22.5$ pp; diagnosis -0.7 pp
Diagnosis gain with process shortfall	7/169 (4.1%)	Acute jaundice-like presentation	The relevant diagnostic family is recovered more often, while specific exposure, associated-symptom, and history targets remain unelicited; the trained policy's trajectory F1 remains 38.6.	Diagnosis $+43.3$ pp; trajectory -1.2 pp

Table S20: Case-level complementarity under the current protocol-admissible coverage endpoint. The first bucket requires a trajectory gain of at least 10 points with cross-run superiority at least 0.70 and no diagnosis loss above 5 points. The second requires low-trajectory/high-diagnosis behavior in at least three of five trained-policy runs, a diagnosis gain of at least 10 points, and diagnosis superiority at least 0.70. Displayed cases are candidates nearest the multivariate bucket center, not maximum-gain cases. Case descriptions are derived from controlled artifacts and conceptually paraphrased; no raw patient text or identifiers are shown.

Appendix I
MedDDC-Eval Reproducibility and Artifact Manifest

I.1 Data Availability and Release Tiers

MedDDC-Eval uses a partial, controlled release strategy. Raw patient records, original consultation dialogues, patient-identifying or re-identifiable information, restricted external data, case-id-level linkage tables, and raw rollout traces are not public. Regardless of source-data authorization, the planned minimum release includes the evaluation schema, approved original Chinese prompt sources, non-executable English prompt translations, matching and scoring code, aggregate tables, and a fully synthetic end-to-end fixture. When permitted by source authorization, de-identification requirements, and institutional review constraints, we additionally plan to release approved subsets of the test data. Upon acceptance, we will prioritize clearance for the held-out evaluation set and runnable source-derived instances; release remains conditional on source authorization, protocol validation, and de-identification review.

Controlled access depends on authorization, privacy, licensing, and data-use terms. The system is for offline research; scores are comparison metrics, not clinical decisions or patient-care judgments.

I.2 Evaluator Procedure

The released evaluator logic follows this sequence:

1. Simulate a bounded consultation and identify the first summary action.
2. Run the shared frozen reader on the elicited history.
3. Extract diagnosis predictions and pre-summary doctor-question units.
4. Generate directional candidate coverage edges for diagnosis predictions and trajectory targets.
5. Filter trajectory targets volunteered before they were asked.
6. Select deterministic one-to-one matchings and compute micro D/T F1.
7. Compute first-summary efficiency and total $0.5D + 0.4T + 0.1E$.

Artifact	Release level	Boundary and role
Evaluation schema	Planned public	No patient text; recreates fields and scoring interfaces.
Annotation format	Planned public	Normalized target fields only; supports independent trajectory scoring.
Evaluation prompts	Public after review	Requires removal of internal identifiers and protected examples.
Prompt manifest	Planned public	Maps runtime roles to authoritative Chinese prompt keys and release tiers.
English translations	Public after bilingual review	Explanatory, non-executable, and not used in experiments.
Scoring code	Public where licensed	Recomputes D/T/E/total from compatible inputs; requires license and protocol checks.
Aggregate tables	Planned public	Model-level statistics for reported numbers and ablations.
Aggregate case buckets	Public after review	Bucket counts and aggregate scores only; no case identifiers or raw text.
Case-id decomposition	Controlled if approved	Linkage and repeated-run assignments may expose restricted artifacts.
Synthetic examples	Planned public	Generated without patient-derived text; demonstrates the schema.
Chinese fixture	Planned public	Fully synthetic, runtime-language example validated by the scorer.
De-identified instances	Conditional controlled	No access promised; depends on authorization, licensing, and privacy review.
Approved test subsets	Conditional public/controlled	De-identified test cases may be released only after source authorization, licensing, and institutional review clearance.
Raw records/dialogues	Not released	Protected by privacy, authorization, and source-license constraints.
Identifying fields	Not released	Direct or re-identifiable patient information is excluded from all tiers.

Table S21: Planned release manifest. Public artifacts reproduce evaluator mechanics; controlled artifacts support restricted audit where permitted; raw and re-identifiable source materials are not released.

I.3 Synthetic Schema Example

Table S22 instantiates the public-facing trajectory format without patient-derived text or identifiers. It is a schema and scoring illustration, not an additional evaluation case.

Synthetic item	Gold information target	Accepted question examples
Initial complaint	Adult patient reports recurrent upper-abdominal discomfort after meals.	n/a
Time course	Onset, duration, and progression of abdominal discomfort.	"When did it start?"; "How long does each episode last?"
Symptom character	Location, quality, severity, and relation to meals.	"Where is the pain and how would you describe it?"; "Is it worse after eating?"
Associated symptoms	Nausea, vomiting, fever, weight loss, black stool, or jaundice.	"Any vomiting, fever, black stool, or yellow eyes?"
Red flags	Severe persistent pain, gastrointestinal bleeding, or systemic deterioration.	"Have you had severe continuous pain or signs of bleeding?"
Past history	Prior gastrointestinal disease and relevant chronic conditions.	"Have you had stomach ulcers, gallbladder disease, or liver disease before?"
Medication exposure	NSAID use, anticoagulants, alcohol, and recent treatment.	"Do you take painkillers like ibuprofen or blood thinners?"
Prior tests	Recent endoscopy, abdominal ultrasound, blood tests, or treatment response.	"Have you had any tests or treatment for this before?"

Table S22: Fully synthetic trajectory example for release and reproducibility documentation. It illustrates the annotation and semantic matching format without exposing any patient-derived case text.