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

Guofeng Zhang[1,*,†], Yizeng Quan[1,2,*], Huaiyi Fang[1], Jianwei Lv[1],
Jinyao Liu[1,3], Xunxu Duan[1,2], Lening An[1], Yu Ouyang[1,†], Junfeng Wang[1]

[1]Baidu, Inc. [2]Peking University [3]Zhejiang University

arXiv:2607.18999v2 [cs.CL] 26 Jul 2026

Abstract

Evaluating multi-turn medical consultation agents requires judging the diagnostic support provided by the histories they elicit through interaction. Yet coupled evaluation lets each policy both elicit the history and generate the terminal diagnosis, so a diagnosis score confounds the elicited history with the policy's own terminal diagnosis generator. We introduce **MedDDC-Eval**, a diagnosis-decoupled evaluation testbed over held-out cases derived from medical records and online consultations. It applies the same frozen shared diagnostic reader to every policy-elicited history, holding terminal diagnosis generation fixed across policies and enabling comparison under the shared diagnostic reader. It reports diagnostic support, information-acquisition coverage, and efficiency. LLM-assisted semantic matching followed by deterministic one-to-one assignment makes the diagnosis–trajectory–efficiency (D/T/E) scores auditable. In a fixed-history audit across eight policies, replacing each policy's own generator with the shared diagnostic reader shifts diagnosis F1 by 2.2–19.0 points and reverses 18% and 36% of pairwise orderings on the Record and Dialogue splits. To examine downstream utility, we use standard Group Relative Policy Optimization (GRPO) with a separate training-time reward that targets the same diagnosis and trajectory dimensions. Relative to its Qwen3-32B initialization, the trained policy gains 9.6 and 4.6 aggregate-score points on the held-out Record and Dialogue splits, respectively, and ablating either feedback signal reduces the aggregate score on both. Together, MedDDC-Eval supports comparison under a shared diagnostic reader and evaluation-informed policy development, while complementing end-to-end evaluation when terminal diagnosis generation is also part of the target capability.

Introduction

Multi-turn medical consultation agents do not merely answer from supplied evidence; they help construct the evidence on which diagnosis is based. Unlike static medical QA and diagnosis benchmarks, where a case description provides the evidence in advance [Singhal et al. 2023, Wang et al. 2024], their questions determine which evidence becomes available for diagnosis [Qiao et al. 2026a, Sanghvi et al. 2026, Lai et al. 2025]. They must adapt to patient responses, screen for missing evidence, and decide when a bounded history is sufficient. Evaluation must therefore measure not only the final diagnosis but also the history elicited by the consultation policy.

Recent evaluation suites increasingly make inquiry visible through multimodal and agentic tracks, physician-authored criteria, consultation rubrics, and information-acquisition measures [Ding et al. 2025, Bedi et al. 2025, Tu et al. 2025, Arora et al. 2025, Gong et al. 2026, Qiao et al. 2026a]. Yet a central attribution problem remains. When the evaluated system both elicits the history and generates the terminal diagnosis, its diagnosis score reflects two components: the history elicited by the policy and the interpretation applied by its own terminal diagnosis generator. How can consultation policies be compared without simultaneously comparing their terminal diagnosis generators? The problem is not that terminal generators are necessarily weak; it is that they vary across policies and therefore change the object being compared. A strong generator can compensate for a thin history, while a weaker one can obscure a rich history. Process scores describe parts of the consultation, but the downstream diagnostic value of two elicited histories remains entangled with their terminal diagnosis generators. Coupled scores can therefore misattribute gains, misrank consultation policies, and direct reward design toward a different component than intended.

Figure 1 exposes this confound by holding the history, diagnostic prompt, and decoding settings fixed while changing the history-to-diagnosis model. This intervention shifts diagnosis F1 by 2.2–19.0 points and reverses 18% and 36% of pairwise policy orderings on the Record and Dialogue splits, respectively, in the complete eight-policy audit. Because the histories are identical, the heterogeneous shifts show that model choice alters between-policy comparisons instead of adding a uniform offset.

Our key idea is simple: hold terminal diagnosis generation fixed across policies and treat the policy-elicited history as the evaluation object. We therefore introduce **MedDDC-Eval**, a diagnosis-decoupled evaluation testbed over 170 held-out cases derived from medical records and online consultations. A frozen shared DeepSeek-v3 (DS-V3) diagnostic

*These authors contributed equally.

†Corresponding authors: zhangguofeng@baidu.com and ouyangyu@baidu.com

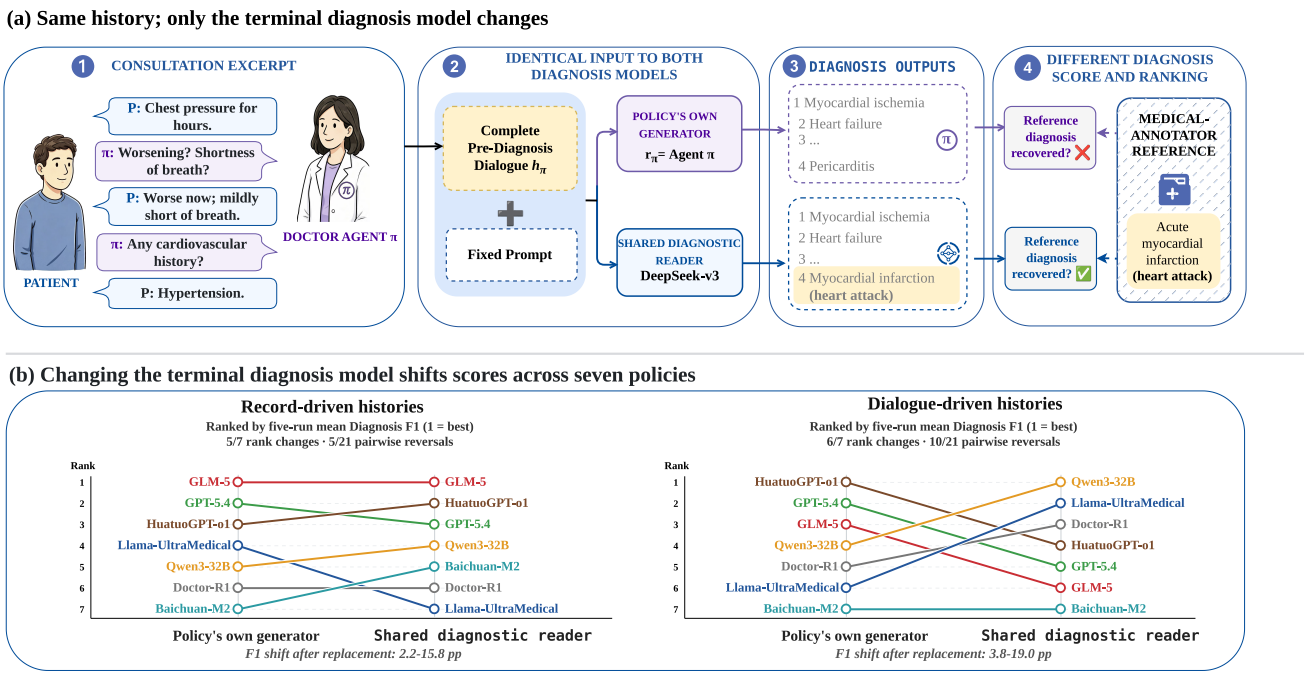

Figure 1: Fixed-history terminal-diagnosis intervention. (a) The history, prompt, and decoding settings are fixed; replacing each policy's own terminal diagnosis generator with the shared diagnostic reader changes which reference diagnoses are recovered. (b) The seven-policy view shows 5/21 Record and 10/21 Dialogue pairwise reversals. Exact endpoints and the complete eight-policy audit (5/28 and 10/28) appear in Supplementary Table S7.

reader [DeepSeek-AI 2025] maps every history to a terminal diagnosis, removing each policy's own terminal diagnosis generator as a source of between-agent variation. This control enables comparison of policy-elicited histories under the shared diagnostic reader and complements end-to-end evaluation when terminal generation is also part of the target capability. The testbed reports diagnostic support, information-acquisition coverage against reviewed reference-question targets, and efficiency separately. LLM-assisted semantic matching recognizes free-form paraphrases; deterministic one-to-one assignment prevents duplicate credit and exposes traceable TP/FP/FN counts.

We next ask whether the evaluation dimensions operationalized by MedDDC-Eval can inform model development. In a controlled evaluation-informed study, standard GRPO [Shao et al. 2024] post-trains Qwen3-32B [Yang et al. 2025] over interactive rollouts using a separate reward with diagnosis and trajectory feedback plus auxiliary behavioral constraints. Under the frozen held-out evaluator, the trained policy gains 9.6 and 4.6 Total points on the two sources; ablating either feedback signal reduces Total on both.

Our contributions are:

1. **Diagnosis-decoupled evaluation.** MedDDC-Eval compares policy-elicited histories under a frozen shared diagnostic reader; a fixed-history intervention shows that the terminal diagnosis generator changes diagnosis scores and rankings.

2. **Auditable elicited-history measurement.** D/T/E characterize diagnostic support, information-acquisition coverage, and efficiency, while semantic matching and one-to-one assignment provide traceable counts.

3. **Evaluation-informed policy development.** A controlled Qwen3-32B study uses a separate training-time reward organized around the diagnosis and trajectory dimensions; the trained policy improves on both held-out sources, and removing either feedback signal lowers joint performance.

Related Work

Medical Evaluation and Inquiry

Conventional static medical QA and diagnosis benchmarks condition on evidence already present in the case [Singhal et al. 2023, Wang et al. 2024]. Broader suites such as MedBench v4 and MedHELM include multimodal, agentic, workflow, and clinician-validated tracks [Ding et al. 2025, Bedi et al. 2025]. Clinical-agent benchmarks instead evaluate sequential decisions, tool use, and workflow execution [Schmidgall et al. 2024, Jiang et al. 2025, Liu et al. 2026, Lee et al. 2025, Qiao et al. 2026b]. For consultation agents, this distinction changes the evaluation object: the inquiry policy helps construct the diagnostic context on which later decisions depend.

The closest interactive benchmarks make different parts of this process observable. AMIE evaluates a complete con-

versational inquiry–diagnosis system, while HealthBench and MedDialogRubrics score outputs against physician-authored criteria or consultation rubrics [Tu et al. 2025, Arora et al. 2025, Gong et al. 2026]. MedConsultBench explicitly evaluates information acquisition across the consultation cycle [Qiao et al. 2026a]; ThReadMed-QA, CP-Bench, and MeDxAgent broaden the setting to patient question threads, challenging patient behavior, and multi-agent consultation [Munnangi and Savage 2026, Li et al. 2026b, Sanghvi et al. 2026]. These benchmarks establish inquiry quality as a capability, but an end-to-end diagnosis score still combines the elicited history with the system's terminal generator.

MedDDC-Eval asks a complementary attribution question: how well does a policy-elicited history support diagnosis when every history is interpreted by the same diagnostic reader? DDXPlus also studies evidence acquisition in a controlled differential-diagnosis simulator [Tchango et al. 2022]; our setting targets open-ended consultation histories and compares them under a frozen shared diagnostic reader. This control holds terminal interpretation fixed across policies, while coupled evaluation remains necessary for judging the complete inquiry–diagnosis pipeline.

Multi-Turn RL for LLM Agents

GRPO is a critic-free grouped-reward method introduced in DeepSeekMath [Shao et al. 2024]. Related work studies reproducible and scalable LLM reinforcement learning, as well as finer-grained credit assignment for multi-turn agents [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]. Doctor-R1 combines process and terminal rewards to train a medical inquiry agent [Lai et al. 2025]. Our evaluation-informed training study instead asks whether a separate reward organized around diagnosis and trajectory feedback improves the history elicited by a fixed Qwen3-32B policy family [Yang et al. 2025]. Interaction spans multiple turns, but reward normalization and credit assignment remain trajectory-level under standard GRPO.

MedDDC-Eval: Diagnosis-Decoupled Evaluation

MedDDC-Eval combines a grounded consultation interface, a frozen shared diagnostic reader, and a fixed D/T/E evaluator. Together they turn free-form policy-elicited histories into controlled comparison objects.

Held-Out Sources and Reference Targets

The testbed uses two Chinese-language data sources: expert-labeled hospital records and online multi-turn consultations. The record pipeline screens 3,616 records spanning 153 source department labels into 2,904 annotated candidates across 139 normalized secondary departments. From this pool, 912 cases support RL training. Held-out evaluation is case-disjoint from training and contains a 100-case **Record split** sampled with departmental stratification and a 70-case **Dialogue split**. Dialogue cases are retained only when the primary intent is diagnostic and the source supports both diagnosis labeling and reference-trajectory construction. The Dialogue split tests whether findings transfer from record-derived cases to cases derived from online consultations. Both splits are converted into simulated consultations; the evaluated doctor never receives the retained source dialogue.

Each case is $c = (s, x, F, G, Q, b)$: data source s, initial context x, de-identified simulator facts F, reference diagnoses G, trajectory targets Q, and turn budget b. A target pairs a reference question with the information it should elicit. Annotators rate candidate targets on a three-level importance scale; only highest-importance targets enter the frozen Q and contribute equally to TP/FP/FN. The target set represents normalized information needs rather than canonical wording or question order, and compatible subquestions may remain bundled as one inquiry unit. Volunteered target facts are excluded from that dialogue's trajectory denominator.

Construction, Governance, and Simulation

Fifteen medical annotators construct G and Q across four batches. The workflow proceeds from case-grounded drafting to medical normalization and importance rating, followed by review for clinical relevance, redundancy, and semantic consistency; disagreements are adjudicated before freezing. Case-level identifiers are deduplicated before split assignment, and training and evaluation are identifier-disjoint. The supplementary material reports the complete construction funnel, frozen inventory, release tiers, and validation evidence.

The grounded patient simulator answers from F: record cases begin from a structured complaint, whereas dialogue-derived cases use an extracted queue of retained facts. The doctor asks questions or emits a structured diagnosis action, and the rollout stops at that action or the turn budget; the main training setup permits at most seven consultation turns. This interface controls information access while allowing policies to choose different questions, orders, and stopping points.

Shared Diagnostic Reader

Each completed history h_π is passed to the same frozen prompt-based DS-V3 diagnostic reader [DeepSeek-AI 2025] r_{shared} with a common prompt and decoding interface. The resulting score measures diagnostic support in h_π under a standardized history-to-diagnosis mapping. The fixed-history intervention replaces each policy's own terminal diagnosis generator with this reader. Policies are compared through elicited histories while retaining different questions and stopping points; coupled evaluation remains complementary when terminal generation is part of the target capability.

D/T/E Measurement and One-to-One Assignment

The evaluator receives the complete dialogue history, the shared diagnostic reader's output, and references (G, Q). It extracts diagnostic predictions and pre-summary doctor-question units, then filters trajectory targets already volunteered by the patient. Figure 2 summarizes the pipeline: LLM-

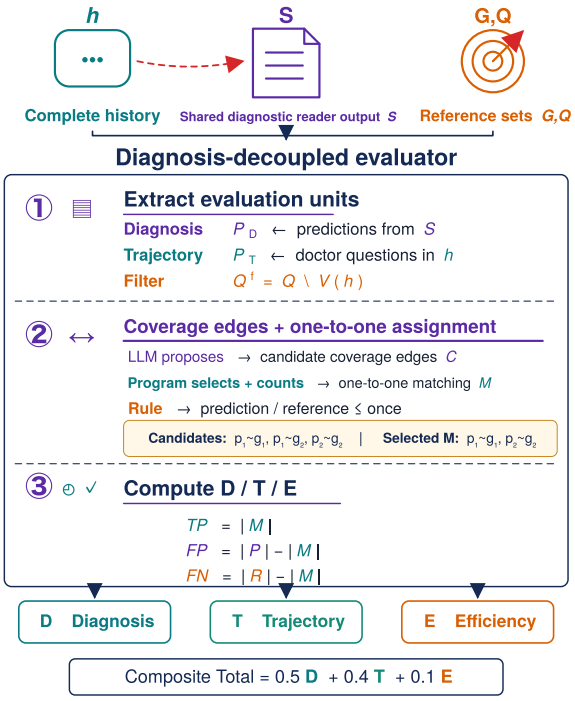

Figure 2: MedDDC-Eval maps each history h to diagnosis output S with a frozen shared diagnostic reader and extracts diagnosis predictions and pre-summary questions. LLM judges propose semantic matches to diagnosis references G and trajectory targets Q; deterministic one-to-one matching yields TP/FP/FN for D and T, while E captures when matched targets are covered. D/T/E are reported separately; Total gives only aggregate ordering.

assisted judges propose candidate semantic matches under the frozen directional-coverage rules, while a deterministic program selects one-to-one matches and computes the reported scores.

For predictions $P = \{p_i\}_{i=1}^{m}$ and references $R = \{r_j\}_{j=1}^{n}$, the semantic judge proposes a protocol-admissible candidate set $C \subseteq P \times R$. An edge records directional coverage under prespecified diagnosis-granularity or inquiry-type rules. A maximum-cardinality bipartite matching $M \subseteq C$ gives

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

followed by standard precision, recall, and F1. Each prediction and reference receives at most one credited match, so repeated paraphrases cannot increase TP beyond the number of distinct nodes. Candidate generation handles open-ended paraphrases and protocol-specified granularity relations, while deterministic assignment makes counting reproducible. This decomposition separates two validity questions: whether credited edges satisfy the protocol and whether candidate generation covers all scoring-relevant admissible relations. The saved candidate correspondences and their selected, excluded, and unmapped outcomes support assignment-level audit.

Diagnosis D is run-level micro F1 between predictions from the shared diagnostic reader and G; it measures the diagnostic support encoded in the elicited history. **Trajectory** T is run-level micro F1 between pre-summary doctor-question units and the filtered targets Q; it measures information-acquisition coverage against the reviewed reference-question targets while allowing flexible wording and question order. Bundling preserves clinically natural inquiry units, with one selected prediction and one selected target consuming one matching capacity. **Efficiency** $E \in [0.1, 1]$ summarizes how early matched trajectory targets are covered and penalizes ineffective turns. Its full timing map appears in the supplementary material. The prespecified aggregate

$$\mathrm{Total} = 0.5D + 0.4T + 0.1E \tag{2}$$

provides a joint system-level ordering, while D/T/E remain visible for component-level interpretation. Tables report all four quantities on a 0–100 scale. The 100 and 70 held-out cases are the statistical units, and matching-operation counts serve as reproducibility audits. All experiments use the frozen splits, simulator, prompts, reader, matching protocol, and weights of MedDDC-Eval v1.0.

Evaluation-Informed Training with Standard GRPO

We test whether MedDDC-Eval's diagnosis and trajectory dimensions can inform policy development. Qwen3-32B [Yang et al. 2025] is post-trained on cases disjoint from held-out evaluation using a separate training-time reward; the evaluator is frozen beforehand, so the study measures downstream utility under a fixed contract.

Interactive GRPO

Standard GRPO is applied to complete doctor–patient trajectories generated with the grounded simulator. Rollouts stop at a diagnosis action or the seven-turn limit; eight are sampled per prompt and rewards are normalized within the group. With SLIME [Zhu et al. 2025], the normalized sequence reward is broadcast across policy-generated tokens. Interaction is multi-turn, while feedback and credit assignment remain trajectory-level.

Feedback and Controls

The reward combines diagnosis and trajectory feedback with auxiliary behavioral constraints. A training-side terminal diagnosis generator scores whether the completed history supports the target, while trajectory feedback scores annotated-target coverage. Turn, format, repetition, question-count, and cross-turn prefix-similarity terms promote valid rollouts.

Training-side generation, extraction, and semantic judging use DeepSeek-v3.2; held-out evaluation uses the frozen DS-V3 shared diagnostic reader [DeepSeek-AI 2025] and offline $0.5D + 0.4T + 0.1E$ evaluator. Behavioral terms absent from the held-out score preserve separation between optimization feedback and reported measurement; coefficients and prompts appear in the supplementary material.

Ablations remove either feedback signal with the simulator, auxiliary constraints, grouped-rollout procedure, and held-out evaluator fixed.

Experiments

Experimental Logic and Common Evaluation Setting

The experiments proceed in three stages. We first isolate whether the terminal diagnosis component changes comparisons among fixed consultation histories. We then test whether the evaluation dimensions support controlled policy development and whether diagnosis and trajectory feedback make distinct contributions. Finally, we characterize case-level heterogeneity and sensitivity to rollout repeats, score weights, and alternative readers.

The main comparison contains the trained policy, its Qwen3-32B initialization, and six external consultation agents listed in Table 1 [Yang et al. 2025, OpenAI 2026, GLM-5 Team 2026, Chen et al. 2024, Zhang et al. 2024, Baichuan-M2 Team et al. 2025, Lai et al. 2025]. Every agent uses the same Chinese-language simulator, source-specific turn budget, frozen DS-V3 shared diagnostic reader, extraction and matching protocol, D/T/E weights, and five-run reporting procedure. We use the trained–base pair as the controlled training contrast; external systems provide reference context under the common evaluation protocol. Repeats sample new rollouts from fixed checkpoints, while confidence intervals resample held-out cases as clusters and recompute the run-level estimators.

Reader Control Exposes Attribution Confounding

The terminal diagnosis component materially changes between-policy comparisons even with fixed consultation histories. Figure 1 holds each history, diagnostic prompt, and decoding setting fixed while replacing each policy's own terminal diagnosis generator with the shared diagnostic reader. Diagnosis F1 shifts by 2.2–19.0 points, reversing 5/28 (18%) Record and 10/28 (36%) Dialogue pairwise orderings in the complete eight-policy audit. Questions, patient answers, trajectory, and efficiency remain fixed, so the heterogeneous shifts and reversals identify terminal diagnosis generation as the source of the comparison change and motivate use of the frozen shared diagnostic reader.

Diagnosis and trajectory are related across systems but are not interchangeable for individual consultations. Across eight system means, Pearson r is 0.566 on Record and 0.303 on Dialogue; across case–model means it falls to 0.147 and 0.101. High-trajectory/low-diagnosis and low-trajectory/high-diagnosis cases both remain common. The axes can improve together across systems, but neither reliably substitutes for the other on individual consultations; D/T component profiles therefore carry information hidden by Total.

Evaluation-Informed Training Improves Elicited Histories

A training reward informed by the diagnosis and trajectory dimensions improves both diagnostic support and information-acquisition coverage. In the controlled trained–base comparison, the architecture family and initialization are shared, so the contrast isolates the effect of post-training under the reported setup. The six external consultation agents in Table 1 provide descriptive context under the common evaluation protocol.

Relative to its initialization, the trained policy improves Record D/T/E/Total by 11.5/9.1/2.7/9.6 points and Dialogue by 2.0/7.9/4.4/4.6 points. The trained policy has the highest Total in the eight-system comparison, while GPT-5.4 retains the strongest Record trajectory score by 1.6 points. This contrast demonstrates why component profiles complement the aggregate ranking.

These profiles change how benchmark results should be used. Diagnosis F1 is the direct comparison when downstream diagnostic support is the priority, whereas trajectory F1 indicates whether the policy elicited the reviewed information needs. Efficiency is most informative after D and T reach an acceptable level: a fast but incomplete consultation can otherwise appear favorable. Models with similar Total can therefore suit different uses, and reporting only the aggregate would hide trade-offs that matter for model selection.

A case-cluster bootstrap over the cases available for each comparison supports the headline mean gains. Each resample re-aggregates TP/pred/gold within every run, recomputes micro D/T, case-mean E, and Total, and then averages the five repeats. The resulting trained–base Total differences are 9.6 points on Record (95% CI [7.6, 11.8]) and 4.6 on Dialogue ([2.4, 6.8]); differences from GPT-5.4 are 3.8 ([1.4, 6.3]) and 3.6 ([1.0, 6.2]). These intervals quantify held-out case-sampling uncertainty for the evaluated checkpoints; training-seed variation lies outside this fixed-checkpoint analysis.

Case-Level Gains Are Heterogeneous

Aggregate gains conceal heterogeneous case-level improvement modes. Figure 3 decomposes the trained–base effect over 170 complete case pairs, using differences between five-run case means and the cross-run stability criteria in the caption. Sixty-two cases (36.5%) meet the S2 stable trajectory-gain criterion; the annotated abdominal-pain case lies near this bucket's center with trajectory $+22.5$ points and diagnosis -0.7. Seven cases (4.1%) meet the S3 diagnosis-gain criterion with recurrently sparse absolute trajectory coverage; the annotated jaundice-like case gains 43.3 diagnosis points while trajectory changes by -1.2.

The two modes also support targeted case analysis. High-trajectory/low-diagnosis cases identify histories that cover reviewed questions yet provide limited diagnostic support under the shared diagnostic reader. Low-trajectory/high-diagnosis cases show that salient clues can support a diagnosis despite limited information-acquisition coverage. These buckets can prioritize case review and indicate whether future reward revisions should emphasize target coverage or diagnostic sup-

Agent / policy	Record-driven (100 cases)				Dialogue-driven (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**	57.2	**82.9**	**53.4**	**41.8**	**52.6**	**82.4**	**50.2**	**+9.6 / +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 1: Main results under the shared diagnostic reader protocol. Values are percentage means over five fixed-checkpoint rollout repeats; within each repeat, D and T use micro-aggregated one-to-one counts and E uses the first summary round. Qwen3-32B + GRPO versus Qwen3-32B is the controlled training comparison; external agents provide descriptive context. Total is $0.5D + 0.4T + 0.1E$, and Δ is computed from the displayed one-decimal Total values for Record / Dialogue. Sample standard deviations appear in the supplementary material.

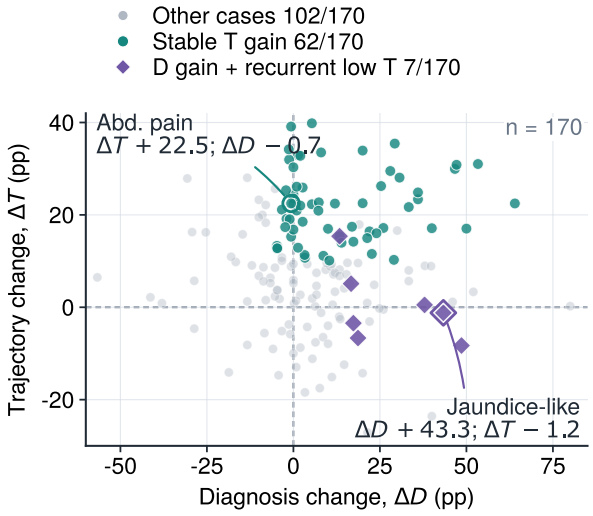

Figure 3: Case-level trained–base changes under the shared diagnostic reader ($n = 170$; five-run means). S is the fraction of the 25 trained–base cross-run pairs in which the trained policy scores higher. S2 (teal; 62/170) requires $\Delta T \geq 10$, $S_T \geq 0.70$, and $\Delta D \geq -5$. S3 (violet; 7/170) requires $\Delta D \geq 10$, $S_D \geq 0.70$, and split-specific bottom-T/top-D membership in at least 3/5 trained-policy runs. For S3, low T denotes an absolute level, not $\Delta T < 0$. Patterns are predefined, non-exhaustive, and may overlap.

port; their role is to localize observed behavior for follow-up analysis.

Both Feedback Signals Contribute

Both feedback signals provide non-redundant gains in this controlled Qwen3-32B setting. Removing diagnosis feedback lowers Total by 4.9/3.6 points on Record/Dialogue, while removing trajectory feedback lowers it by 4.1/4.2 points. Case-cluster bootstrap confidence intervals for all four differences

Setting	D	T	E	Total
Record-driven				
Full reward	**44.6**	**57.2**	**82.9**	**53.4**
w/o diagnosis feedback	36.4	53.2	90.8	48.5
w/o trajectory feedback	38.8	55.4	78.1	49.4
Dialogue-driven				
Full reward	**41.8**	**52.6**	**82.4**	**50.2**
w/o diagnosis feedback	35.3	50.1	89.0	46.6
w/o trajectory feedback	36.3	50.3	77.0	46.0

Table 2: Reward-component ablations for Qwen3-32B policies trained with GRPO. Values are percentage means over five fixed-checkpoint rollout repeats; the full-reward rows match the trained policy in Table 1. Ablations change only the training reward, while the held-out evaluator and $0.5D + 0.4T + 0.1E$ score remain fixed.

remain above zero. With the simulator, auxiliary constraints, grouped-rollout procedure, and held-out evaluator fixed, the removals are directly comparable. The no-diagnosis variant nevertheless increases E while reducing D and T, showing why efficiency must be interpreted jointly.

Evaluator Stability Under the Fixed Protocol

The main ordering is stable across the tested evaluator perturbations, although separation narrows in the most sensitive setting. Across the 16 model–split summaries, the median Total SD over five rollout repeats is 1.41 points, and the top mean ordering is preserved across all seven tested weight settings and both alternative readers. Under the GPT-5.4 Dialogue reader, however, the leading margin narrows to 0.44 points; this result supports ordering stability in that setting, but not decisive pairwise separation.

These checks characterize sensitivity to rollout randomness, aggregate weights, and the history-to-diagnosis reader. They do not establish prospective clinical validity or exhaustive coverage of all admissible semantic matches; those questions remain outside this offline benchmark.

Limitations

MedDDC-Eval evaluates simulated Chinese-language consultations rather than clinical deployment. Grounded simulators provide controlled, case-consistent interactions but do not capture the full diversity of patient behavior, and valid consultations may differ from the reviewed reference trajectories in ordering and phrasing.

Results are conditional on the frozen shared diagnostic reader and candidate-generation protocol. Broader clinical and cross-setting validation, as well as independent training-seed studies, remain future work. The reward-ablation study evaluates one Qwen3-32B training configuration with repeated rollouts from fixed checkpoints; its conclusions therefore apply to the reported controlled setting.

Ethical Statement

This work uses de-identified source-derived cases to drive simulated consultations for offline evaluation. Experimental inputs contain no direct personal identifiers; raw records, original dialogues, linkable identifiers, and case-level traces are not released, as detailed in the supplementary data-construction and validation appendix. No evaluated model is intended for patient-facing or autonomous clinical use, and automatic scores are not safety certification.

Planned public artifacts are limited to reviewed prompts, schema, scoring code, aggregate results, and synthetic examples. Source-derived subsets remain conditional on authorization, licensing, institutional review, and de-identification; these constraints take precedence over unrestricted release.

Conclusion

MedDDC-Eval reframes the evaluation of multi-turn medical consultation agents around the histories that policies elicit. A frozen shared diagnostic reader holds the history-to-diagnosis mapping constant across policies, while one-to-one D/T/E measurement characterizes diagnostic support, information-acquisition coverage, and efficiency. Fixed-history interventions reveal score shifts and ranking reversals when policy-specific terminal diagnosis generators are replaced, and the controlled GRPO study shows gains in held-out elicited-history measurements using a separate reward organized around diagnosis and trajectory feedback. Together, these results position diagnosis-decoupled evaluation as a complement to end-to-end evaluation: the former compares elicited histories under a common reader, whereas the latter assesses the complete system, including terminal diagnosis generation. Future work should examine broader reader ensembles, adaptive diagnosis–trajectory feedback, and additional language and evaluation settings.

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.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2025. doi: 10.48550/arXiv.2412.19437. URL https://arxiv.org/abs/2412.19437.

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.

GLM-5 Team. Glm-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. doi: 10.48550/arXiv.2602.15763. URL https://arxiv.org/abs/2602.15763.

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*, 2025. 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. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6227–6243, Rabat, Morocco, 2026a. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.328. URL https://aclanthology.org/2026.findings-eacl.328/.

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.

OpenAI. Gpt-5.4 model documentation. https://platform.openai.com/docs/models/gpt-5.4, 2026. Accessed 2026-07-25.

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.

A. Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

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 37*, pages 26045–26081. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-0819. URL https://doi.org/10.52202/079017-0819.

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, shared 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 support the evaluation-informed training study and 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. Semantic near-duplicate screening across the heterogeneous source text remains a follow-up to the completed identifier-level deduplication.

A.2 Case Schema and Reference Construction

The common case object contains a data-source label, 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 semantic matching under directional coverage rules separates information-acquisition coverage 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 targets 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. This workflow produces a reviewed consensus reference rather than a matrix of independent parallel labels; construction validity is therefore assessed through staged medical review.

The staged review above provides construction validity by defining trajectory targets as reviewed normative information needs while preserving flexibility in consultation sequence. Scoring integrity is addressed through the evaluator contract and sensitivity analyses below.

(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.

Case field	Construction meaning	Simulation, scoring, and release role
Data source s	Record-derived or dialogue-derived provenance.	Defines the evaluation split; only aggregate source 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 semantic matching under directional coverage and 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.

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 candidate semantic matches from the LLM-assisted judge under the protocol's directional coverage rules, 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.

The saved candidate correspondences record the candidate, selected, excluded, and unmapped outcomes used for counting. The proposal set follows the frozen protocol, and MedDDC-Eval v1.0 micro-aggregates the resulting counts within each run before reporting the mean and sample SD across five fixed-checkpoint rollout repeats. The held-out shared diagnostic reader, extractor, and candidate-match judge use DeepSeek-v3 through frozen role-specific gateways; the training-side terminal diagnosis generator, extractor, and judge 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 $\lambda = 0.15$,

$$w_i = \exp[-\lambda(t_i - 1)], \qquad \bar{t} = \frac{\sum_i t_i w_i}{\sum_i w_i}. \tag{3}$$

The timing map is

$$\mathrm{timing}(\bar{t}) = \begin{cases} 1.0, & \bar{t} \le 2, \\ 0.95 - 0.05(\bar{t} - 2), & 2 < \bar{t} \le 3, \\ 0.90 - 0.15(\bar{t} - 3), & 3 < \bar{t} \le 5, \\ 0.60 - 0.10(\bar{t} - 5), & 5 < \bar{t} \le 8, \\ \max(0.1, 0.30 - 0.02(\bar{t} - 8)), & \bar{t} > 8. \end{cases} \tag{4}$$

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.

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

$$r_{\text{ineff}} = \max\left(0, \frac{T-m}{T}\right), \tag{5}$$

$$E = \min\left(1, \max\left(0.1, \text{timing}(\bar{t})\left(1 - 0.3 r_{\text{ineff}}\right)\right)\right). \tag{6}$$

Appendix C
Fixed-History Diagnostic-Reader Intervention

Supplementary Table S4 provides the complete eight-policy fixed-history audit underlying the main-paper intervention figure. Each rollout history and the diagnostic prompt are held fixed; only the terminal diagnosis component changes from each policy's own terminal diagnosis generator to the frozen DS-V3 shared diagnostic reader. The table retains all exact diagnosis-F1 endpoints and rank changes, including the trained policy excluded from the figure's seven-policy aggregate braid.

Policy rollout	Record diagnosis F1 Own generator → shared diagnostic reader	Record D rank Own generator → shared diagnostic reader	Dialogue diagnosis F1 Own generator → shared diagnostic reader	Dialogue D rank Own generator → shared diagnostic reader
Qwen3-32B + GRPO	40.4 → 44.6 (+4.1)	1 → 1	36.5 → 41.8 (+5.3)	1 → 1
Qwen3-32B	26.5 → 33.1 (+6.6)	6 → 5	29.6 → 39.8 (+10.2)	5 → 2
GPT-5.4	33.2 → 36.3 (+3.1)	3 → 4	30.9 → 35.5 (+4.6)	3 → 6
GLM-5	33.9 → 38.0 (+4.1)	2 → 2	30.5 → 34.3 (+3.8)	4 → 7
HuatuoGPT-o1-70B	33.0 → 37.7 (+4.7)	4 → 3	32.4 → 37.2 (+4.7)	2 → 5
Llama-3-70B-UltraMedical	29.2 → 31.4 (+2.2)	5 → 8	20.8 → 39.5 (+18.6)	7 → 3
Baichuan-M2-32B	16.6 → 32.3 (+15.8)	8 → 6	14.5 → 33.5 (+19.0)	8 → 8
Doctor-R1	24.9 → 31.5 (+6.6)	7 → 7	27.8 → 37.4 (+9.6)	6 → 4

Table S4: Fixed-history terminal-diagnosis intervention. The own-generator condition applies each policy's own terminal diagnosis generator under a common diagnostic prompt; the shared-reader condition applies the frozen DS-V3 diagnostic reader to the identical history. Questions, patient answers, trajectory score, and efficiency are held fixed. The heterogeneous diagnosis-F1 shifts and rank changes show that the terminal diagnosis component materially affects diagnosis-based comparisons of elicited histories.

Appendix D
Compact Measurement Summaries

Tables S5–S6 summarize the compact inventory and measurement-contract details that support the main-paper protocol description.

Component	Role	Inventory / scope	Boundary
Grounded consultation interface	Produces bounded policy-elicited histories from two data sources.	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 diagnostic reader	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 conditional on the reader.
Auditable D/T/E harness	Measures diagnostic support, information-acquisition coverage, and efficiency.	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 S5: 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.

M	Question answered	Evaluation unit	Aggregation	Main boundary
D	Does the elicited history support reference diagnoses under the shared diagnostic reader?	Diagnosis item	Run-level micro F1	Conditional on the frozen reader and proposed semantic edges.
T	Does the consultation cover the reviewed reference-question targets?	Doctor question–target unit	Run-level micro F1 before first summary	Bundled-target resolution; volunteered facts are filtered.
E	How early are matched trajectory targets covered?	Matched question turn	Deterministic timing score	Coupled to targets matched for 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 S6: Self-contained D/T/E measurement contract. Semantic models propose protocol-admissible candidate edges; deterministic one-to-one assignment selects and counts them.

Appendix E
Evaluation-Informed Training Details

Tables S7–S8 and Figure S1 specify the rollout, reward, and update controls used for the evaluation-informed training study.

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 feedback and trajectory feedback
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 S7: Process-aware GRPO training configuration.

E.1 Rollout, Reward, and Ablation Controls

Table S8 summarizes the interaction unit, termination rule, grouped sampling, reward timing, credit scope, optimizer, and train–evaluation separation used in the evaluation-informed training study.

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 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 S8: 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.

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

$$\begin{aligned} R = R_{\mathrm{diag}} + R_{\mathrm{traj}} + \lambda_{\mathrm{turn}} R_{\mathrm{turn}} + \lambda_{\mathrm{format}} R_{\mathrm{format}} \\ + \lambda_{\mathrm{rep}} R_{\mathrm{rep}} + R_{\mathrm{qcount}} + \lambda_{\mathrm{prefix}} R_{\mathrm{prefix}}. \end{aligned} \tag{7}$$

This separate training-time reward targets the diagnosis and trajectory dimensions but is distinct from the held-out $0.5D + 0.4T + 0.1E$ aggregate Total formula. It also includes behavioral constraints that are absent from the held-out score. Reward ablations remove only diagnosis feedback or trajectory feedback while preserving the simulator, auxiliary constraints, frozen shared diagnostic reader, held-out evaluator, and score coefficients.

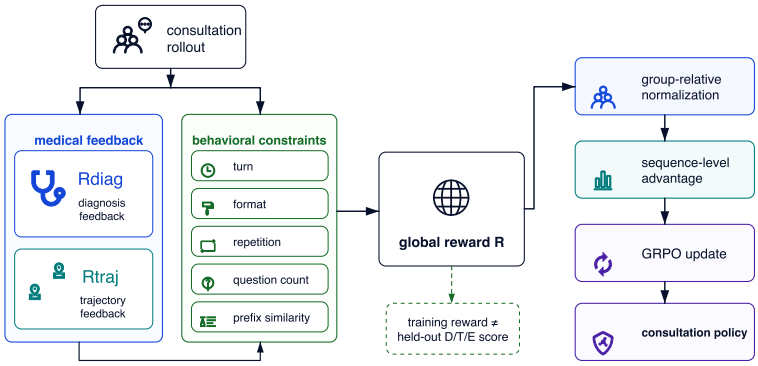

Figure S1: Process-aware reward and standard GRPO update. Diagnosis feedback, trajectory feedback, and behavioral constraints form the training reward. Group-relative normalization yields a sequence-level advantage for the policy update; this reward is distinct from the held-out D/T/E score.

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 S9: 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.

Appendix F
Main Result Uncertainty and Robustness

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 yields information-acquisition coverage

Table S10: Baseline protocol and interpretation. Trained same-family comparisons are separated from external references so closed-model comparisons are not treated as optimizer ablations.

F.1 Case-Cluster Uncertainty and Weight Sensitivity

The case-cluster bootstrap resamples the held-out cases available for each comparison with replacement and applies the same draw to every policy and repeated run. Within each draw, it re-aggregates TP/pred/gold to recompute run-level micro D/T, case-mean E, and Total before averaging the five repeats. The resulting point estimates match the main-table contrasts. The 10,000 resamples quantify held-out case-sampling uncertainty for fixed trained checkpoints; training-seed variability lies outside this analysis.

Agent / policy	Diagnosis F1 (%)	Trajectory F1 (%)	Efficiency (%)	Total (%)
Qwen3-32B + GRPO	**44.6**±3.0	57.2±1.7	**82.9**±1.0	**53.4**±1.3
GPT-5.4	36.3±2.3	**58.8**±0.7	79.7±0.7	49.7±1.3
GLM-5	38.0±7.7	54.4±1.6	66.4±16.6	47.4±5.4
HuatuoGPT-o1-70B	37.7±1.5	46.2±2.0	77.9±0.5	45.1±1.1
Llama-3-70B-UltraMedical	31.4±2.4	50.5±1.7	81.2±1.0	44.0±1.0
Qwen3-32B	33.1±0.9	48.1±1.3	80.2±1.1	43.8±1.0
Baichuan-M2-32B	32.3±3.2	46.3±2.9	81.8±1.2	42.9±2.1
Doctor-R1	31.5±1.8	40.3±2.7	75.0±2.1	39.3±1.5

Table S11: Component results on 100 record-driven held-out cases, ordered by total score. Values are mean ± sample standard deviation over five repeated runs. Diagnosis and trajectory F1 use strict one-to-one semantic 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**±2.1	**52.6**±1.1	**82.4**±1.2	**50.2**±1.3
GPT-5.4	35.5±2.2	52.4±0.7	79.1±0.3	46.6±1.0
Llama-3-70B-UltraMedical	39.5±3.6	46.9±2.2	79.4±1.5	46.4±2.5
Qwen3-32B	39.8±1.4	44.7±1.9	78.0±2.0	45.6±1.5
Doctor-R1	37.4±2.4	41.2±2.6	78.6±3.4	43.1±1.5
GLM-5	34.3±4.2	44.7±1.7	74.9±2.0	42.5±2.2
Baichuan-M2-32B	33.5±2.8	44.2±1.2	80.8±2.2	42.5±1.3
HuatuoGPT-o1-70B	37.2±2.8	40.8±2.6	75.9±2.3	42.5±2.2

Table S12: Component results on 70 dialogue-driven held-out cases, ordered by total score. Values are mean ± sample standard deviation over five repeated runs. Diagnosis and trajectory F1 use strict one-to-one semantic 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	HuatuoGPT-o1-70B	70	7.7	[5.2, 10.0]	>0.9999	47/7/16
Dialogue	Baichuan-M2-32B	70	7.7	[5.2, 10.2]	>0.9999	58/2/10
Dialogue	GLM-5	70	7.6	[5.3, 10.0]	>0.9999	57/4/9
Dialogue	Doctor-R1	70	7.1	[4.8, 9.4]	>0.9999	53/2/15
Dialogue	Qwen3-32B	70	4.6	[2.4, 6.8]	>0.9999	50/5/15
Dialogue	Llama-3-70B-UltraMedical	70	3.7	[1.4, 6.1]	0.9993	43/4/23
Dialogue	GPT-5.4	70	3.6	[1.0, 6.2]	0.9973	41/4/25
Record	Doctor-R1	100	14.1	[11.7, 16.6]	>0.9999	88/4/8
Record	Baichuan-M2-32B	100	10.6	[8.1, 13.1]	>0.9999	72/8/20
Record	Qwen3-32B	100	9.6	[7.6, 11.8]	>0.9999	76/7/17
Record	Llama-3-70B-UltraMedical	100	9.4	[7.0, 11.8]	>0.9999	77/8/15
Record	HuatuoGPT-o1-70B	100	8.3	[5.9, 11.0]	>0.9999	65/8/27
Record	GLM-5	100	6.0	[3.7, 8.5]	>0.9999	77/4/19
Record	GPT-5.4	100	3.8	[1.4, 6.3]	0.9995	54/7/39

Table S13: Case-cluster bootstrap over the cases available for each trained-policy comparison. Each resample draws held-out cases and recomputes run-level micro D/T, case-mean E, and Total before averaging the five repeats, so point differences match the main table. Differences and confidence intervals are percentage points from 10,000 resamples. W/T/L is a descriptive count from available repeated-run case means with a one-point tie tolerance.

Split	Baseline	Component	Mean diff.	95% CI	$P(\Delta > 0)$
Dialogue	GPT-5.4	Diagnosis	6.3	[1.6, 11.1]	0.9957
Dialogue	GPT-5.4	Efficiency	3.3	[2.1, 4.4]	>0.9999
Dialogue	GPT-5.4	Trajectory	0.2	[-3.0, 3.3]	0.5390
Dialogue	Qwen3-32B	Diagnosis	2.0	[-1.9, 5.8]	0.8457
Dialogue	Qwen3-32B	Efficiency	4.4	[2.1, 6.9]	>0.9999
Dialogue	Qwen3-32B	Trajectory	7.9	[5.0, 10.8]	>0.9999
Record	GPT-5.4	Diagnosis	8.2	[4.0, 12.6]	0.9999
Record	GPT-5.4	Efficiency	3.2	[1.8, 4.6]	>0.9999
Record	GPT-5.4	Trajectory	-1.7	[-4.5, 1.2]	0.1292
Record	Qwen3-32B	Diagnosis	11.5	[8.2, 15.0]	>0.9999
Record	Qwen3-32B	Efficiency	2.7	[1.0, 4.3]	0.9990
Record	Qwen3-32B	Trajectory	9.1	[6.5, 11.8]	>0.9999

Table S14: Case-cluster bootstrap over the cases available for the controlled Qwen3-32B comparison and strongest external total-score reference. Each draw recomputes the run-level component estimators before averaging repeats. 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	3.76	Qwen3-32B + GRPO	3.56
0.55 / 0.35 / 0.10	Qwen3-32B + GRPO	4.26	Qwen3-32B + GRPO	3.58
0.45 / 0.45 / 0.10	Qwen3-32B + GRPO	3.27	Qwen3-32B + GRPO	3.25
0.45 / 0.40 / 0.15	Qwen3-32B + GRPO	3.51	Qwen3-32B + GRPO	3.41
0.50 / 0.35 / 0.15	Qwen3-32B + GRPO	4.01	Qwen3-32B + GRPO	3.61
0.55 / 0.40 / 0.05	Qwen3-32B + GRPO	4.02	Qwen3-32B + GRPO	3.71
0.50 / 0.45 / 0.05	Qwen3-32B + GRPO	3.52	Qwen3-32B + GRPO	3.41

Table S15: 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.

F.2 Alternative Readers

Table S16 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. The closest setting supports ordering stability, with a 0.44-point margin that leaves pairwise separation unresolved.

Policy	Record total (%)		Dialogue total (%)	
	Qwen3-32B reader	GPT-5.4 reader	Qwen3-32B reader	GPT-5.4 reader
Qwen3-32B + GRPO	**49.9**$\pm$**2.1**	**50.4**$\pm$**0.5**	**46.0**$\pm$**0.9**	**44.8**$\pm$**1.2**
Qwen3-32B	40.5$\pm$1.3	42.8$\pm$0.8	40.5$\pm$1.0	40.0$\pm$1.5
GPT-5.4	47.4$\pm$1.8	48.1$\pm$1.2	44.1$\pm$2.3	44.3$\pm$0.7
GLM-5	44.5$\pm$4.3	45.6$\pm$3.7	40.0$\pm$1.2	40.6$\pm$1.4
HuatuoGPT-o1-70B	41.9$\pm$0.7	42.3$\pm$1.0	38.7$\pm$3.4	39.3$\pm$2.4
Llama-3-70B-UltraMedical	43.9$\pm$0.7	45.1$\pm$1.0	41.1$\pm$1.4	42.8$\pm$1.3
Baichuan-M2-32B	38.3$\pm$2.1	38.8$\pm$1.8	37.6$\pm$0.8	39.1$\pm$0.4
Doctor-R1	37.8$\pm$1.5	37.2$\pm$1.0	37.7$\pm$2.4	40.3$\pm$2.4

Table S16: Alternative shared diagnostic reader analysis on the same consultation histories. Values are percentage mean $\pm$ sample standard deviation over five runs. Only the frozen post-consultation diagnostic reader changes; trajectory, efficiency, and score weights remain fixed. The trained Qwen3-32B + GRPO policy has the highest total under both readers on both sources.

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.

Level	Split	n	Pearson r	Spearman ρ	Off-diagonal HT/LD; LT/HD
Model means	Record	8	0.566	0.238	–
Model means	Dialogue	8	0.303	0.381	–
Case–model means	Record	800	0.147	0.139	71; 76
Case–model means	Dialogue	560	0.101	0.086	50; 63

Table S17: Diagnosis–trajectory association under strict one-to-one scoring. 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 not interchangeable.

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	58.7	54.7	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 S18: 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 leaderboard.

Pattern	Bucket frequency	Case-derived setting	Evidence summary	Five-run change
Stable process improvement	62/170 (36.5%)	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	7/170 (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 S19: 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 H
Interpretive and Deployment Boundaries

The semantic judge and frozen shared diagnostic reader are model components of the measurement pipeline. Candidate semantic matches follow protocol-defined directional coverage relations, and D/T/E scores are interpreted under the frozen candidate-generation and matching protocol. Public prompts, matching code, and aggregate results make this dependency inspectable. Alternative shared diagnostic readers characterize dependence on the history-to-diagnosis model, while exhaustive candidate-generation coverage remains a separate measurement question.

Trajectory targets are reviewed normative information needs expressed as clinically natural question units. Related 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 design provides coarser within-bundle resolution than atomic-fact recall while allowing semantically compatible questions and removing facts already volunteered by the simulated patient. The targets cover reviewed information needs across flexible consultation strategies. Diagnostic support, information-acquisition coverage, and efficiency should therefore be read together, with efficiency measuring how early matched trajectory targets are covered.

These boundaries mean that D/T/E should be read as protocol-conditioned research metrics rather than clinical endpoints. Prospective clinical validity, broader reader ensembles, and deployment safety require study designs beyond this offline evaluation.