DiagBench: Evaluating Diagnostic Revision in Large Language Models with Longitudinal Inpatient Records

Huawei Feng[1] **Dongsheng Shi**[1] **Linlin Wang**[1,2]*
[1]East China Normal University [2]City University of Hong Kong
{hwfeng,dongsheng}@stu.ecnu.edu.cn llwang@cs.ecnu.edu.cn

Abstract

Reliable clinical decision support requires tracking evidence over a hospital stay and revising a disease diagnosis when new findings challenge the initial assessment. Strong performance on medical question answering does not establish that large language models (LLMs) can make these longitudinal decisions. We introduce DIAGBENCH,[1] a benchmark of 2,865 Chinese inpatient admissions that evaluates diagnostic revision as clinical evidence accumulates. It pairs admission-only prediction with staged evidence disclosure, distinguishing an unchanged diagnosis, refinement of an existing hypothesis, and replacement by a different diagnosis. Development and test cases have human-annotated reference labels. The best LLM configuration in the main evaluation achieves only 52.6% macro recall; 80.3–87.7% of errors across five DeepSeek-Chat (DeepSeek-V3.2, non-thinking mode) prompting strategies occur between adjacent revision categories. Strategies that improve final-diagnosis prediction do not consistently improve revision decisions. We further use DIAGAGENT to trace diagnoses and evidence across rounds, revealing failures to replace an initial hypothesis despite recognizing contradictory findings. On DeepSeek-Chat, a single agent with the same state outperforms the full system on macro recall.

1 Introduction

Inpatient diagnosis requires reasoning over evidence collected across days or weeks. Laboratory results, progress notes, and pathology may confirm an initial hypothesis, make it more specific, or reveal a different disease. A model supporting clinical decisions must retain relevant earlier findings and revise its working diagnosis when later evidence challenges the initial explanation. Failure to reconsider that explanation can contribute to diagnostic error (Graber et al., 2005). Although large language models (LLMs) perform increasingly well on medical question answering (Singhal et al., 2025; Wang et al., 2024), these results leave their ability to revise diagnoses over a hospital stay insufficiently assessed.

Consider two patients whose discharge diagnosis is lung cancer. One admission note already suspects a lung tumor; the other attributes the symptoms to a different disease. Later pathology confirms the existing suspicion in the first case but requires a change in the disease explanation in the second. An LLM must recognize this difference even when it predicts the same final diagnosis for both patients. This requires relating new findings to earlier suspicions and exclusions, then updating the diagnosis accordingly. Evaluating only the final disease name misses this essential aspect of longitudinal clinical reasoning.

Existing evaluations include knowledge benchmarks such as CMExam and CMB (Liu et al., 2023; Wang et al., 2024). Record-based benchmarks such as emrQA, CliMedBench, and ClinicBench evaluate questions and clinical decisions (Pampari et al., 2018; Ouyang et al., 2024; Liu et al., 2024). Temporal and interactive benchmarks assess broader workflows: TIMER studies longitudinal records (Cui et al., 2025); AgentClinic and MedAgentBench evaluate actions in clinical environments (Schmidgall et al., 2024; Jiang et al., 2025); and INA-Bench evaluates when to read another inpatient note before predicting later documented actions (Feng et al., 2026). These evaluations do not directly distinguish whether newly available evidence warrants preserving, refining, or replacing the initial diagnosis.

We introduce DIAGBENCH, built from 2,865 Chinese inpatient admissions from two tertiary teaching hospitals. It captures the admission-to-discharge diagnostic relationship through three ordered labels: NO REVISION, WEAK RE-

*Corresponding author.
[1]https://github.com/hwfengcs/DiagBench

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3821–3832
November 6-10, 2026 ©2026 Association for Computational Linguistics

FINEMENT, and STRONG REVISION. A separate SPECIAL/MANAGEMENT category covers management-focused admissions. Task 1 predicts the final diagnosis and revision label from admission information alone. Task 2 reveals first-course, progress, and, when available, postoperative notes in stages, requiring the model to maintain, refine, revise, or defer its judgment. The discharge record is withheld from model inputs. The tasks assess both initial predictions and subsequent updates.

Our contributions are threefold:

- We introduce DIAGBENCH with paired admission-only and staged-evidence tasks, independently annotated and adjudicated development and test labels, and separate scoring of diagnosis prediction and revision decisions.
- We evaluate prompting strategies and multiple model backbones. DeepSeek-Chat Self-Consistency achieves the highest macro recall in the main evaluation, at only 52.6%; 80.3–87.7% of errors across five DeepSeek-Chat strategies confuse adjacent categories. Stronger diagnosis prediction does not consistently improve revision decisions.
- We introduce DIAGAGENT to trace diagnoses and evidence across rounds, revealing failures to revise unsupported hypotheses despite recognizing new findings. On DeepSeek-Chat, Single-Agent State achieves 45.4% macro recall versus 37.7% for the full system.

2 The DIAGBENCH Benchmark

2.1 Data Sources and Selection

The source corpus contains 9,576 admissions from two tertiary teaching hospitals, organized into three data domains. Each benchmark case is one admission. Eight linked tables contain cover-sheet metadata, admission records, first-course notes (the initial inpatient assessment), daily progress notes, surgery records, postoperative notes, discharge records, and short-stay admission/discharge records. We link documents by admission identifier. Of these admissions, 3,907 have entries in all five required tables: cover sheet, admission, first-course, daily notes, and discharge.

We retain cases with an admission record of at least 500 characters, a stay of 2–60 days, extractable first-course diagnostic analysis, and an available weak revision label. The resulting 2,865 admissions have a median admission-note length of 835 characters and a median stay of 9 days. Before human annotation, we split them into 2,004 training, 286 development, and 575 test cases, stratified by domain and weak label. Every retained case has daily notes; 819 also have surgery and postoperative records. The online technical companion gives detailed source counts and domain distributions.

2.2 Diagnostic Revision Categories

The four labels describe how the final documented diagnosis relates to the initial assessment (Table 1). Annotators read the diagnostic analysis and later notes to decide whether the same explanation is retained, made more specific, or replaced. Labels depend on the record text, including what was suspected, what was ruled out, and when each finding became available.

For weak supervision (Ratner et al., 2017), we compare International Classification of Diseases (ICD) codes at admission and discharge: exact codes, their first three characters (ICD-3), and their leading letters, with a separate check for management codes. A code transition may reflect diagnostic confirmation, increased specificity, a different disease, or the purpose of admission. The clinical text determines which interpretation applies. An initial suspicion of malignancy may be confirmed by pathology, whereas a definite benign diagnosis may make later malignant pathology a substantial revision. We use the code-derived labels for training and human-reviewed labels for development and test.

For NoR versus WR, added detail must change the medical interpretation: laterality or wording alone may leave it unchanged, whereas pathological confirmation or a meaningful subtype refines an earlier hypothesis. Annotators compare explanations across notes, including suspicions absent from the admission diagnosis field.

2.3 Annotation and Adjudication

We developed the guideline through a 100-case internal pilot with AI-assisted screening and faculty review of 28 cases. The pilot clarified the distinction between diagnostic change and the purpose of admission; it is separate from the independent human annotation reported below.

Two medical-NLP researchers independently labeled every development and test case with a confidence level and rationale. The annotators and adjudicator were master's students or faculty advisors with medical-benchmark experience; none was a licensed physician. Primary annotators could

Label	Decision rule	Illustrative example
No Revision (NoR)	The main diagnosis is unchanged; wording or incidental detail may differ.	Type 2 diabetes → type 2 diabetes with poor glucose control.
Weak Refinement (WR)	Later evidence confirms a suspected diagnosis or makes it more specific.	Lung mass, suspected tumor → lung adenocarcinoma confirmed by pathology.
Strong Revision (SR)	The final diagnosis replaces the initial explanation.	Lumbar disc herniation → spinal tuberculosis.
Special/Management (S/M)	The code transition reflects a management-focused admission.	Admission for tumor chemotherapy → the known underlying malignancy.

Table 1: The four revision labels. Examples illustrate the annotation rules and are not verbatim patient records. S/M is analyzed separately from the ordered NoR–WR–SR categories.

	Train (weak)	Dev	Test
No Revision	645	93	202
Weak Refinement	732	84	267
Strong Revision	427	80	47
Special/Management	200	29	59
Total	2,004	286	575
Surgery records	568	73	178

Table 2: Split sizes. Development and test counts use the final human labels; training counts use ICD-derived labels.

read all notes, including discharge records and ICD fields, but not weak labels or each other's answers. The adjudicator also saw weak labels. We report raw agreement and Cohen's κ (Artstein and Poesio, 2008).

All 77 development disagreements were adjudicated. On test, the adjudicator reviewed 71 disagreements and 140 low-confidence consensus cases; the remaining 364 cases retained the independent consensus. Raw agreement is 73.1% on development ($\kappa = 0.621$) and 87.7% on test ($\kappa = 0.813$). Table 2 reports the final distributions. The technical companion documents the pilot and adjudication audit.

2.4 Tasks and Evidence Sequence

Task 1: Admission-only prediction. The model receives patient information, the admission diagnosis, and the admission note. It predicts the final diagnosis, its ICD-3 code, and the revision label, in that order. The revision label can therefore depend on the diagnosis the model has just generated.

Task 2: Revision with accumulating evidence. The model receives cumulative evidence in up to four rounds: (1) admission information; (2) first-course diagnostic analysis and auxiliary findings; (3) daily progress notes; and (4) surgery and post-operative records, when available. At each round, it reports a working diagnosis and one action: MAINTAIN, REFINE, REVISE, or DEFER. The first three indicate the degree of diagnostic change; DEFER means that the decision remains unresolved. Daily notes are used both as model inputs and during annotation. Discharge records are used only to construct the reference.

Each case has one reference label covering admission to discharge. We combine the model's actions into one prediction using the rule in Section 4.2. Task 1 tests what the admission record alone reveals about the eventual diagnosis and its relation to the initial assessment. Task 2 tests how the model handles that relation as findings arrive. The shared reference makes the tasks comparable while preserving their different information settings.

The diagnosis and revision outputs answer different questions. A model might anticipate lung cancer from an admission note yet label it as a strong revision, even though the note already suspects a tumor. Conversely, a model might correctly recognize that a different disease is needed to explain later findings while predicting the wrong disease. We therefore score the diagnosis and the revision decision separately in both tasks.

3 DIAGAGENT: Diagnosis Tracking

DIAGAGENT keeps a shared record, `DxState`, of its ranked diagnoses, ICD codes, supporting and conflicting evidence, unresolved questions, and previous actions. Reading this record across rounds connects each decision to the evidence and diagno-

sis that produced it. In particular, we can inspect whether a new finding changes the leading diagnosis or is simply added to the existing explanation.

A Triage Agent initializes the record. An Exam Agent reads new evidence and proposes an update. When confidence is low or diagnoses conflict, Specialist Agents provide additional assessments. A Revision Agent then chooses one of the four actions and updates `DxState`. All roles use the same underlying LLM, following role-based collaboration in medical LLM systems (Kim et al., 2024; Wang et al., 2025).

We compare the full system with Single-Agent State, which keeps the same record but uses one agent to update it. Three further ablations remove, respectively, the revision trigger, the evidence ledger, and specialist routing. These components require an explicit revision decision, store supporting and conflicting evidence, and determine when specialists are called. The comparisons test whether each component helps the system track diagnoses or choose revision actions.

DIAGAGENT initializes its state with the admission note and first-course analysis, records a required DEFER action, and reads the first-course evidence again at its first update. Prompt baselines begin with admission information alone. We compare the systems using decisions aggregated over each case. Removing the initialization action leaves all 575 test case-level predictions unchanged. The repository provides the exact role prompts, initialization code, and state fields.

4 Experimental Setup

4.1 Models and Baselines

We evaluate DeepSeek-V3.2 in non-thinking mode (API alias `deepseek-chat`, denoted DeepSeek-Chat).[2] We compare five prompting strategies: Direct, chain-of-thought (CoT) (Wei et al., 2022), Structured Output, Self-Refine (Madaan et al., 2023), and Self-Consistency (Wang et al., 2023). Direct requests a JSON answer; Structured Output provides an explicit template. Self-Refine generates, critiques, and revises an answer. Self-Consistency votes over five samples. All scores use the saved outputs from the original runs.

We also evaluate DeepSeek-R1-Distill-Qwen-14B-AWQ (R1-14B), DeepSeek-R1-Distill-Qwen-7B, GLM-4-9B-Chat, InternLM3-8B-Instruct, and Qwen3.5-4B. Direct is run on all five open-weight models, and Self-Refine on the first four. HuatuoGPT-o1-7B, a medical reasoning model (Chen et al., 2024), is tested with Direct, Structured, Self-Refine, and full DIAGAGENT. R1-14B, GLM-4-9B, and InternLM3-8B also have full DIAGAGENT runs. All LLMs are used without fine-tuning. Checkpoint names, decoding settings, and a complete list of runs are provided in the repository.

Two simple baselines help interpret the LLM scores. Always-Weak predicts WEAK REFINEMENT for every case. A TF-IDF classifier uses admission text and the admission diagnosis with character 1–3-grams (`char_wb`, minimum document frequency 2, at most 50,000 features) and class-balanced logistic regression ($C = 1$). It is trained on the 2,004 weakly labeled cases.

4.2 Evaluation Metrics

Case-level revision. Our main comparison uses one revision prediction per case. For Task 2, the aggregation rule selects REVISE if it occurs in any round; otherwise, it selects REFINE if present. If neither occurs, it uses the last action, with DEFER for an empty sequence. Thus, Maintain–Refine–Maintain becomes Refine, whereas Defer–Maintain–Defer remains Defer. This rule preserves a diagnostic change made before the last round, but can also retain an early mistake that is later withdrawn. We examine alternative rules in Section 6 and the technical companion.

NoR, WR, and SR correspond to MAINTAIN, REFINE, and REVISE. Primary test scores use the 516 cases with these reference labels, excluding S/M. We report accuracy (Acc.), macro recall (Macro-R), and macro-F1. Macro recall is the unweighted mean of the three class recalls; macro-F1 is the unweighted mean of their F1 scores. Both assign equal importance to each class despite the class imbalance. A DEFER prediction is counted as an error and as a false negative for its reference class. Unresolved or empty outputs remain in the evaluation.

Eff-M is a secondary, benchmark-specific measure: the mean of the three class recalls and the proportion of S/M cases assigned any non-DEFER action. It uses all 575 test cases. Its S/M component measures whether the model commits to an action. We therefore report exact four-class macro-F1 separately for annotation consistency; the companion also provides stricter S/M evaluations.

[2] Version inferred from April 2026 run dates and the API changelog: `https://api-docs.deepseek.com/updates`. No immutable snapshot ID was retained.

Comparison	Acc.	Macro-R	Macro-F1	
			3-class	4-class
A → final	91.5	92.4	87.9	89.0
B → final	96.3	95.6	94.9	94.7
A → B	87.8	86.3	83.5	–
B → A	87.1	80.4	83.1	–

Table 3: Test-set annotation consistency (%). Arrows indicate prediction → reference. Accuracy, macro recall, and three-class macro-F1 exclude reference S/M cases; four-class macro-F1 uses all labels.

Diagnosis and round-level decisions. We score the top-ranked diagnosis ($k = 1$) by exact name, ICD-3, and leading ICD-letter match. The last measure is called Prefix; it compares the first letter, not an entire ICD chapter. Task 1 Rev-F1 is macro-F1 over all four revision labels. Task 2 diagnosis scores use the last available round. Its decision metrics compare actions with a sequence constructed from the case label: Maintain throughout for NoR; Maintain until a final Refine or Revise for WR or SR; and Defer throughout for S/M. Dec-Acc averages within-case round accuracy across cases; Dec-F1 pools rounds and computes macro-F1 over the four actions. These constructed sequences provide secondary diagnostics; the primary evaluation uses the human case-level labels.

Statistical comparisons. Paired comparisons use 1,000 bootstrap resamples of cases. We report 95% confidence intervals and approximate two-sided bootstrap p-values without correction for multiple comparisons. Domain and component comparisons are exploratory. Differences are calculated from unrounded scores and reported in percentage points.

4.3 Memory and Late Evidence

Prompt baselines receive all clinical evidence available so far, but not their previous outputs. DIAGAGENT also retains its own state. To examine this difference, a separate experiment adds previous outputs to Direct prompts using DeepSeek-V4-flash. It compares 574 shared cases; one case has an empty sequence in the run without memory. Because V4-flash is a later API version, these results are reported separately from the main evaluation.

Every prompt excludes the discharge record. Surgery and postoperative records can contain the final diagnosis: postoperative ICD-3 matches discharge ICD-3 in 109 of the 178 test cases with surgery records. We measure the contribution of Round 4 actions by removing them from saved sequences before applying the case-level aggregation rule. This analysis changes the actions used for scoring, with the original model inputs and outputs held fixed.

5 Results

5.1 Validation of ICD-Derived Labels

ICD-derived labels agree with the final human labels in 475 of 575 test cases (82.6%; bootstrap 95% CI 79.5–85.6; Figure 1). Most corrections reduce the degree of revision. Of 122 weak SR cases, 71 become WR, six become NoR, and two become S/M; only 43 retain SR. The rule identifies most of the 47 reference SR cases (91.5% recall), but only 35.2% of its SR predictions are correct (95% CI 26.8–43.9).

The pattern differs on development: overall agreement is 92.3%, and 19 weak WR cases become SR. Weak SR precision is 100.0% on development but 35.2% on test. The split was fixed before annotation, with similar weak-label proportions. Thus, similar code-based distributions do not guarantee similar agreement with judgments based on the text. Checking the rule only on development would have missed its tendency to overstate revision on test.

5.2 Overall Revision Performance

Using each annotator's labels in turn as the reference yields 80.4–86.3% macro recall and 83.1–83.5% macro-F1 (Table 3). Scores differ by direction because the reference labels determine class frequencies and which cases are non-Special. Scores against final labels are higher, as adjudication selects an annotator's answer in each disagreement. These scores measure retrospective annotation consistency with access to the discharge note.

DeepSeek Self-Consistency leads the main LLM evaluation with 52.6% Macro-R and 42.8% Macro-F1 (Table 4). On the 364 cases that required no adjudication, its Macro-R is 54.0%. Revision remains difficult even among cases with clear initial agreement between annotators.

Always-Weak reaches 51.7% accuracy because WR is the largest class, but only 33.3% macro recall. The TF-IDF classifier raises macro recall to 41.6% and predicts SR for 121 cases, compared with 47 in the reference. Training on the weak labels yields a classifier that uses all three revision categories but predicts strong revisions too often.

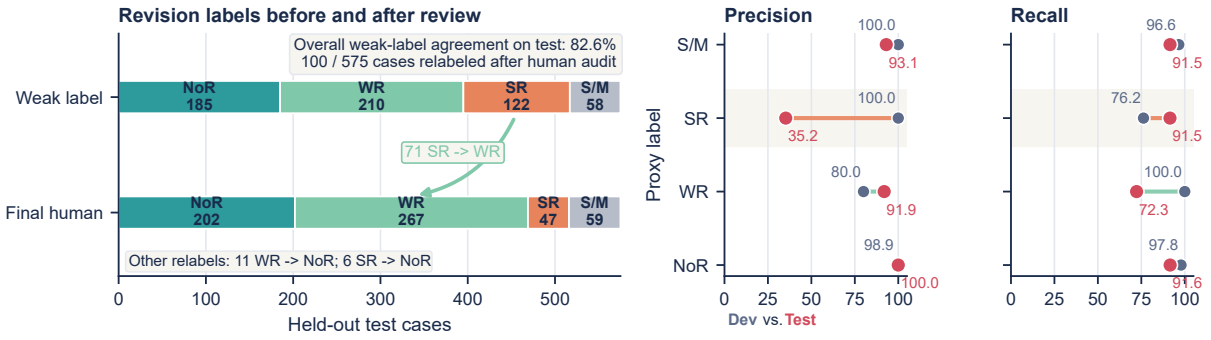


Figure 1: Audit of ICD-derived labels. Left: the test-label distribution before and after review; 71 weak SR cases become WR. Right: classwise precision and recall of the weak labels against the human reference on development and test. The code-based rule behaves differently across the two splits.

System	Acc.	Macro-R	Macro-F1	NoR	WR	SR	Eff-M
Always-Weak	51.7	33.3	22.7	0.0	100.0	0.0	50.0
Weak-label TF-IDF	45.2	41.6	40.6	46.5	46.4	31.9	35.9
DeepSeek-Chat Direct	48.6	51.7	40.3	12.9	71.9	70.2	63.7
DeepSeek-Chat CoT	49.0	49.1	40.0	14.4	73.4	59.6	61.8
DeepSeek-Chat Structured	48.1	50.5	39.3	10.4	73.0	68.1	62.9
DeepSeek-Chat Self-Refine	46.7	47.5	41.0	26.7	60.3	55.3	60.6
DeepSeek-Chat Self-Consistency	49.8	**52.6**	**42.8**	15.3	72.3	70.2	63.2
R1-14B Direct	45.0	50.7	42.8	58.4	31.8	61.7	63.0
GLM-4-9B Direct	48.3	37.9	34.2	11.4	80.9	21.3	53.4
InternLM3-8B Direct	47.7	43.1	41.1	80.2	27.7	21.3	57.3
HuatuoGPT-o1-7B Direct	49.4	40.3	37.0	82.2	32.2	6.4	55.2
HuatuoGPT-o1-7B Self-Refine	35.5	44.3	34.2	26.7	36.0	70.2	57.8
HuatuoGPT-o1-7B Structured	44.2	39.8	36.3	75.2	25.1	19.1	54.9
DIAGAGENT Full (DeepSeek-Chat)	34.1	37.7	29.8	63.9	10.9	38.3	53.3
Single-Agent State (DeepSeek-Chat)	32.8	45.4	32.4	25.7	29.6	80.9	58.6

Table 4: Case-level test results (%). Acc., Macro-R, and Macro-F1 use 516 non-Special cases; NoR/WR/SR are class recalls. Eff-M uses all 575 cases with the permissive S/M rule. Bold marks the highest macro scores among the LLM rows shown before rounding. The later-version memory experiment is reported separately.

HuatuoGPT-o1-7B reaches 44.3% Macro-R with Self-Refine. Changing from Direct to Self-Refine raises its SR recall from 6.4% to 70.2%, while lowering NoR recall from 82.2% to 26.7%.

DeepSeek-Chat Direct recognizes many refinements and strong revisions but seldom keeps an unchanged diagnosis. InternLM3-8B Direct and HuatuoGPT-o1-7B Direct do the reverse: they identify most NoR cases while missing many cases that require change. R1-14B Direct has more balanced NoR and SR recall, with refinement remaining its weakest category.

5.3 Diagnosis and Revision Performance

Better diagnosis prediction does not necessarily mean better revision decisions (Table 5). Direct leads Task 1 prefix match, Self-Consistency leads Task 1 ICD-3 match, and Structured leads Task 2 ICD-3 match. Self-Refine leads the round-level decision metrics, while Self-Consistency leads case-level macro recall.

The same pattern appears across models using Direct prompts. DeepSeek-Chat has higher Task 2 ICD-3 accuracy than InternLM3-8B (54.4 vs. 44.9), but lower Dec-F1 (20.4 vs. 31.5). R1-14B achieves 21.7% ICD-3 accuracy and 26.0%

	Task 1			Task 2		
DeepSeek-Chat strategy	**Prefix**	**ICD-3**	**Rev-F1**	**ICD-3**	**Dec-Acc**	**Dec-F1**
Direct	74.4	53.9	25.7	54.4	29.6	20.4
CoT	73.9	52.7	23.3	55.1	31.0	20.6
Structured	72.2	50.4	26.3	56.5	27.6	19.2
Self-Refine	73.7	52.2	26.1	55.0	46.1	24.3
Self-Consistency	73.6	54.1	25.7	53.9	30.4	21.6

Table 5: Task-specific scores on 575 test cases (%). All diagnosis-code scores use rank 1 ($k = 1$). Prefix compares the leading ICD letter. Task 2 code prediction includes postoperative evidence; round-level decision metrics use action sequences constructed from the case label.

Dec-F1. DeepSeek-Chat therefore leads this comparison in diagnosis-code prediction, while both other models score higher on round-level revision decisions.

Prompting effects also depend on the model. Self-Refine lowers Eff-M from 57.3 to 49.2 for InternLM3-8B, from 63.0 to 54.5 for R1-14B, and from 53.4 to 52.0 for GLM-4-9B. Its advantage on DeepSeek-Chat's round-level scores thus coexists with lower case-level scores on these open-weight models.

5.4 Effects of Agent Components

On DeepSeek-Chat, Full DIAGAGENT achieves 37.7% Macro-R, compared with 45.4% for Single-Agent State ($p = 0.018$); Eff-M is 53.3 versus 58.6 ($p = 0.024$). One agent therefore performs better on these macro metrics, although its Acc. is slightly lower. Self-Refine also exceeds Full on Macro-R by 9.8 percentage points ($p = 0.004$). Recording state helps us inspect decisions, but additional roles do not improve the main macro scores in this comparison.

Full's class recall varies strongly with the underlying model (Table 6). DeepSeek-Chat correctly keeps many NoR cases unchanged but misses most WR cases. GLM-4-9B recognizes WR much more often and rarely identifies NoR. InternLM3-8B has high SR recall. Across these four matched models, Full never exceeds Direct on Eff-M. HuatuoGPT-o1 Full likewise reaches 34.7% Macro-R and 51.0% Eff-M, below its Direct scores.

Component ablations also give mixed results. Removing the revision trigger reduces SR recall to 5.0% on development, but its effect on test Eff-M is not significant ($p = 0.246$). Removing the evidence ledger improves some development decision scores while worsening diagnosis-code prediction.

	Eff-M		Full class recall		
Backbone	**Direct**	**Full**	**NoR**	**WR**	**SR**
DeepSeek-Chat	63.7	53.3	63.9	10.9	38.3
R1-14B	63.0	46.9	46.5	7.9	46.8
GLM-4-9B	53.4	52.7	6.4	74.5	29.8
InternLM3-8B	57.3	57.2	5.9	41.9	80.9

Table 6: Matched Direct and Full DIAGAGENT results (%). Full's class recall varies strongly across models.

Domain	**System**	**Macro-R**	**Macro-F1**	**Eff-M**
B	Self-Refine	45.0	41.2	58.7
B	Full DIAGAGENT	49.0	41.4	61.7
C	Self-Refine	38.4	34.3	53.8
C	Full DIAGAGENT	47.6	39.7	60.7

Table 7: Separate runs on two data domains with DeepSeek-Chat (%). Both systems use the same final human labels within each domain.

The technical companion provides complete ablation and cost tables and identifies which scores use weak labels and which use human labels.

5.5 Performance across Data Domains

Separate runs on Domains B and C contain 88 and 185 test cases. Full DIAGAGENT has higher macro recall than Self-Refine in both domains (Table 7), reversing the full-test ordering. The paired differences are not significant (Macro-R $p = 0.509$ on B and 0.074 on C). These within-corpus comparisons show how the numerical ordering varies with the cases being evaluated.

5.6 Memory and Late-Evidence Effects

In the separate V4-flash experiment, adding previous outputs to Direct prompts increases Macro-R from 55.1 to 58.3% on shared cases. The paired estimate is an increase of 3.3 percentage points

Variant	Macro-R	Macro-F1	Eff-M	Dec-F1
Without memory	55.1	49.7	66.3	29.7
With memory	58.3	48.2	68.7	24.5

Table 8: Output-memory sensitivity on 574 shared test cases (%), including 515 non-Special cases. Both variants use the later V4-flash version.

(95% CI $[-2.1, 8.8]$, $p = 0.250$). Eff-M rises by 2.5 percentage points, while Dec-F1 falls by 5.2 percentage points. The macro-recall gain is not statistically significant, and the two scoring levels respond differently to output memory.

Removing Round 4 actions from saved sequences changes Eff-M by at most 1.17 percentage points and Macro-R by at most 1.55 percentage points across the tested systems. The case-level revision scores are therefore largely preserved when the final action is omitted. Task 2 diagnosis scores, including the best ICD-3 result of 56.5%, cover the full evidence sequence with postoperative records.

6 Analysis of Diagnostic Revision

6.1 Errors at Category Boundaries

Across the five original DeepSeek prompting strategies, 80.3–87.7% of incorrect non-Special predictions confuse neighboring labels: NoR with WR, or WR with SR. In all five runs, the most frequent error is labeling a NoR case as WR. Direct, for example, assigns WR to 137 of the 202 NoR cases. The model often treats added information as a reason to refine the diagnosis, even when the main explanation remains unchanged.

The boundary between refinement and revision requires a different comparison: whether the final disease was already part of the initial assessment. "Consider malignancy" records a suspicion that later pathology can confirm. "Malignancy excluded" assigns the same disease a different status. Reading the disease name without its surrounding wording loses this distinction. Annotators use the first-course analysis to recover the initial hypothesis and later notes to determine whether it was supported or replaced.

The timing of a diagnosis also changes the interpretation. In an admission for planned chemotherapy, a known cancer can appear in both the medical history and the discharge record, while the admission code describes the treatment visit. The final cancer code then refers to an established condition. The annotation follows that history across notes, rather than treating each diagnosis field as a new assessment. This accounts for the separate S/M category and for the guideline's emphasis on the purpose of admission.

Strategy	Adjacent	NoR–SR	Defer
Direct	84.5	15.5	0.0
CoT	87.1	12.9	0.0
Structured	87.7	11.9	0.4
Self-Refine	81.5	18.5	0.0
Self-Consistency	80.3	13.1	6.6

Table 9: Error composition for the original DeepSeek prompt sweep (% of incorrect non-Special cases). Each row partitions all errors.

Table 9 separates neighboring-label errors, direct NoR–SR errors, and unresolved decisions. Self-Consistency has the smallest share of neighboring-label errors and the largest share of DEFER errors.

6.2 Longitudinal Diagnosis Trajectories

The DeepSeek Full system assigns MAINTAIN to 185 of 267 WR cases and 24 of 47 SR cases. Its records show how new findings can be absorbed into an existing diagnosis instead of changing it. One reviewed SR case begins with rheumatoid arthritis and ends with enterococcal septicemia. The model first lists active rheumatoid arthritis as its leading diagnosis. After blood cultures identify infection, it adds infection as a complication and chooses REFINE. Rheumatoid arthritis remains the leading diagnosis despite the new evidence.

A second case has bronchiectasis with infection at both admission and discharge. The model initially describes a bacterial infection, then adds a fungal component after reading later findings and antifungal treatment. It chooses REFINE, while the human label is NoR. The added detail changes the generated description without changing the principal diagnosis captured by the reference label.

In a successful SR case, the admission diagnosis is fever. DIAGAGENT initially considers adult-onset Still's disease, then infective endocarditis, and finally tuberculosis. Its last update cites negative blood cultures and improvement after antituberculous treatment as evidence for changing the leading diagnosis. The final action is REVISE. Self-Refine, by comparison, keeps adult-onset Still's disease throughout the same case.

Some Task 1 and agent instructions associate an ICD-prefix change with strong revision. In a chemotherapy admission for known pancreatic cancer, the Exam Agent recommends MAINTAIN, but the Revision Agent cites the Z-to-C code transi-

tion and chooses REVISE. The diagnosis remains pancreatic cancer throughout the record. The final action follows the code rule despite agreement on the underlying disease. This mismatch between the prompt and the text-based guideline explains why retrieving the correct diagnosis can still produce the wrong revision decision.

6.3 Evaluation and Reproducibility

The label audit and model errors involve different boundaries. The largest audit correction moves weak SR cases to WR: a change in code often confirms an existing suspicion. The dominant DeepSeek prompting error instead moves NoR cases to WR: the model refines a diagnosis that the annotation leaves unchanged. Revising the ICD rule and improving model decisions therefore require attention to different comparisons.

Action aggregation changes which decisions contribute to the score. The main rule preserves any revision or refinement made during a case; final-action scoring measures where the model ends. In the saved aggregation analysis, switching Single-Agent State to final-action scoring raises Acc. from 32.8 to 51.2%, while Macro-R falls from 45.4 to 43.0%. Final accuracy improves even though mean class recall falls. The released action sequences support either scoring rule.

The repository provides exact prompts, parsers, scoring and bootstrap code, and per-case records without clinical text. These records retain both annotators' original labels, model labels, round actions, and diagnosis-match indicators under new study keys, allowing rescoring without clinical text or API access. Synthetic examples demonstrate input and output formats and are excluded from experiment counts. The companion provides extended audit, model, domain, and ablation results. Clinical narratives and original linkage keys are not public. New inference or training requires record-level access; researchers may contact the corresponding author for review.

7 Related Work

Clinical text and temporal evaluation. MIMIC-III and MIMIC-IV support a wide range of clinical NLP research (Johnson et al., 2016, 2023). emrQA and EHRSQL evaluate questions over clinical records (Pampari et al., 2018; Lee et al., 2022); CMExam and CMB evaluate medical knowledge and reasoning in Chinese (Liu et al., 2023; Wang et al., 2024). CliMedBench covers clinical scenarios, including inpatient tasks (Ouyang et al., 2024). TIMER, ER-Reason, and MedR-Bench further examine reasoning over time, emergency-room decisions, and clinical cases (Cui et al., 2025; Mehandru et al., 2025; Qiu et al., 2025). INA-Bench studies adaptive information acquisition as inpatient notes are revealed in sequence (Feng et al., 2026). DIAGBENCH focuses on whether later evidence supports the initial diagnosis, makes it more specific, or replaces it.

Agents and correction of clinical notes. AI Hospital, AgentClinic, and MedAgentBench evaluate interactions and actions in clinical environments (Fan et al., 2025; Schmidgall et al., 2024; Jiang et al., 2025). MDAgents and ColaCare study collaboration among medical agents (Kim et al., 2024; Wang et al., 2025). Our agent comparisons test whether dividing work among roles helps diagnostic revision. MEDEC and MedRECT evaluate error detection and correction in completed clinical text (Ben Abacha et al., 2025; Iwase et al., 2025). In DIAGBENCH, the target is the change between an early assessment and a later diagnosis. A diagnosis may reasonably change as evidence arrives without the earlier assessment having been an error.

8 Conclusion

DIAGBENCH evaluates long-horizon diagnostic revision over inpatient records, distinguishing confirmation and refinement from replacement of an initial disease explanation. The best LLM configuration in the main evaluation achieves 52.6% macro recall, with errors concentrated at category boundaries. Accurate final-diagnosis prediction does not consistently imply appropriate revision, and state traces reveal failures to update a leading diagnosis despite recognizing new evidence. These findings motivate explicit checks of whether new evidence supports the current diagnosis or requires a different disease explanation. Future systems should be assessed on both diagnostic accuracy and the appropriateness of their updates, including how they respond to contradictory findings across successive rounds.

Limitations

The reference labels are retrospective judgments by medical-NLP researchers. Physician-panel validation would provide an additional clinical assessment of the category boundaries. Annotators read ICD transitions and complete notes, including discharge records; the adjudicator also saw weak labels. This information can influence annotation, and the resulting agreement scores measure label consistency rather than human performance under the model's input conditions. The AI-assisted pilot served to develop the guideline and is separate from the independent development and test annotation.

The benchmark covers two hospitals and three local domains. Splits are by admission, and the available identifiers do not establish patient-disjointness across separate hospitalizations. Evaluation in other institutions and languages would test transfer beyond this collection. The training split uses weak labels; supervised evaluation here is limited to the TF-IDF classifier.

Some original prompts tie revision to ICD changes, introducing a mismatch with the text-based guideline. Results describe complete model and prompt configurations, including their decoding settings and API versions. The memory study uses the later V4-flash version, and medical-model coverage consists of one checkpoint. Prompt baselines and agents also differ in initialization, memory, and inference cost.

Each admission has one reference label rather than expert labels for individual rounds. Aggregation can retain a revision that the model later withdraws. Postoperative fields can disclose the final diagnosis; testing their effect on inference requires reruns with those fields masked. Diagnosis-code scores measure agreement with documentation, and Eff-M accepts any committed action for S/M. Public records support rescoring; new inference requires separately approved access to clinical text.

Ethical Considerations

This retrospective study uses de-identified inpatient records from two tertiary teaching hospitals. All records were de-identified before benchmark construction, with direct identifiers and other personally identifying information removed. The study was reviewed and approved by the ethics committees of the participating hospitals before any experiments were conducted, and the data were used solely for research purposes. For API-based experiments, only de-identified text was transmitted, and no identifiable patient information was provided to external model providers. External API use followed the applicable institutional data-governance requirements and provider terms for research use.

Generative AI Assistance

We used generative AI tools to assist with wording, presentation, and preparation of reproducibility materials. The preliminary pilot also included AI-assisted internal screening, as described in Section 2, separately from human annotation of the development and test sets. Research records document the reported experiments and results. The authors are responsible for the final content, citations, code, and numerical claims.

Acknowledgments

This work was supported by the Computational Biology Program (Grant No. 25JS2830402) of Science and Technology Commission of Shanghai Municipality (STCSM), and the Shanghai Municipal Science and Technology Major Project (Grant No. 2025SHZDZX025G06).

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A benchmark for medical error detection and correction in clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550, Vienna, Austria. Association for Computational Linguistics.

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

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H. Shah. 2025. TIMER: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on*

Computational Linguistics, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Huawei Feng, Dongsheng Shi, and Linlin Wang. 2026. Benchmarking large language models for adaptive information acquisition and diagnosis with sequentially revealed inpatient records. In *Findings of the Association for Computational Linguistics: EMNLP 2026*.

Mark L. Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165(13):1493–1499.

Naoto Iwase, Hiroki Okuyama, and Junichiro Iwasawa. 2025. MedRECT: A medical reasoning benchmark for error correction in clinical texts. *arXiv preprint arXiv:2511.00421*.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents. *arXiv preprint arXiv:2501.14654*.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. EHRSQL: A practical text-to-sql benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, volume 35.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking large language models on CMExam: A comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594.

Nikita Mehandru, Niloufar Golchini, Namrata Garg, Kathy T. LeSaint, Christopher J. Nash, Anu Ramachandran, Travis Zack, Liam G. McCoy, Adam Rodman, David Bamman, Melanie Molina, and Ahmed Alaa. 2025. ER-Reason: A benchmark dataset for LLM clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A large-scale chinese benchmark for evaluating medical large language models in clinical scenarios. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16(1):9799.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen,

Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. 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*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen M. Harrison, Chengwei Pan, Junyi Gao, and Liantao Ma. 2025. ColaCare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *WWW '25: Proceedings of the ACM on Web Conference 2025*, pages 2250–2261. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.