CLINTRACE: BENCHMARKING CLINICAL REASONING AS EVIDENCE ACCUMULATES

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-assisted clinical decision support requires large language models (LLMs) to reason and make decisions as evidence accumulates, not merely predict final diagnoses. However, existing benchmarks rarely provide expert reasoning traces aligned with the evidence available at each stage. To address this gap, we introduce CLINTRACE, a longitudinal clinical reasoning benchmark built from live, multi-expert discussions of real-world cases. Unlike retrospective rationales, these discussions capture physicians' reasoning before subsequent findings are revealed. Using a physician co-designed taxonomy and an audited data-extraction pipeline, we structure these discussions into time-aligned trajectories that link available evidence to expert interpretations, hypothesis updates, and proposed actions. The dataset includes 1,464 public cases with 84,044 atomic reasoning steps, plus 184 non-public cases. We formulate five complementary tasks to evaluate longitudinal clinical reasoning: progressive diagnosis and reasoning, action planning, hypothesis tracking, evidence–hypothesis relation prediction, and clinical statement typing. Evaluation of 11 general-purpose and medical LLMs shows that strong diagnostic performance coexists with incomplete coverage of documented expert reasoning. The leading model achieves 76.7% top-5 diagnostic accuracy but only 33.3% reasoning comprehensiveness. Supervised fine-tuning improves longitudinal reasoning capabilities and transfers beyond CLINTRACE, with diagnostic gains on two external benchmarks. Together, CLINTRACE supports evaluation and learning of clinical reasoning as evidence accumulates, beyond final-answer accuracy. We will release the public dataset and retain the non-public cases as a protected evaluation set.

1 INTRODUCTION

Clinical diagnosis is not a one-shot prediction task. As patient information accumulates, clinicians interpret new evidence, revise diagnostic hypotheses, and select appropriate next actions (Norman & Eva, 2010; Sox et al., 2024). Final-diagnosis accuracy captures only the endpoint of this process: a model may reach the correct diagnosis while overlooking important evidence, retaining hypotheses contradicted by new findings, or proposing inappropriate actions (Hager et al., 2024; Qiu et al., 2025; Chen et al., 2026). This distinction matters for clinical decision support, as randomized trials show that strong standalone model performance does not consistently translate into improved clinician performance when LLM assistance is introduced (Goh et al., 2024; 2025; Qazi et al., 2026; Everett et al., 2026; Rounding et al., 2026). Evaluating clinical AI therefore requires assessing not only what models ultimately predict, but also how their reasoning and decisions evolve as a case unfolds.

Evaluating this process requires aligning three elements at each stage: the information available at that time, the judgments made before future findings are known, and the resulting updates to hypotheses and actions. Existing benchmarks rarely preserve all three. Endpoint-oriented benchmarks primarily evaluate predictions from complete cases or accumulated electronic health records (EHRs) (Jin et al., 2021; Wornow et al., 2023; Cui et al., 2025); rationale-based benchmarks often reconstruct reasoning after the outcome is known (Wu et al., 2025; Qiu et al., 2025; Zhou et al., 2025); and sequential or interactive benchmarks generally lack aligned step-level references from clinicians reasoning under the same partial information (Schmidgall et al., 2026; Jiang et al., 2025; Chiu et al., 2025; Molina et al., 2025). Consequently, these settings cannot directly test whether a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

model updates its reasoning prospectively as evidence arrives or produces a plausible explanation only after sufficient diagnostic information is available (Table 1).

Live, staged clinical case discussions provide a natural source of such time-aligned expert reasoning traces. We introduce CLINTRACE (**Clin**ical **T**rajectories of **R**easoning **A**cross **C**umulative **E**vidence), a longitudinal clinical reasoning benchmark built primarily from Virtual Morning Report (VMR) sessions hosted by the Clinical Problem Solvers (CPSolvers), a global community for clinical reasoning education. In each session, a presenter sequentially reveals a de-identified real-world case while multiple physician discussants verbalize their reasoning without access to subsequent findings. We structure these discussions into trajectories linking each stage's available evidence to expert interpretations, hypothesis updates, and proposed actions. To complement the public VMR corpus, we include 184 non-public cases from RLR, a subscription-only CPSolvers series in which two expert physician-educators present cases to one another in stages, with the discussant verbalizing their reasoning at each stage before further evidence is revealed. Across 11 general-purpose and medical LLMs, our evaluation shows that strong diagnostic performance coexists with incomplete coverage of documented expert reasoning.

Our contributions are:

- **Live, staged expert reasoning dataset.** We construct CLINTRACE from multi-expert case discussions using a physician co-designed taxonomy and audited extraction pipeline, providing time-aligned reasoning traces in a public corpus and a protected evaluation set (§3).
- **Longitudinal evaluation framework.** We formulate five complementary tasks and a time-aware protocol to assess how models interpret evidence, revise hypotheses, and select actions as cases unfold, extending evaluation beyond final-diagnosis accuracy (§4).
- **Systematic analysis of model behavior.** We evaluate 11 LLMs under three prompting strategies, showing that strong diagnostic performance coexists with incomplete coverage of expert reasoning. Further analyses characterize differences in reasoning composition and timing (§5).
- **Learning from expert trajectories.** We demonstrate the dataset's value as supervision: fine-tuning improves reasoning coverage and added hypothesis recall, with Llama also achieving diagnostic gains on two external benchmarks (§5).

Table 1: Comparison of representative clinical reasoning benchmarks.

Benchmark	N	Source	Flow	Reference	Targets	Access
Static endpoint evaluation						
MedQA (Jin et al., 2021)	12,723	USMLE	Complete	–	Dx	Open
DiagnosisArena (Zhu et al., 2025)	1,113	Reports	Complete	Endpoint	Dx	Open
Retrospective reasoning evaluation						
MedThink-Bench (Zhou et al., 2025)	500	QA/exams	Complete	Post hoc	Dx/Rat	Open
DiReCT (Wang et al., 2024)	511	MIMIC-IV	Complete	Post hoc tree	Dx/Ev	Gated
MedR-Bench (Qiu et al., 2025)	1,453	Reports	Comp./Inter.	Post hoc report	Exam/Dx/Tx/Rat	Open
Sequential or interactive evaluation						
MIMIC-IV-Ext CDM (Hager et al., 2024)	2,400	MIMIC-IV	Interactive	Guidelines	Tests/Dx/Tx	Gated
MediQ (Li et al., 2024)	1,412	QA/vignettes	Interactive	–	Questions/Dx	Open
SCT-Bench (McCoy et al., 2025)	750	Vignettes	Single update	Elicited panel	Hyp	Mixed
ER-REASON (Molina et al., 2025)	3,984	EHR	Staged	Post hoc subset	Hyp/Act/Dx	Gated
MedChain (Liu et al., 2025)	12,163	Online	Sequential	Workflow only	Hx/Exam/Dx/Tx	Open
VivaBench (Chiu et al., 2025)	1,762	QA/vignettes	Interactive	–	Hx/Test/Dx	Open
SDBench (Nori et al., 2025)	304	NEJM CPC	Interactive	–	Questions/Dx/Cost	Limited
ClinDiag-Benchmark (Chen et al., 2026)	4,421	Mixed	Interactive	–	Hx/Exam/Test/Dx	Mixed
Ours						
CLINTRACE	**1,464 (+184)**	**CPSolvers**	**Stages**	**Live multi exp**	**Dx/Ev/Act/Hyp/Type**	**O+P**

Abbreviations: Act, action planning; exp, expert; Dx, diagnosis; Ev, evidence interpretation; Hx, history; Hyp, hypothesis tracking; O+P, Open+Proprietary; Rat, rationale; Tx, treatment; Type, statement typing. "+184" denotes our proprietary evaluation set.

2 RELATED WORK

Clinical reasoning benchmarks. Clinical benchmarks have assessed medical knowledge through multiple-choice examinations and question-answering tasks (Jin et al., 2021; Zuo et al., 2025; Siam et al., 2026), while more recent studies incorporate richer context from clinical vignettes, case reports, and longitudinal electronic health records (EHRs) (McDuff et al., 2025; Zhu et al., 2025; Wornow et al., 2023; Fleming et al., 2024). Reasoning-oriented benchmarks further provide ratio-

nales or decompose diagnosis into intermediate tasks: MedCaseReasoning and MedR-Bench extract reasoning from published case reports (Wu et al., 2025; Qiu et al., 2025), and MedThink adds expert-authored rationales to medical exam questions (Zhou et al., 2025). Longitudinal and interactive benchmarks introduce temporal structure through accumulated EHRs or simulated encounters in which agents acquire information and select actions (Wornow et al., 2023; Pellegrini et al., 2026; Schmidgall et al., 2026; Jiang et al., 2025; Liu et al., 2026; Chiu et al., 2025). However, these settings typically either expose the complete case, focus on prediction or retrospective reasoning generation tasks from fixed inference prompts. Therefore, they provide limited supervision for how expert reasoning evolves as new evidence is revealed. CLINTRACE is built from de-identified cases presented sequentially in live CPSolvers sessions, aligning each stage of case release with multi-expert clinical discussion. This design yields multi-perspective reasoning trajectories, enabling direct evaluation of both diagnostic outcomes and the evolution of clinical reasoning over time.

Process-level evaluation and model–expert comparison. Clinical model evaluation commonly emphasizes endpoint diagnosis accuracy, aggregate rationale quality, task completion, or holistic physician ratings (Hager et al., 2024; Qiu et al., 2025; Palepu et al., 2025; Tu et al., 2025; Yao et al., 2026). As shown in Table 1, existing benchmarks span static endpoint, retrospective reasoning, and sequential or interactive settings, but their reference signals are limited to endpoints, post hoc rationales or trees, guidelines, or workflow structure. ER-REASON evaluates multiple stages of an emergency-care workflow, yet its expert rationales cover only a subset of cases and are reconstructed retrospectively (Molina et al., 2025). Consequently, these benchmarks provide limited support for identifying where and when model reasoning diverges from clinician reasoning as evidence unfolds. CLINTRACE uses live multi-expert reasoning as the reference and composes five complementary tasks that compare 11 LLMs with clinicians across diagnosis, evidence attribution, hypothesis revision, action planning, and reasoning composition as cases unfold.

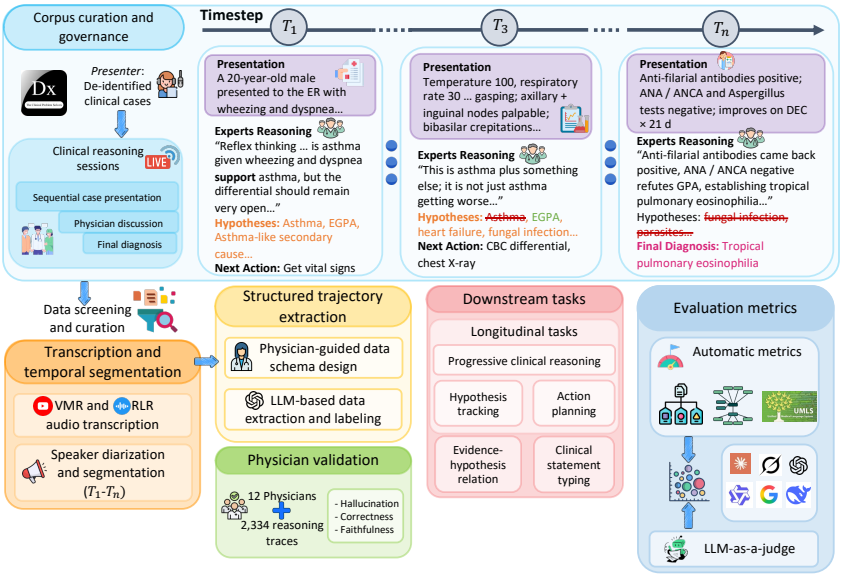


Figure 1: Overview of CLINTRACE. Expert reasoning traces are aligned with sequentially revealed evidence from real-world clinical cases. The figure illustrates the physician-audited curation pipeline, structured timestep-level annotations, five downstream tasks, and evaluation metrics.

3 DATASET CONSTRUCTION AND VALIDATION

To preserve temporal fidelity and annotation reliability, we construct CLINTRACE through four stages (Figure 1): corpus curation and governance, transcription and temporal segmentation, structured trajectory extraction with temporal constraints, and physician validation. We describe each stage below and then characterize the resulting public VMR dataset.

Corpus curation and governance. We collected 1,498 recordings from the CPSolvers VMR series. After excluding 7 private sessions and 27 sessions containing only informal conversation without

clinical case discussion, we retained 1,464 eligible cases. These cases form the public component of CLINTRACE and are partitioned by case into training, validation, and test splits. We additionally collected 184 cases from the subscription-only RLR series, reserved exclusively for evaluation and withheld from the public release. The study was approved by our Institutional Review Board and conducted with permission from CPSolvers.

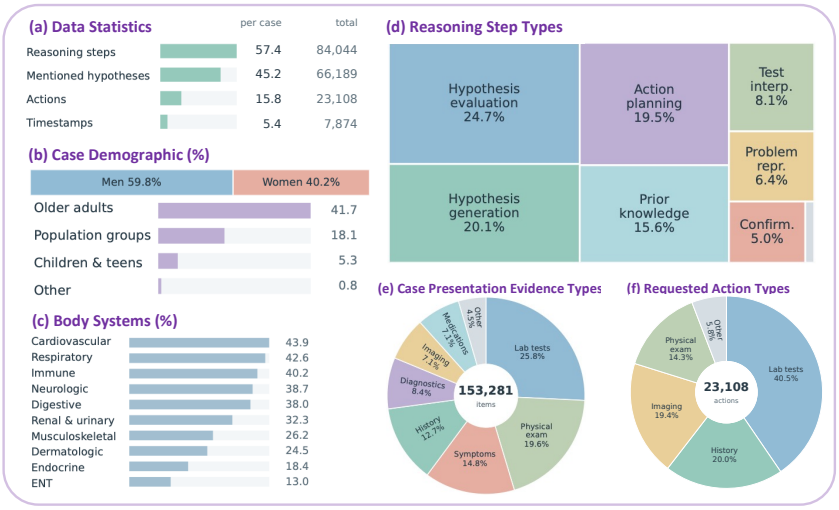


Figure 2: Composition and summary statistics of the public VMR component of CLINTRACE.

Transcription and temporal segmentation. We convert each recording into a speaker-attributed transcript using diarization to preserve speaker identity across long sessions. We segment each case whenever the designated presenter reveals new clinical information; the presenter does not participate in the diagnostic discussion. Each clinical fact and reasoning statement is anchored to the timestamp of the corresponding presenter turn. We define a *timestep* as one information-release stage, a *reasoning step* as an atomic reasoning statement within a timestep, and a *trajectory* as the ordered sequence of timesteps for a case. At timestep t_i, the state includes only information revealed through t_i, with no access to future turns ($t_{>i}$). Implementation details are provided in A.1.

Structured trajectory extraction and temporal integrity. For a case with timesteps (t_1 to t_n), the representation at timestep t_i includes the case information available through t_i, newly introduced evidence, atomic reasoning steps, evidence–hypothesis relations, requested actions, the cumulative set of active diagnostic hypotheses, and whether the final diagnosis has been revealed. The schema and label taxonomy are co-designed with physicians through iterative review and pilot extraction to capture how discussants interpret evidence, generate and evaluate hypotheses, plan actions, and revise the differential diagnosis. We use `GPT-5.4` through the Microsoft Azure API to conduct data extraction and labeling from speaker-attributed transcripts using physician-reviewed prompts. Each extracted item is anchored to its source timestep, and the representation at t_i excludes any information introduced after t_i. We further separate the clinical evidence and record how evidence supports or refutes hypotheses, as well as when hypotheses are added or removed. These annotations provide structured, multi-perspective trajectories rather than a single, uniquely correct reasoning path. Detailed field definitions, taxonomies, and extraction examples are provided in App. A.

Physician validation. We assess extraction fidelity through a field-level physician audit of 48 randomly sampled VMR cases by 12 CPSolvers-affiliated physicians. Six cases are independently reviewed by two physicians and 42 by one, with reviewers judging both faithfulness to the source discussion and label correctness. Across 1,532 annotation units in the double-reviewed cases, physicians achieve 95.7% agreement, and 97.7% of individual annotator judgments rate the extracted units as correct. Across 9,450 annotation units in the remaining 42 cases, pooled extraction accuracy is 96.6%, with 9 of the 11 evaluated fields exceeding 95% accuracy. These results support the high fidelity of the extracted trajectories to the source discussions. Detailed protocols and results are provided in App. A.6, with the annotation platform shown in App. A.7.

Dataset composition. Figure 2 summarizes the scale and diversity of the public VMR dataset. The 1,464 cases contain 7,874 timesteps, 84,044 reasoning steps, 66,189 hypothesis mentions, and

23,108 requested actions. Cases span diverse patient groups and organ systems, with broad representation across cardiovascular, respiratory, immune, neurologic, digestive, and other conditions. The annotations cover multiple evidence and action types, while reasoning steps are categorized into functions such as hypothesis generation and evaluation, action planning, test interpretation, and problem representation. Laboratory tests are the most common evidence type in case presentations and the most frequently requested action by experts (see App. A.2 for data composition details).

Table 2: Summary of the five downstream tasks of CLINTRACE.

Task	Model input → prediction	Reference signal
Progressive diagnostic reasoning	Case presentation through t_i → ranked differential, most likely diagnosis, and supporting reasoning	Final confirmed diagnosis; reasoning documented through t_i
Action planning	Case presentation through t_i → next diagnostic, monitoring, consultation, or management actions and their purposes	Actions requested before their outcomes were revealed
Hypothesis tracking	Prior active hypotheses and information newly revealed at t_{i+1} → hypotheses to add, retain, or remove and the updated differential	Hypothesis updates documented after the new information
Clinical evidence and reasoning typing	Evidence span or reasoning statement → functional type or reasoning tags	Annotated statement types and reasoning tags
Evidence–hypothesis relation prediction	Evidence item and candidate hypothesis → support or refute relation with justification	Physician-verified relation and documented justification

4 TASKS AND EVALUATION PROTOCOL

4.1 TASK FORMULATION

Clinical reasoning is an iterative process of gathering and interpreting clinical evidence, testing hypotheses, and making diagnostic and management decisions under uncertainty (Sox et al., 2024). To assess the different components of this process, we formulate five complementary tasks using sequence-indexed case trajectories (Table 2). Three *longitudinal tasks* assess reasoning as evidence unfolds, rather than only after a complete case is available: (1) *progressive diagnostic reasoning* assesses how the differential diagnosis and supporting reasoning evolve over time; (2) *action planning* evaluates prospective diagnostic, monitoring, consultation, and management decisions under partial information; and (3) *hypothesis tracking* compares model-generated hypothesis additions and removals with expert-documented updates following new evidence. Two *pointwise tasks* evaluate local reasoning operations: (4) *clinical evidence and reasoning typing* classifies evidence by information type (e.g., demographic, patient history, clinical symptom, or physical evaluation) and reasoning statements by functional category (e.g., hypothesis generation, hypothesis evaluation, or interpretation of test); and (5) *evidence–hypothesis relation prediction* determines whether a finding supports or refutes a hypothesis. The complete label taxonomy is provided in Table 5, while per-case counts and token-length statistics are summarized in Table 6.

4.2 EVALUATION PROTOCOL

We evaluate both final predictions and longitudinal clinical reasoning traces. At each timestep t_i, future case data is strictly masked for both model inputs and evaluation references. For evidence–hypothesis relation prediction, we evaluate only physician-verified "support" and "refute" labels, omitting subjective neutral or weak relations. All task prompts are physician co-designed. To ensure fair comparisons, all models share identical instructions, output schemas, and parsing rules. Task scores are computed per case or timestep. Metrics marked with * use `GPT-5.6 Luna` as an LLM judge (LLJ) with default settings. LLJ prompts appear in D.2.

Diagnosis. Evaluating diagnostic correctness is challenging because clinically equivalent diagnoses may differ in terminology and coding specificity. To evaluate complementary levels of diagnostic correctness, we report ICD-10 exact match for code-level agreement and top-k differential diagnosis (DDx) accuracy ($k \in \{1, 3, 5\}$) using an LLJ to determine whether the reference diagnosis is clinically equivalent to any of the first k diagnoses in the ranked differential (Gaber et al., 2025).

Clinical reasoning and planning. Because similar clinical inferences can be expressed in different ways, we use an LLJ to compare model-generated reasoning with expert trajectories. We assess (1) *reasoning trace recall*, which measures coverage of reference reasoning steps through semantic

matching (Wu et al., 2025), and (2) *reasoning comprehensiveness*, which applies a stricter criterion, crediting a reference rationale only when its key content is explicitly covered without contradictory statements (Zhou et al., 2025). We additionally measure *clinical concept recall*, defined as $|\mathrm{CUI}_{\mathrm{pred}} \cap \mathrm{CUI}_{\mathrm{ref}}|/|\mathrm{CUI}_{\mathrm{ref}}|$, where $\mathrm{CUI}_{\mathrm{pred}}$ and $\mathrm{CUI}_{\mathrm{ref}}$ are the sets of Unified Medical Language System (UMLS) Concept Unique Identifiers extracted from the predicted and reference reasoning using scispaCy (Molina et al., 2025; Neumann et al., 2019; Bodenreider, 2004). For action planning, we report LLJ-based *action recall* and *clinical concept recall* to assess coverage of reference actions and their associated clinical concepts, respectively. Reference actions include only those requested before their outcomes are revealed, preserving the prospective nature of the evaluation.

Hypothesis tracking. Given the information available at each timestep and the previously active hypotheses, models predict which hypotheses to add or remove. We evaluate these updates against expert trajectories using two metrics. *Added-hypothesis recall* measures the proportion of newly introduced reference hypotheses recovered by the model, with matches assessed by an LLJ. *Ruled-out hypothesis recall* measures the proportion of hypotheses removed by experts that the model eliminates, using string matching. Together, these metrics quantify how well models recover expert-documented additions to and removals from the differential diagnosis as evidence accumulates.

Statement and relation classification. We evaluate evidence types and reasoning step types as single-label multiclass classification, reasoning attributes as multi-label classification, and evidence–hypothesis relations as binary classification (*support* or *refute*). We report macro-F_1 separately for each prediction target to give equal weight to classes despite label imbalance.

LLM judge validation. To calibrate LLM-based evaluation, one physician scores 44 randomly sampled VMR cases that are also evaluated by two LLJs, `GPT-5.6 Luna` and `Claude Sonnet 5`. Both judges agree with the physician on 90.9% of top-1 diagnosis judgments and 97.6% of action-recall judgments, with lower agreement for reasoning comprehensiveness (88.0–90.4%) and reasoning trace recall (75.4–76.9%). Both LLJs assign scores that are equal to or lower than the physician's, indicating their conservative scoring. Given the single physician reviewer, we interpret this analysis as a calibration check rather than definitive evidence of inter-rater reliability (App. B.6).

5 Experiments and Results

5.1 Experimental Setup

Evaluation data. We report benchmark performance on 131 cases from the VMR test split and 184 non-public RLR cases, which have lower public exposure. Supervised fine-tuning (SFT) uses only the VMR training split and is evaluated with a single seed. All longitudinal evaluations follow the temporal constraints defined in Sec. 4.2; hypothesis tracking begins at the second timestep because it requires a previously established hypothesis set.

Models and inference settings. Following prior clinical-reasoning evaluations (Wang et al., 2026; Wu et al., 2025), we evaluate 11 LLMs in three groups: (1) three *proprietary general-purpose models*: `GPT-5.6 Sol` (`GPT`), `Claude Opus 5` (`Claude`), and `Grok 4.3` (`Grok`); (2) five *open-source general-purpose models*: `Kimi K2.6` (`Kimi`), `Deepseek V4 Pro` (`Deepseek`), `Llama 3.1 8B Instruct` (`Llama`), `Qwen 3.5 9B` (`Qwen9B`), and `Qwen 3.8 27B` (`Qwen27B`); and (3) three *open-source medical models*: `Lingshu 7B` (`Lingshu`), `HuatuoGPT 3 8B` (`Huatuo`), and `MedGemma 1.5 4B` (`MedGemma`). We use recommended decoding settings when available and provider defaults otherwise. `Llama` follows prior work (Wu et al., 2025), while `Qwen27B` uses low reasoning effort due to compute constraints. Detailed experiment configurations are provided in App. B.1.

Prompting conditions. We compare three prompting conditions to examine the effects of case presentation and access to prior expert reasoning. (1) *Full-case prompting* concatenates case presentations from all timesteps into a single static vignette, without expert reasoning traces. (2) *Cumulative-context prompting*, our primary setting, provides case presentations from t_1 through the evaluated timestep t_i, without expert reasoning traces or case presentations from later timesteps. (3) *Full-oracle prompting* augments the same case presentations used in cumulative-context prompting with expert reasoning traces from earlier timesteps $(t_1, \ldots, t_{i-1})$, while withholding expert reasoning

Table 3: Results under cumulative-context prompting on the VMR test split across five tasks, pooled over all evaluated timesteps. For each metric, the best mean score is in **bold** and the second-best is shaded. * denotes evaluation by an LLM judge. All scores are percentages, reported as means across three random seeds, with sample standard deviations shown as subscripts.

Metric	GPT	Claude	Grok	Kimi	DeepSeek	Llama	Qwen9B	Qwen27B	Lingshu	Huatuo	MedGemma
Task 1: Progressive Diagnostic Reasoning											
Top-1 DDx Acc (ICD-10)	**29.9** ± 1.4	28.1 ± 3.9	26.6 ± 1.2	26.0 ± 0.9	25.2 ± 3.0	13.9 ± 1.6	22.8 ± 1.7	23.0 ± 1.4	11.7 ± 1.1	17.3 ± 1.0	13.2 ± 1.0
Top-1 DDx Acc*	50.7 ± 1.9	**54.1** ± 4.5	45.5 ± 2.6	45.8 ± 2.7	46.3 ± 2.4	29.2 ± 2.7	41.2 ± 0.6	41.0 ± 1.8	31.3 ± 2.5	34.2 ± 1.9	29.1 ± 1.8
Top-3 DDx Acc*	64.3 ± 1.8	**69.8** ± 5.4	57.7 ± 2.1	59.3 ± 3.0	59.2 ± 2.0	41.6 ± 2.2	53.5 ± 0.4	54.1 ± 1.8	46.5 ± 1.9	42.9 ± 1.5	44.5 ± 3.0
Top-5 DDx Acc*	70.7 ± 1.1	**76.7** ± 4.9	63.5 ± 2.0	66.4 ± 2.5	64.8 ± 1.9	48.3 ± 2.1	60.5 ± 1.4	60.2 ± 0.2	56.0 ± 1.8	47.0 ± 1.9	52.6 ± 3.6
Clinical concept recall	24.7 ± 0.1	**35.1** ± 0.7	21.3 ± 0.3	26.3 ± 0.2	31.2 ± 0.7	17.6 ± 0.4	22.2 ± 0.6	27.1 ± 0.4	20.0 ± 0.6	21.6 ± 0.4	27.7 ± 1.2
Reasoning trace recall *	58.2 ± 0.9	**76.9** ± 0.1	50.8 ± 0.5	56.8 ± 0.5	58.4 ± 0.4	29.9 ± 0.6	49.0 ± 0.5	52.6 ± 0.4	32.9 ± 0.6	50.2 ± 1.0	42.7 ± 0.9
Reasoning comprehensiveness *	18.5 ± 0.2	**33.3** ± 0.5	14.3 ± 0.3	18.5 ± 0.8	20.1 ± 0.3	3.8 ± 0.6	11.8 ± 0.4	16.5 ± 1.0	5.0 ± 0.3	12.7 ± 1.6	9.6 ± 0.7
Task 2: Action Planning											
Clinical concept recall	20.4 ± 0.5	**26.7** ± 1.1	7.6 ± 0.2	10.6 ± 1.1	10.9 ± 0.7	8.7 ± 1.2	6.7 ± 0.2	10.5 ± 0.3	4.8 ± 0.3	11.6 ± 1.1	3.3 ± 1.4
Action recall*	53.3 ± 0.5	**67.9** ± 2.1	29.0 ± 0.9	37.3 ± 0.6	40.7 ± 0.3	16.2 ± 1.5	28.7 ± 1.0	35.6 ± 2.2	13.6 ± 0.4	34.9 ± 0.6	8.8 ± 4.6
Task 3: Hypothesis Tracking											
Added hypothesis recall*	26.2 ± 2.5	**37.1** ± 0.7	17.7 ± 1.3	19.2 ± 0.6	14.3 ± 1.0	13.6 ± 0.5	9.5 ± 0.6	12.7 ± 1.0	9.7 ± 3.7	18.9 ± 1.9	11.0 ± 1.8
Ruled-out hypothesis recall	60.4 ± 1.3	56.8 ± 4.2	64.0 ± 3.7	**68.1** ± 3.7	53.5 ± 4.8	33.3 ± 2.8	62.8 ± 6.2	58.6 ± 2.7	33.7 ± 0.5	62.1 ± 6.1	28.9 ± 3.9
Task 4: Clinical Evidence and Reasoning Typing											
Evidence type macro-F_1	**67.2** ± 1.5	67.6 ± 1.8	62.7 ± 2.8	61.2 ± 1.2	58.5 ± 1.3	17.6 ± 1.8	58.9 ± 1.0	62.8 ± 5.4	52.5 ± 2.8	61.4 ± 3.1	24.7 ± 1.5
Step type macro-F_1	**48.1** ± 5.2	42.6 ± 4.4	42.6 ± 2.5	41.5 ± 2.8	37.4 ± 3.5	13.6 ± 0.5	38.0 ± 5.7	33.9 ± 1.1	16.4 ± 0.7	31.9 ± 8.0	9.3 ± 0.2
Reasoning attributes macro-F_1	**40.6** ± 2.0	39.0 ± 1.7	39.7 ± 3.0	36.9 ± 2.4	30.0 ± 4.2	7.4 ± 1.4	30.4 ± 2.0	33.1 ± 2.9	6.4 ± 2.4	29.6 ± 1.3	0.3 ± 0.0
Task 5: Evidence–Hypothesis Relation Prediction											
Relation macro-F_1	**87.3** ± 0.5	77.7 ± 6.9	57.7 ± 0.7	57.4 ± 0.4	47.2 ± 4.7	64.5 ± 0.9	85.5 ± 0.8	84.7 ± 0.5	67.6 ± 0.7	82.2 ± 0.7	38.1 ± 3.3

traces at t_i and all later timesteps. For each task, all conditions use the same output schema and parsing rules defined in Sec. 4.2.

5.2 Results and Analysis

We use CLINTRACE to characterize model performance as clinical evidence accumulates and to assess the value of expert trajectories as training supervision. We first examine task-level performance, temporal patterns, and differences from documented expert reasoning across VMR and RLR. We then evaluate the effects of prompting conditions and supervised fine-tuning, including transfer to external diagnostic benchmarks.

Task-specific strengths and incomplete expert-reasoning coverage. Table 3 reports cumulative-context results pooled across evaluated timesteps on VMR. `Claude` achieves the highest mean top-5 diagnostic accuracy (76.7%) and reasoning trace recall (76.9%), but its reasoning comprehensiveness falls to 33.3% under the stricter coverage criterion. Model rankings also differ across tasks: `Qwen9B` exceeds `Claude` in evidence–hypothesis relation prediction (85.5% vs. 77.7% macro-F_1), whereas `Claude` recovers more expert-documented hypothesis additions (37.1% vs. 9.5%). These results motivate evaluating a profile of reasoning and decision-making tasks rather than relying on a diagnostic score alone.

Longer outputs do not consistently correspond to higher scores. Figure 3 compares aggregate VMR performance with mean generated reasoning length per timestep. `Claude` achieves the highest aggregate score (52.6%) with 837 tokens per timestep, while `GPT` scores 48.0% with 303 tokens. In contrast, `MedGemma` generates 493 tokens but achieves the lowest aggregate score (22.9%). Thus, longer outputs do not consistently correspond to better performance across the evaluated models. This cross-model comparison does not establish the effect of increasing the reasoning budget for an individual model.

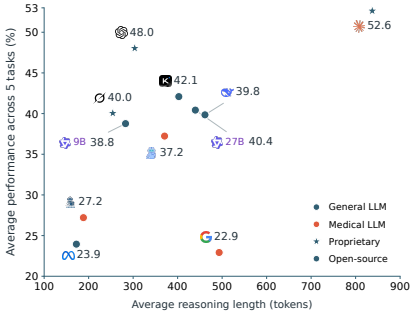


Figure 3: Overall VMR performance and model-generated clinical reasoning length across 11 LLMs. Clinical reasoning length is per timestep.

Performance varies across stages of evidence disclosure. As evidence accumulates, diagnostic accuracy generally improves across timesteps. Reasoning coverage often declines at earlier stages before recovering (Figures 14). In contrast, action planning and ruled-out hypothesis performance varies across timesteps (Figure 4). Additionally, at the final stage of each case, eight models exceed 90% top-5 diagnostic accuracy on VMR, whereas reasoning comprehensiveness remains much lower, ranging from 8.9% to 50.1% (Table 12). These results show that

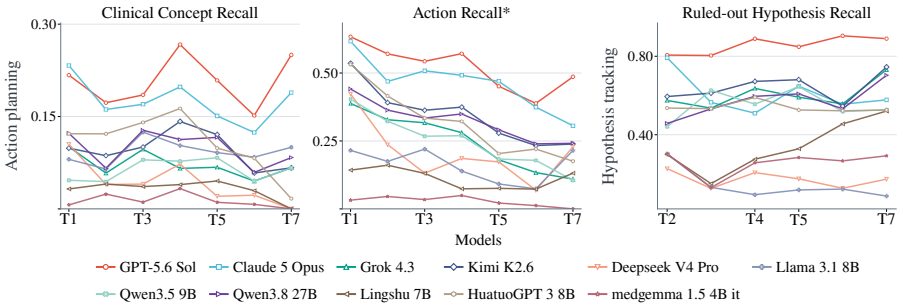


Figure 4: Per-timestep trajectories for clinical concept recall, action recall, and ruled-out hypothesis recall. T1–T7 mark timesteps that new information is revealed. Hypothesis tracking begins at T2 because it measures changes from the preceding timestep. *: LLM judge.

final-stage performance and timestep-level trajectories capture different aspects of model behavior. We interpret absolute-timestep comparisons descriptively because the set of contributing cases may differ across timesteps.

Diagnostic accuracy is lower on protected RLR cases for most models. In pooled cumulative-context evaluation, ICD-10 accuracy is lower on RLR than on VMR for eight of 11 models, while LLJ-based top-1 and top-5 accuracy each decrease for seven models (Tables 3 and 14). For example, top-5 accuracy falls from 76.7% to 70.0% for `Claude` and from 56.0% to 44.1% for `Lingshu`. In contrast, reasoning trace recall and reasoning comprehensiveness differ by at most 2.0 and 2.1 percentage points, respectively, across sources. These results suggest that RLR presents additional diagnostic challenges for many evaluated models, while incomplete coverage of expert reasoning persists on both sources. The comparison does not isolate the effects of case characteristics, discussion format, or prior exposure.

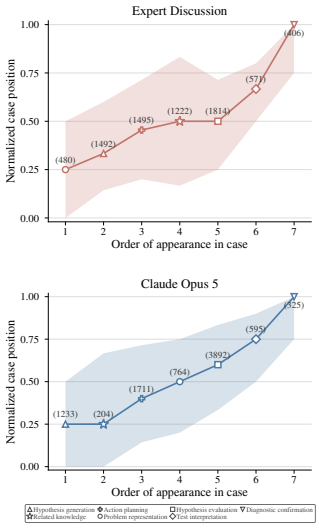


Figure 5: Temporal distributions of reasoning step types in expert discussions (top) and `Claude` outputs (bottom). Bands show the interquartile range.

Generated reasoning differs in composition and sequence. To characterize differences beyond reference coverage, we compare the distribution and normalized positions of reasoning step types in expert discussions and `Claude` outputs (Figure 5). Hypothesis evaluation accounts for 44.6% of model statements, compared with 24.3% of expert statements, while prior-knowledge statements account for 2.3% and 16.3%, respectively. Problem representations also appear later in the model's normalized reasoning sequence.

Prompting effects depend on the model and prediction target. Relative to final-stage cumulative-context prompting, full-case prompting changes diagnostic accuracy in a model- and source-dependent manner (Figure 6). On VMR, top-1 accuracy decreases for eight models, including a 7.1-point decrease for `Kimi`, whereas `Lingshu` gains 20.1 points in top-5 accuracy. On RLR, `Lingshu` gains 10.8 points in top-5 accuracy. Providing earlier expert reasoning through full-oracle prompting increases reasoning comprehensiveness for nine models on each source, but does not produce uniform gains in diagnosis. These findings distinguish performance with access to prior expert traces from performance based on case presentations alone. Experimental setup of final-stage evaluation is provided in App. B.1.

Expert trajectories provide useful, but task-dependent, supervision. Under cumulative-context evaluation, SFT increases reasoning trace recall from 30.4% to 42.7% for `Llama` and from 48.1% to 52.2% for `Qwen9B`. Added-hypothesis recall also increases, from 13.3% to 28.6% and from 10.1% to 21.2%, respectively (Table 15). Improvements are not uniform: relation macro-F_1 decreases for both models, and `Qwen9B` also declines in action recall and ruled-out-hypothesis recall. The one-seed experiments therefore demonstrate gains in selected prediction targets rather than a general improvement across all capabilities.

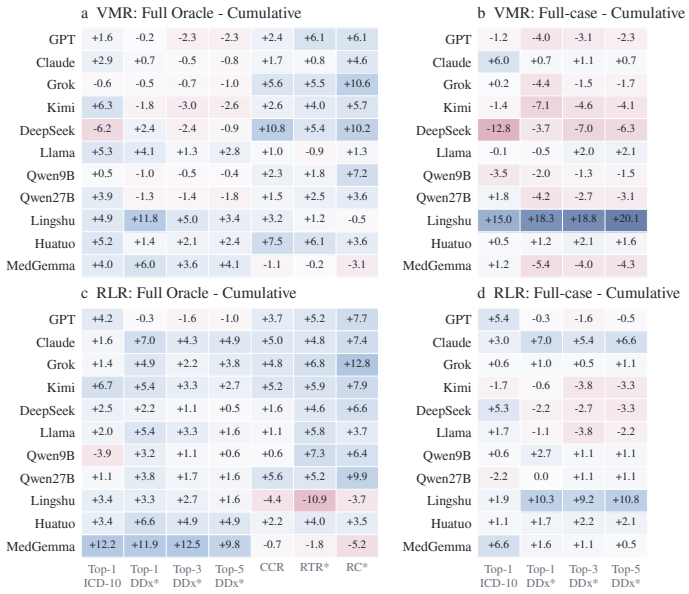


Figure 6: Endpoint performance differences, in percentage points, relative to cumulative-context prompting on VMR (top) and RLR (bottom). Left: full-oracle prompting in progressive diagnostic reasoning task. Right: full-case prompting for diagnosis. CCR: clinical concept recall; RTR: reasoning trace recall, RC: reasoning comprehensiveness.

Preliminary transfer to external diagnostic benchmarks. To assess whether the supervision benefits diagnosis beyond CLINTRACE, we evaluate the fine-tuned `Llama` model under full-case prompting on 100 randomly sampled cases from each of MedCaseReasoning (Wu et al., 2025) and ER-Reason (Molina et al., 2025) (Figure 7). These tasks assess diagnosis prediction from published clinical case reports and emergency-department records, respectively. Top-1/top-5 accuracy increases from 16%/40% to 19%/51% on MedCaseReasoning and from 19%/36% to 21%/41% on ER-Reason. These results provide preliminary evidence that supervision from CLINTRACE can benefit external diagnostic tasks, while broader evaluation is needed to establish the consistency of the gains.

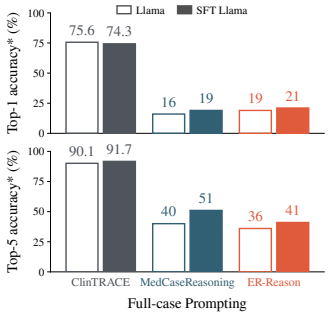


Figure 7: Top-1/5 diagnostic accuracy of `Llama` and SFT `Llama` on CLINTRACE, MedCaseReasoning, and ER-Reason on full-case prompting.

Illustrative case analysis. Table 17 compares expert traces with `Claude` outputs for an unfolding case of enterococcal endocarditis. Although the model reaches the correct diagnosis, earlier outputs introduce details not established by the available case presentation, such as "night-sweat-equivalent symptoms." Its hypothesis updates also differ from those documented by the experts. This example shows how stage-specific inspection can identify unsupported statements and reference-discordant updates that a final diagnosis score does not describe.

6 CONCLUSION

Final diagnosis accuracy alone does not capture how model reasoning evolves as evidence accumulates. We introduce CLINTRACE with time-aligned expert reasoning trajectories from live case discussions and five complementary tasks. Across 11 LLMs, strong diagnostic performance occurs alongside incomplete recovery of expert reasoning, hypothesis updates, and proposed actions. Longer outputs do not consistently yield better performance, and full-case prompting produces model-dependent changes in diagnostic accuracy. Fine-tuning on CLINTRACE improves reasoning coverage and added hypothesis recall, though gains are not uniform across tasks. Fine-tuned `Llama` also improves diagnosis on two external benchmarks, providing preliminary evidence of transfer. We will release the public VMR corpus while retaining RLR for protected evaluation to support research toward more reliable clinical AI systems.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript language polishing, LLM judging, and restructure the VMR and RLR discussion for CLINTRACE construction. We have not used generative AI tools for research design, experiment implementation, and synthetic data generation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study is approved by our Institutional Review Board and conducted with permission from the CPSolvers. Source cases are de-identified, and data use and release follow the CPSolvers license. The public release contains the VMR component, while the subscription-only RLR cases remain non-public and are used only for protected evaluation. CLINTRACE is intended for research on clinical reasoning and model evaluation, not for direct clinical decision-making.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of dataset construction, temporal segmentation, annotation schema, data annotation, model configurations, prompting conditions, and LLM judge settings in the main paper and appendices.

REFERENCES

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Xi Chen, Hanyu Zhou, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Ziyuan Qin, Hairui Li, Xue Zhang, Yingman Guo, Songlin Li, Yunan Hu, Qianjiang Xiong, Rongxiao Li, Lei Fan, Qicheng Lao, Weili Fu, Jian Li, and Kang Li. Grounding large language models in clinical diagnostics. *Nature Communications*, 17(1):4401, March 2026. doi: 10.1038/s41467-026-70274-w. URL https://doi.org/10.1038/s41467-026-70274-w.

Christopher Chiu, Silviu Pitis, and Mihaela van der Schaar. Simulating viva voce examinations to evaluate clinical reasoning in large language models. *arXiv preprint arXiv:2510.10278*, 2025.

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

Selin S. Everett, Bryan J. Bunning, Priyank Jain, Ivan Lopez, Anup Agarwal, Manisha Desai, Robert Gallo, Ethan Goh, Vinay B. Kadiyala, Zahir Kanjee, Jacob M. Koshy, Andrew Olson, Adam Rodman, Kevin Schulman, Eric Strong, Jonathan H. Chen, and Eric Horvitz. From tool to teammate in a randomized controlled trial of clinician–AI collaborative workflows for diagnosis. *npj Digital Medicine*, 9(1):409, March 2026. doi: 10.1038/s41746-026-02545-1. URL https://doi.org/10.1038/s41746-026-02545-1.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher D. Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22021–22030, 2024. Preprint arXiv:2308.14089, 2023.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263, 2025.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, October 2024. doi: 10.1001/jamanetworkopen.2024.40969. URL https://doi.org/10.1001/jamanetworkopen.2024.40969.

Ethan Goh, Robert J. Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A. Freed, Joséphine A. Cool, Zahir Kanjee, Kathleen P. Lane, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Jason Hom, Jonathan H. Chen, and Adam Rodman. GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nature Medicine*, 31(4):1233–1238, February 2025. doi: 10.1038/s41591-024-03456-y. URL https://doi.org/10.1038/s41591-024-03456-y.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, July 2024. doi: 10.1038/s41591-024-03097-1. URL https://doi.org/10.1038/s41591-024-03097-1.

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

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0908. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/32b80425554e081204e5988ab1c97e9a-Abstract-Conference.html.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael R. Lyu, and Wenting Chen. MedChain: Bridging the gap between LLM agents and clinical practice with interactive sequence. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2498. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6bc02900c74234aa5a0c1a7148ed933f-Abstract-Datasets_and_Benchmarks_Track.html.

Yunsong Liu, Zunamys I. Carrero, Xiaofeng Jiang, Dyke Ferber, Georg Wölflein, Li Zhang, Sanddhya Jayabalan, Tim Lenz, Zhouguang Hui, and Jakob Nikolas Kather. Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digit. Medicine*, 9(1), 2026. doi: 10.1038/S41746-026-02443-6. URL https://doi.org/10.1038/s41746-026-02443-6.

Liam G. McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel C. K. Tan, Kevin Tan, Thomas A. Buckley, Peter G. Brodeur, Leo A. Celi, Arjun K. Manrai, A. Humbert, and Adam Rodman. Assessment of large language models in clinical reasoning: A novel benchmarking study. *NEJM AI*, 2(10), 2025. doi: 10.1056/AIdbp2500120.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457, 2025. doi: 10.1038/s41586-025-08869-4.

Mel Molina, Nikita Mehandru, Niloufar Golchini, and Ahmed Alaa. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *PhysioNet*, 2025. doi: 10.13026/55s7-3c27.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task*, pp. 319–327, 2019.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P. Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025. doi: 10.48550/arXiv.2506.22405. URL https://arxiv.org/abs/2506.22405.

Geoffrey R Norman and Kevin W Eva. Diagnostic error and clinical reasoning. *Medical education*, 44(1):94–100, 2010.

Anil Palepu, Valentin Liévin, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S. Sara Mahdavi, Joëlle Barral, Dale R. Webster, Katherine Chou, Avinatan Hassidim, Yossi Matias, James Manyika, Ryutaro Tanno, Vivek Natarajan, Adam Rodman, Tao Tu, Alan Karthikesalingam, and Mike Schaekermann. Towards conversational AI for disease management. *arXiv preprint arXiv:2503.06074*, 2025.

Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Matthias Keicher, and Nassir Navab. Ehr2path: Comprehensive pathway-level modeling of longitudinal patient trajectories from multimodal electronic health records. 2026. URL https://arxiv.org/abs/2506.04831.

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, et al. Vibevoice-asr technical report. *arXiv preprint arXiv:2601.18184*, 2026.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.

Ihsan Ayyub Qazi, Ayesha Ali, Asad Ullah Khawaja, Muhammad Junaid Akhtar, Ali Zafar Sheikh, and Muhammad Hamad Alizai. Large language model diagnostic assistance for physicians in a lower-middle-income country: A randomized controlled trial. *Nature Health*, 1(2):198–205, February 2026. doi: 10.1038/s44360-025-00007-8. URL https://doi.org/10.1038/s44360-025-00007-8.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16(1):9799, November 2025. doi: 10.1038/s41467-025-64769-1. URL https://doi.org/10.1038/s41467-025-64769-1.

Nicholas Rounding, Luthfi Saiful Arif, Janine Berg, Jochen Cals, Diederik De Boer, Eefje De Bont, Sander Dijksman, Ardi Findyartini, Didier Fouarge, Marie-Christine Fregin, Pawel Gmyrek, Nadia Greviana, Ralph Leijenaar, Soraiya Manji, Annastacia Mbithi, Norah Obungu, Arierta Pujitresnani, Roselyter Riang'a, Diantha Soemantri, Sairabanu Mohamed Rashid Sokwalla, Sanne Steens, Lucia Velasco, Ardy Wildan, Prasandhya Astagiri Yusuf, and Mark Levels. Impact of LLM assistance on physician decision-making: A multi-country randomized controlled trial. *npj Digital Medicine*, September 2026. doi: 10.1038/s41746-026-03111-5. URL https://doi.org/10.1038/s41746-026-03111-5.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, Michael Moor, et al. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 9, 2026. doi: 10.1038/s41746-026-02674-7.

Md Kamrul Siam, Angel Varela, Md Jobair Hossain Faruk, Jerry Q. Cheng, Huanying Gu, Abdullah Al Maruf, and Zeyar Aung. Benchmarking large language models on the united states medical licensing examination for clinical reasoning and medical licensing scenarios. *Scientific Reports*, 16:1387, 2026.

Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. *Medical decision making*. John Wiley & Sons, 2024.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. DiReCT: Diagnostic reasoning for clinical notes via large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/892850bf793e03b5c410dfd9425b94c8-Abstract-Datasets_and_Benchmarks_Track.html.

Xidong Wang, Shuqi Guo, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, and Benyou Wang. Liveclin: A live clinical benchmark without leakage. In *International Conference on Learning Representations*, volume 2026, pp. 97981–98011, 2026.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam Shah. EHRSHOT: an EHR benchmark for few-shot evaluation of foundation models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d42db1f74df54cb992b3956eb7f15a6f-Abstract-Datasets_and_Benchmarks.html.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports, 2025. URL https://arxiv.org/abs/2505.11733.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. MedQA-CS: Objective structured clinical examination (osce)-style benchmark for evaluating LLM clinical skills. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 6183–6257, 2026. Preprint arXiv:2410.01553, 2024.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. Automating expert-level medical reasoning evaluation of large language models. *npj Digit. Medicine*, 9(1), 2025. doi: 10.1038/S41746-025-02208-7. URL https://doi.org/10.1038/s41746-025-02208-7.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zuo25a.html.

CONTENTS OF APPENDIX

A Details of CLINTRACE 15
A.1 Dataset Construction 15
A.2 Dataset Schema 15
A.3 Dataset Statistics and Label Distribution 15
A.4 Case Taxonomy 16
A.5 Provenance of VMR and RLR Reasoning Trajectories 17
A.6 Dataset Quality Verification 17
A.7 Data Quality Evaluation Platform 18

B Experimental Results 19
B.1 Experimental Setup 19
B.2 Final-Timestep Results 22
B.3 RLR Cumulative-Context Prompting Results 23
B.4 Timestep-wise Performance 24
B.5 Supervised Fine-tuning Results 25
B.6 LLM-as-a-Judge Validation 25

C Case Study 27

D Prompt Templates 29
D.1 Downstream Tasks Prompts 29
D.2 LLM-as-a-Judge Prompts 32

Table 4: Dataset schema of CLINTRACE.

Field	Description
Case	
`case_id`	Unique case identifier.
`gold_diagnosis`	Reference final diagnosis.
`reasoning_trajectory`	Ordered records at timestamps $t_1, \ldots, t_n$.
Timestep	
`timestep`	Temporal index of the presentation.
`case_presentation`	Information available at the current timestamp.
`observed_evidence`	Spans with id, text, and type.
`reasoning_steps`	Reasoning with step type, reasoning attributes, and hypothesis updates.
`evidence_hypothesis_relations`	Evidence IDs, hypothesis, relation type, and justification.
`requested_actions`	Action, action type, and purpose.
`cumulative_active_hypotheses`	Differentials still active at the timestamp.
`case_include_final_dx`	Whether the presentation reveals the final diagnosis.

A DETAILS OF CLINTRACE

A.1 DATASET CONSTRUCTION

We apply a speech-to-text model to transcribe audio files into speaker-identified clean transcripts. Specifically, after downloading all audio files, we employ `pyannote/speaker-diarization-community-1`[1] (Plaquet & Bredin, 2023) as a speaker diarization model to record all unique speakers in each discussion for the ones that beyond the 1-hour long case. Given the speech-to-text model `microsoft/VibeVoice-ASR-HF`[2] only accept input less than 1 hour, we chunk the audio files into maximum two parts for audios that beyond 1-hour limit. We need to guarantee which speaker reveals the case presentation throughout the discussion, therefore, directly chunking the audio into multiple parts cannot memorize which speaker is speaking in different chunks. `VibeVoice-ASR-HF` is a multilingual speech-to-text model designed to process 60-minute long-form audio in a single pass, generating structured transcriptions containing speaker, timesteps, and content (Peng et al., 2026). We use `GPT-5.4`[3] for VMR and RLR data extraction.

A.2 DATASET SCHEMA

We label our dataset into multiple fine-grained categories across different fields. As depicted in Table 4, each case is split into multiple timesteps, with extracted clinical evidence, reasoning trajectories, actions, and cumulative hypotheses. We report the detailed taxonomy of our extracted data in Table 5, including clinical evidence type, reasoning step type, reasoning attributes, evidence-hypothesis relation, and requested action type.

A.3 DATASET STATISTICS AND LABEL DISTRIBUTION

As shown in Table 6, we split the VMR data into train/validation/test sets and use the test set for evaluation. The RLR data statistics are reported in the final column, with 184 evaluation cases and slightly more average reasoning timesteps (6.0 per case) over all cases. Figure 8 shows our curated dataset label distributions among different categories. In our experiments, we filter out all the traces labeled as "Other" in step types given those traces are irrelevant to the case.

[1] `https://huggingface.co/pyannote/speaker-diarization-community-1`

[2] `https://huggingface.co/microsoft/VibeVoice-ASR-HF`

[3] Model version: `gpt-5.4-2026-03-05`, `https://developers.openai.com/api/docs/models/gpt-5.4`

Table 5: Taxonomies used to structure the extracted longitudinal clinical reasoning trajectories. Each observed evidence item is assigned exactly one evidence type; each reasoning step is assigned exactly one step type and one or more reasoning attributes. Evidence–hypothesis relations and requested actions use their respective controlled label sets.

Label	Definition
Evidence type *(one per observed evidence item)*	
`demographic`	Age, sex, race, ethnicity, and other identifying characteristics.
`patient_history`	Prior medical, surgical, social, or personal history (incl. pertinent negatives).
`family_history`	Familial medical history.
`clinical_symptom`	Reported symptoms or complaints, including pertinent negative symptoms.
`physical_evaluation`	Physical examination findings and signs.
`laboratory_testing`	Laboratory values and test results (e.g., chemistry, troponin).
`pharmacologic_therapy`	Medications the patient is taking, was started or stopped on, or was given.
`diagnostic_testing`	Non-imaging studies and results (e.g., ECG, echocardiogram, endoscopy, biopsy).
`imaging`	Imaging studies and results (e.g., CT, MRI, X-ray, ultrasound).
`other`	Evidence that fits none of the above categories.
Reasoning step type *(one per reasoning step)*	
`problem_representation`	Concisely summarizes presenting problems, key findings, and context.
`hypothesis_generation`	Formulates possible diagnostic hypotheses from the available case information.
`hypothesis_evaluation`	Assesses or compares hypotheses against evidence (rule-in, rule-out, pattern).
`interpretation_of_test`	Interprets diagnostic test results to inform ongoing reasoning.
`diagnostic_confirmation`	Establishes a working or final diagnosis by synthesizing evidence and reasoning.
`action_planning`	Proposes next clinical actions: tests, treatment, monitoring, or follow-up.
`prior_knowledge`	General disease knowledge: illness scripts, epidemiology, mechanisms, teaching.
`other`	Reasoning that fits none of the above categories.
Reasoning attributes *(one or more per reasoning step)*	
`history_extraction`	Interprets or extracts key history elements already presented in the case.
`history_inquiry`	Requests more history (e.g., prior episodes, medications, social/family history).
`physical_exam_interpretation`	Interprets physical examination findings.
`lab_results_analysis`	Interprets laboratory results.
`imaging_interpretation`	Interprets imaging findings.
`risk_factor_identification`	Identifies risk factors present in the available clinical information.
`fast_pattern_recognition`	Recognizes an illness pattern directly, without extensive analytic reasoning.
`slow_pattern_recognition`	Recognizes an illness pattern through extensive analytic reasoning.
`diagnostic_foothold`	Identifies a key clue that anchors subsequent diagnostic reasoning.
`hypothesis_revision`	Updates previously considered diagnostic hypotheses.
`parallel_hypothesis_tracking`	Considers or tracks multiple diagnostic hypotheses simultaneously.
`uncertainty_articulation`	Explicitly expresses diagnostic uncertainty or a diagnostic dilemma.
`test_selection`	Selects a specific laboratory, imaging, biopsy, or other test to obtain next.
`treatment_plan`	Recommends a therapy, procedure, supportive care, monitoring, or follow-up.
`other`	Reasoning attribute that fits none of the above categories.
Evidence–hypothesis relation *(one per evidence–hypothesis link)*	
`supporting`	Evidence directly supports the hypothesis.
`weakened`	Evidence lowers the likelihood of the hypothesis but does not rule it out.
`refuting`	Evidence effectively rules out the hypothesis.
`neutral`	Evidence is relevant but neither clearly supports nor weakens the hypothesis.
Requested action type *(one per requested action)*	
`clinical_history`	Requests additional history from the patient.
`laboratory_testing`	Requests laboratory testing.
`electrocardiogram`	Requests an ECG/EKG.
`imaging`	Requests an imaging study (e.g., ultrasound, CT, MRI, X-ray).
`physical_exam`	Requests a physical examination maneuver or finding.
`other`	Requested action that fits none of the above categories.

The data statistics of RLR data is reported in Figure 9. Overall, RLR data share the similar label distribution on our extracted trajectories, showing the consistency and high quality of our curated benchmark.

A.4 Case Taxonomy

To understand the general body system distribution of our curated dataset, we select which body parts the case involves through an LLM from MedlinePlus (e.g., Blood/Heart, Brain/Nerves). Additionally, each case will have one disease type and multiple age, sex, and any named population labels from MedlinePlus labeled by the LLM. Following MedR-Bench (Qiu et al., 2025), we further categorize the VMR cases into rare disease. Specifically, we apply a two-stage rare disease categorization strategy adapted from a previous study. First, the disease name in each case must match a

Table 6: Summary statistics of the curated intermediate clinical reasoning benchmark. We split VMR into train/validation/test sets with the same ratio. RLR serves as a proprietary test set. Values are reported as mean ± standard deviation. Token counts use the OpenAI `tiktoken` encoding `o200k_base`. Case presentation length is measured per timestamp; diagnostic reasoning length is measured per reasoning step and after summing steps within each timestamp.

Statistic	Train	Validation	Test	Overall	RLR
Dataset size					
Cases	1,249	84	131	1,464	184
Reasoning trajectory timesteps	6,643	464	767	7,874	1,103
Diagnostic reasoning steps	71,701	4,822	7,521	84,044	8,888
Per-case statistics					
Timesteps	5.3 ± 1.6	5.5 ± 1.3	5.9 ± 1.6	5.4 ± 1.5	6.0 ± 2.1
Diagnostic reasoning steps	57.4 ± 16.1	57.4 ± 15.8	57.4 ± 16.1	57.4 ± 15.9	48.3 ± 12.0
Observed evidence items	104.6 ± 41.9	106.6 ± 47.0	105.8 ± 37.7	104.7 ± 41.4	69.1 ± 32.4
Evidence–hypothesis relations	43.8 ± 15.4	45.0 ± 15.8	44.6 ± 16.4	44.0 ± 15.5	22.6 ± 8.5
Requested action items	15.6 ± 7.8	16.4 ± 7.2	16.6 ± 8.0	15.8 ± 7.7	13.8 ± 5.2
Mentioned hypotheses per case	44.9 ± 21.7	47.6 ± 20.9	46.1 ± 20.1	45.3 ± 21.5	40.2 ± 16.6
Token-length statistics					
Case presentation / timesteps	158.1 ± 106.9	160.9 ± 110.4	155.5 ± 105.7	165.2 ± 107.2	111.9 ± 79.4
Case presentation / case	887.9 ± 377.9	890.9 ± 367.2	890.2 ± 336.3	888.5 ± 372.9	670.7 ± 313.2
Diagnostic reasoning / step	33.3 ± 13.0	33.9 ± 13.7	33.7 ± 13.4	33.4 ± 13.1	34.4 ± 13.3
Diagnostic reasoning / timesteps	359.0 ± 166.4	353.2 ± 175.5	331.0 ± 169.5	356.3 ± 167.8	277.0 ± 161.6
Diagnostic reasoning / case	1911.4 ± 542.8	1946.9 ± 471.1	1935.2 ± 552.2	1916.5 ± 535.9	1660.3 ± 426.0

disease listed in Orphanet[4], a rare disease database. Second, an LLM (i.e., `GPT-5.4`) must confirm that the case is indeed about that disease. If either criterion is not met, the case is categorized as "not rare." As shown in Figure 10, the case taxonomy distribution shows that the men patients take over 50% for all sets, with 126 rare disease cases. The taxonomy of RLR data is shown in Figure 11, while no cases are classified as rare disease using our current grouping protocol.

Following MedR-Bench (Qiu et al., 2025), we categorize the curated cases into different medical aspects. Specifically, we first classify each curated case based on the body locations/systems, disorders and conditions, and demographic groups using MedlinePlus[5]. Following the prompts used in MedR-Bench, we further include Demographic Groups as another dimension. Cases that do not fit into the pre-defined categories will be labeled as "Other." In Figure 10, we report the distribution of the case taxonomy across train, validation, and test splits on the VMR data, including body system, disorder and condition, patient demographic group, and rare-disease relevance.

A.5 Provenance of VMR and RLR Reasoning Trajectories

While Sec. 3 describes the provenance of VMR and RLR separately, this section directly compares their clinical content and reasoning structure. VMR is curated from the CPSolvers' multi-participant Virtual Morning Report sessions (a case presenter plus multiple physician discussants), whereas RLR is curated from a subscription podcast in which two board-certified physicians discuss each case. We characterize how this difference in discussion format is reflected in (i) the clinical topics discussed and (ii) the structure of the reasoning process itself.

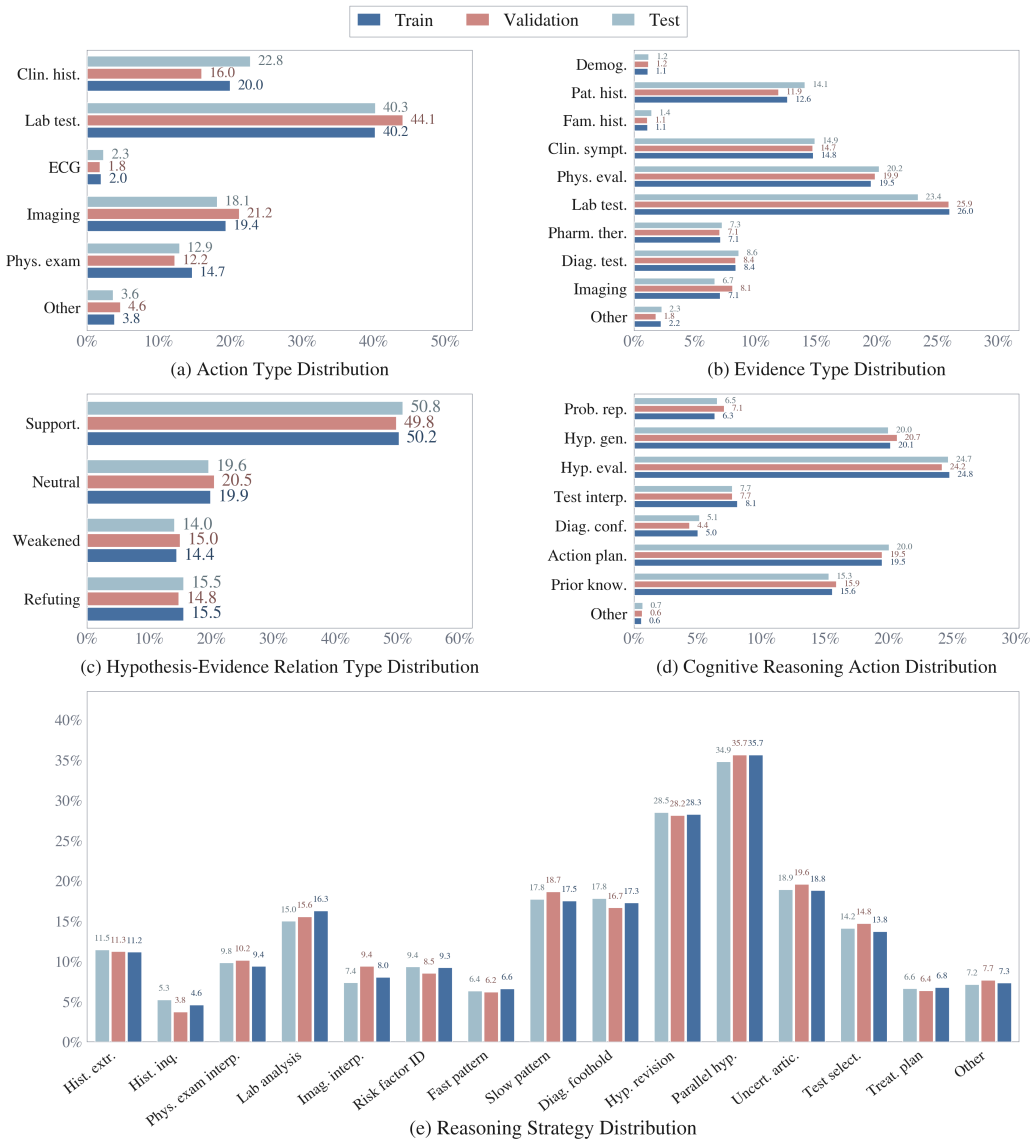
A.6 Dataset Quality Verification

As we introduced in Sec. 3, we recruit 12 medical experts to evaluate the data quality of our curated benchmark. We randomly sample 48 VMR cases, including 6 cases independently annotated by two physicians for inter-annotator agreement (IAA) calculation and 42 cases evaluated by a single expert (10 physicians in total, each is compensated with a $15 Amazon gift card). We provide the annotation questions in Table 7.

Across 1,532 annotation units in the six dual-annotated IAA cases, annotators achieve 95.7% raw agreement, with a pooled annotation accuracy of 97.7%. Across the remaining 9,450 annotation

[4] `https://www.orpha.net/`

[5] `https://medlineplus.gov/healthtopics.html`



(a) Action Type Distribution

(b) Evidence Type Distribution

(c) Hypothesis-Evidence Relation Type Distribution

(d) Cognitive Reasoning Action Distribution

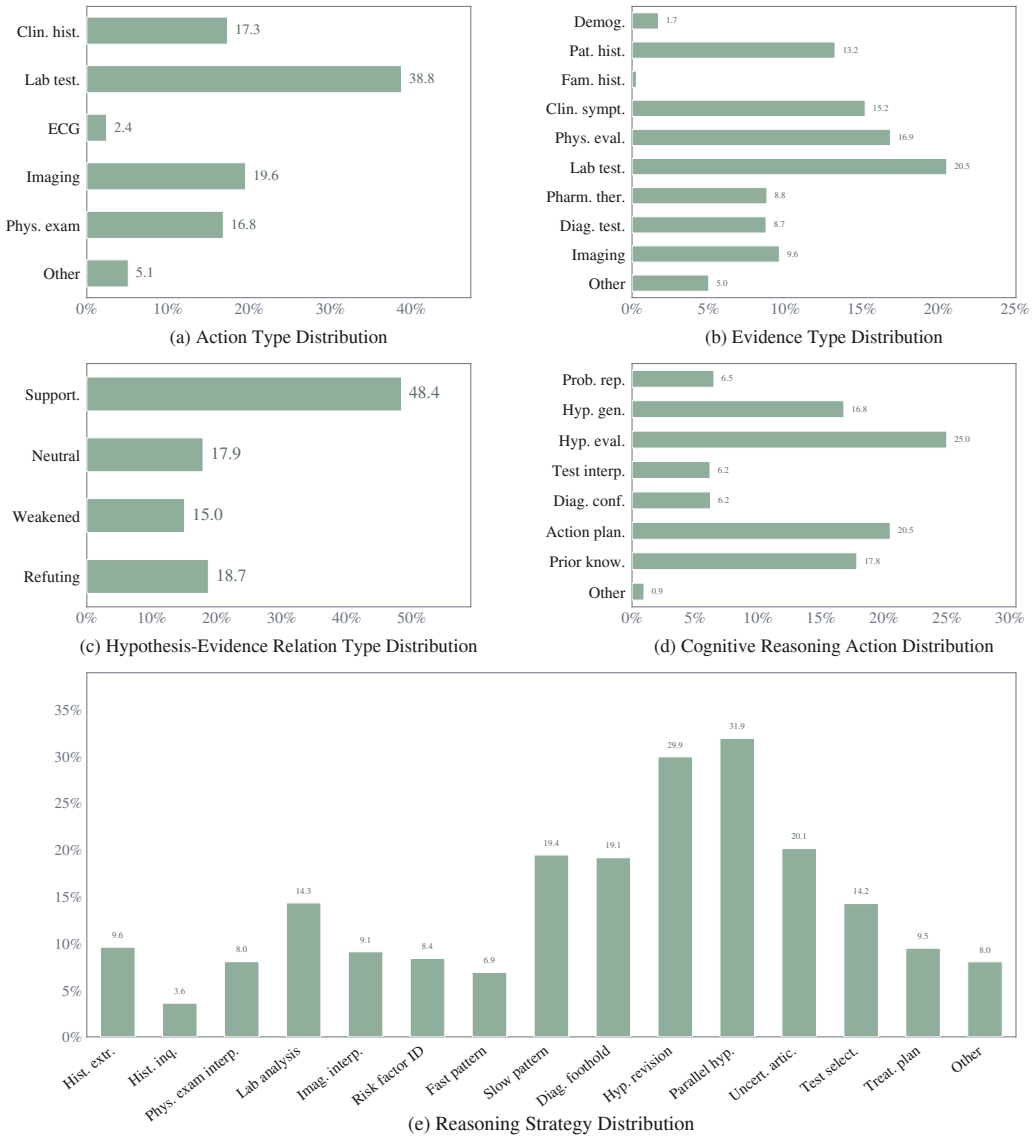
(e) Reasoning Strategy Distribution

Figure 8: Within-split label distributions for the VMR data. (a) requested action types; (b) observed evidence types; (c) evidence–hypothesis relation types; (d) cognitive reasoning step types; (e) fine-grained reasoning strategy tags. Colors indicate train, validation, and test splits.

units from 42 single-annotated cases, pooled accuracy is 96.6%, with most fields exceeding 95% accuracy (Table 8–Table 11). These results indicate that the extracted reasoning trajectories closely reflect the source discussions and that annotator agreement is consistently high.

A.7 Data Quality Evaluation Platform

To validate the quality of our curated data, we develop a human evaluation protocol to validate our curated data quality. As shown from Figure 12 to Figure 13, we present the main annotation interface of both VMR data and LLM-judges validation tasks.



(a) Action Type Distribution

(b) Evidence Type Distribution

(c) Hypothesis-Evidence Relation Type Distribution

(d) Cognitive Reasoning Action Distribution

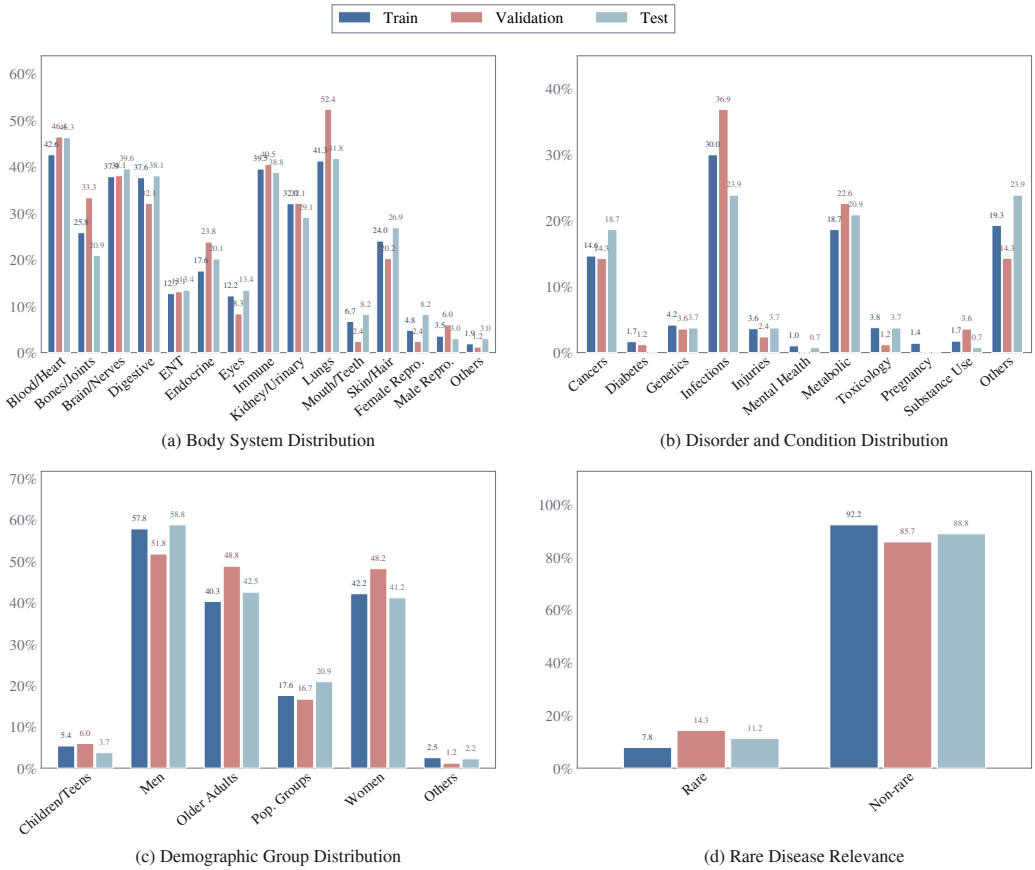
(e) Reasoning Strategy Distribution

Figure 9: Overall label distributions for the RLR data. (a) requested action types; (b) observed evidence types; (c) evidence–hypothesis relation types; (d) cognitive reasoning step types; (e) fine-grained reasoning strategy tags.

B Supplementary Experiment Results

B.1 Experiment Details

Evidence type and step type typing are multi-class classification tasks, while reasoning tag typing is a multi-label classification task. The final timestep evaluation only support progressive diagnostic reasoning because it produces timestep-specific predictions conditioned on the accumulated case state without further actions or hypothesis. In contrast, clinical evidence and reasoning typing and evidence–hypothesis relation prediction operate on individual statements or relations and therefore do not have a natural case-level endpoint. Full-case prompting supports only endpoint evaluation because the complete case is presented in a single step, leaving no endpoint-level trajectory for reasoning evaluation (Figure 6). For the supervised fine-tuning on `Llama` and `Qwen9B`, due to



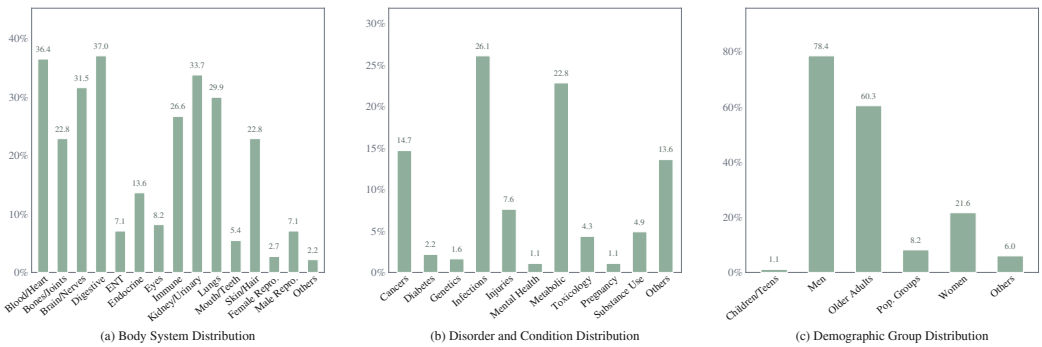
(a) Body System Distribution

(b) Disorder and Condition Distribution

(c) Demographic Group Distribution

(d) Rare Disease Relevance

Figure 10: Case taxonomy distributions across train, validation, and test splits of VMR data. **(a)** Body system, **(b)** disorder and condition, **(c)** demographic group, and **(d)** rare-disease relevance. Categories in (a–c) follow MedlinePlus Health Topics and are assigned by `GPT-5.4` from each case's final diagnosis and case presentations. Rare-disease labels in (d) were obtained via SciSpaCy/UMLS–Orphanet Rare Disease Ontology (ORDO) candidate matching and GPT verification. Bars indicate split-specific percentages of cases.



(a) Body System Distribution

(b) Disorder and Condition Distribution

(c) Demographic Group Distribution

Figure 11: Case taxonomy distributions of RLR data. **(a)** Body system, **(b)** disorder and condition, and **(c)** patient demographic group. Categories in (a–c) follow MedlinePlus Health Topics and are assigned by `GPT-5.4` from each case's final diagnosis and case presentations.

Table 7: Physician annotation protocol for evaluating extracted clinical reasoning trajectories of VMR. All questions are answered as *Yes*/*No*. Questions Q3–Q6 are evaluated for each extracted reasoning step, and Q1–Q2 and Q7–Q11 are evaluated per trajectory.

ID	Annotation level	Evaluated field	Question shown to physician annotators
Observed clinical evidence			
Q1	Timestamp	Observed evidence	Is the observed evidence correctly decomposed into atomic, faithful clinical facts from the case presentation?
Q2	Timestamp	Evidence type	Is the evidence-type label assigned to each observed evidence item correct?
Clinical reasoning steps			
Q3	Reasoning step	Reasoning fidelity	Is the extracted reasoning faithful to the source clinical discussion and reasonably inferable from it?
Q4	Reasoning step	Step type	Is the step type assigned to the reasoning step correct?
Q5	Reasoning step	Reasoning attributes	Are the reasoning attributes assigned to the reasoning step correct?
Q6	Reasoning step	Hypothesis extraction	Are the current hypotheses and ruled-out hypotheses correctly extracted from the diagnostic reasoning?
Structured trajectory fields			
Q7	Relation instance	Evidence–hypothesis relation	Is each evidence–hypothesis relation correctly linked, including its relation type and supporting reason?
Q8	Action instance	Requested action	Is each requested action aligned with the corresponding action planning reasoning?
Q9	Timestamp	Active hypotheses	Does cumulative active hypotheses accurately reflect the hypotheses that remain active in the discussion up to this timestamp?
Case-level quality			
Q10	Case	Hallucination	Is the extracted output free of information unsupported by the source discussion?
Q11	Case	Temporal organization	Are the extracted clinical reasoning traces organized according to the correct temporal sequence of the source discussion?

Table 8: Overall statistics of the human evaluation annotations, grouped by annotation level and data field. The human inter-annotator agreement (IAA) subset contains six dual-annotated cases, while the remaining 42 cases were single-annotated.

Level	Field	IAA (6 cases)	Single (42 cases)	Total
Case	Hallucination free	6	38	44
Case	Time course correct	6	42	48
Timestamp	Observed evidence correct	21	90	111
Timestamp	Observed evidence type correct	20	90	110
Timestamp	Cumulative active hypotheses correct	21	84	105
Item	Evidence–hypothesis relations	186	712	898
Item	Requested actions	36	294	330
Step	Faithful to discussion	309	2,025	2,334
Step	Step type correct	309	2,025	2,334
Step	Tags correct	309	2,025	2,334
Step	Hypotheses extraction correct	309	2,025	2,334
Total		**1,532**	**9,450**	**10,982**

computing constrains, we only run the evaluation on one seed of VMR data. For external benchmark evaluation as depicted in Figure 7, we randomly select 100 cases from both MedCaseReasoning (Wu

Table 9: Inter-annotator agreement (IAA) on the six dual-annotated cases. Agreement denotes the raw proportion of agreement, and AC1 denotes Gwet's agreement coefficient. CI: confidence interval.

Level	Field	n	Agreement	95% CI
Timestamp	Cumulative active hypotheses correct	21	1.000	1.000
Timestamp	Observed evidence correct	21	1.000	1.000
Step	Faithful to discussion	309	0.990	[0.971, 1.000]
Step	Step type correct	309	0.990	[0.982, 0.997]
Step	Tags correct	309	0.977	[0.961, 0.993]
Item	Evidence–hypothesis relations	186	0.952	[0.913, 0.979]
Timestamp	Observed evidence type correct	20	0.950	[0.826, 1.000]
Item	Requested actions	36	0.944	[0.826, 1.000]
Step	Hypotheses extraction correct	309	0.876	[0.760, 0.942]
Case	Hallucination free	6	0.833	[0.200, 1.000]
Case	Time course correct	6	0.833	[0.200, 1.000]
Pooled		**1,532**	**0.957**	**[0.940, 0.969]**

Table 10: Field-level annotation accuracy for the 42 single-annotated cases. Accuracy is reported with the 95% Wilson score confidence interval.

Level	Field	n judged	Yes	No	Accuracy [95% CI]
Step	Faithful to discussion	2,025	2,016	9	99.6% [99.2, 99.8]
Step	Step type correct	2,025	2,004	21	99.0% [98.4, 99.3]
Timestamp	Observed evidence correct	90	89	1	98.9% [94.0, 99.8]
Timestamp	Observed evidence type correct	90	88	2	97.8% [92.3, 99.4]
Case	Hallucination free	38	37	1	97.4% [86.5, 99.5]
Step	Tags correct	2,025	1,965	60	97.0% [96.2, 97.7]
Item	Requested actions	294	284	10	96.6% [93.9, 98.1]
Item	Evidence–hypothesis relations	712	682	30	95.8% [94.0, 97.0]
Case	Time course correct	42	40	2	95.2% [84.2, 98.7]
Timestamp	Cumulative active hypotheses correct	84	79	5	94.0% [86.8, 97.4]
Step	Hypotheses extraction correct	2,025	1,846	179	91.2% [89.8, 92.3]
Pooled		**9,450**	**9,130**	**320**	**96.6% [96.2, 97.0]**

et al., 2025) and ER-Reason (Molina et al., 2025) for evaluation, while CLINTRACE evaluation uses full VMR test set. We report top-1 and top-5 diagnosis accuracy using LLJ for evaluation.

For the SFT setyp, we perform multi-task supervised fine-tuning of `Llama-3.1-8B-Instruct` and `Qwen3.5 9B` using a shared pool of examples covering all five downstream tasks. Each example pairs a task-specific instruction with a structured reference response, and training minimized next-token cross-entropy over response tokens while masking prompt and padding tokens. We use LoRA with rank 32, scaling factor 64, and dropout 0.05, together with AdamW, a learning rate of 2e-4, linear learning-rate decay, and BF16 precision. Both models use an effective batch size of 32: `Llama` uses a per-device batch size of 16 with two gradient-accumulation steps, whereas `Qwen` uses 8 with four accumulation steps. Training was configured for up to two epochs, with aggregate and per-task validation, checkpointing every 200 steps, and early-stopping patience of three evaluations. Both runs ended at step 1,400, approximately 1.63 epochs, and restored checkpoint 800. The selected checkpoints achieve aggregate validation losses of 1.188 and 0.884 for `Llama` and `Qwen`, respectively. All experiments are conducted using maximum 2 A100 GPUs.

B.2 FINAL-TIMESTEP RESULTS

VMR cumulative-context prompting final timestep results. Other than the pooled results over timesteps, we also report the final timestep performance across 11 models and analyze the perfor-

Table 11: Field-level pooled annotation accuracy for the six dual-annotated IAA cases.

Level	Field	n units	Pooled accuracy	95% CI
Timestamp	Cumulative active hypotheses correct	21	100.0%	[91.6, 100.0]
Step	Faithful to discussion	309	99.5%	[98.6, 99.8]
Step	Step type correct	309	99.5%	[98.6, 99.8]
Step	Tags correct	309	98.9%	[97.7, 99.5]
Item	Evidence–hypothesis relations	186	97.6%	[95.5, 98.7]
Timestamp	Observed evidence type correct	20	97.5%	[87.1, 99.6]
Item	Requested actions	36	97.2%	[90.4, 99.2]
Timestamp	Observed evidence correct	21	95.2%	[84.2, 98.7]
Step	Hypotheses extraction correct	309	93.5%	[91.3, 95.2]
Case	Hallucination free	6	91.7%	[64.6, 98.5]
Case	Time course correct	6	91.7%	[64.6, 98.5]
Pooled		**1,532**	**97.7%**	**[97.2, 98.2]**

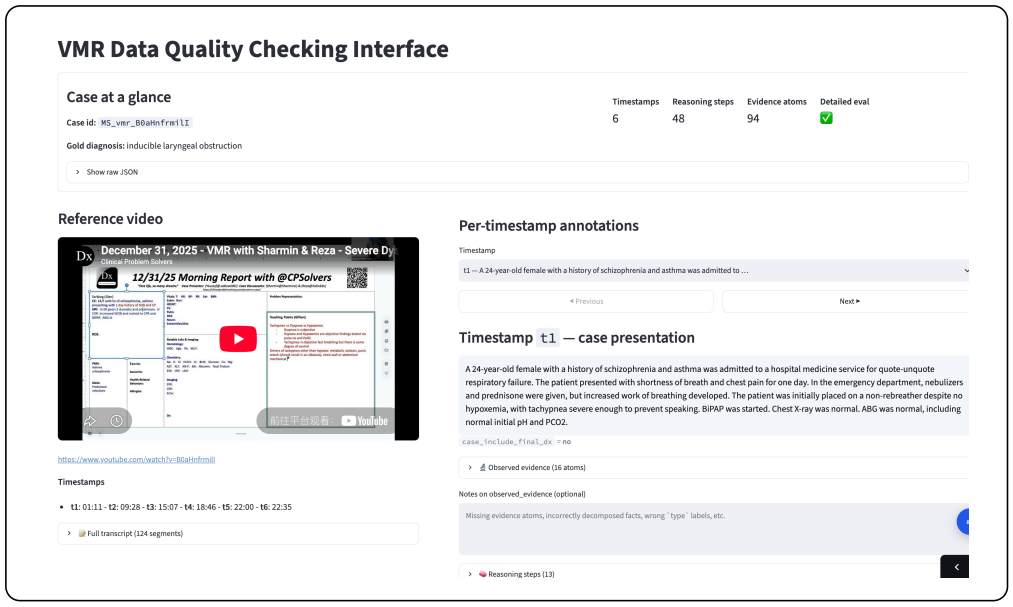


Figure 12: The annotation main page for VMR data. Each case will be paired with the VMR video and the split timesteps.

mance of LLMs when providing full cumulative case information. As shown in Table 12, `GPT` reaches 98.5% accuracy on top-3/5 DDx, showing the strong diagnosis prediction performance on clinical cases. Notably, the smaller `Qwen27B` model also achieves 96.2% top-5 DDx accuracy. Several stronger models approach ceiling-level top-k accuracy at the final stage, while substantial gaps remain across the full model set.

RLR cumulative-context prompting final timestep results. Following the VMR protocol, we report final-timestep RLR performance across all LLMs in Table 13. GPT achieves the highest top-5 DDx accuracy (98.3%), while Claude leads in reasoning comprehensiveness (48.3%). Notably, `Qwen9B` achieves the best final ICD-10 diagnosis accuracy, surpassing the second best `Grok` model.

B.3 RLR CUMULATIVE-CONTEXT PROMPTING RESULTS

Except for the VMR results on cumulative-context prompting setting (Table 3), we show the RLR data results in Table 14. Consistently, `Claude` and `GPT` achieve leading performance on all five tasks, while these two models' ICD-10 final diagnosis accuracy decreases.

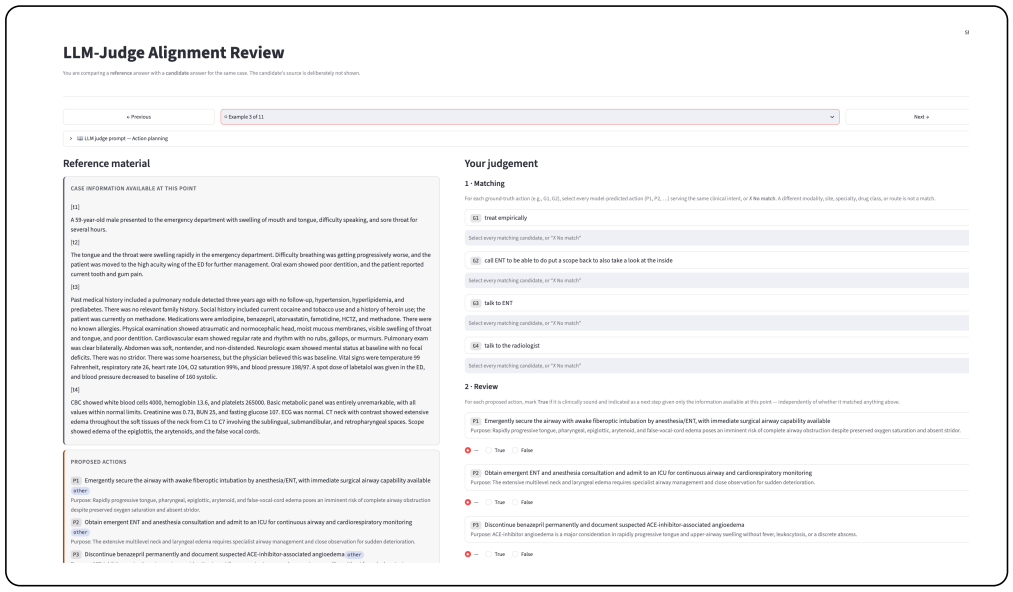


Figure 13: An example of LLM-judge annotation schema and questions.

Table 12: Final-timestep results of cumulative-context prompting on VMR across three tasks. Each case contributes only its own final timestamp. Best and second-best means are **bold** and shaded. * denotes an LLM judge. RC: reasoning comprehensiveness; RTR: reasoning trace recall; CCR: clinical concept recall. All models run three random seeds. Action planning and hypothesis tracking tasks are not applicable at the final timestep.

Metric	GPT	Claude	Grok	Kimi	DeepSeek	Llama	Qwen9B	Qwen27B	Lingshu	Huatuo	MedGemma
TASK 1: PROGRESSIVE DIAGNOSTIC REASONING											
Top-1 DDx Acc (ICD-10)	**63.5** ± 3.3	56.8 ± 4.8	62.9 ± 0.8	59.1 ± 0.2	62.8 ± 3.9	47.8 ± 0.2	62.4 ± 2.0	59.7 ± 1.8	29.6 ± 1.9	52.6 ± 5.2	45.7 ± 3.1
Top-1 DDx Acc*	**96.4** ± 3.1	94.0 ± 3.0	92.9 ± 3.8	92.6 ± 4.3	88.4 ± 3.1	76.1 ± 3.1	92.1 ± 3.4	92.0 ± 3.8	52.7 ± 3.8	85.8 ± 4.5	70.3 ± 2.8
Top-3 DDx Acc*	**98.5** ± 1.3	95.9 ± 3.1	95.4 ± 2.8	95.4 ± 3.5	94.8 ± 0.9	85.0 ± 2.9	94.4 ± 1.2	95.8 ± 1.6	62.1 ± 1.6	90.3 ± 2.5	81.1 ± 2.5
Top-5 DDx Acc*	**98.5** ± 1.3	97.0 ± 4.0	96.4 ± 1.8	95.7 ± 3.1	94.8 ± 0.9	88.0 ± 3.1	95.4 ± 0.8	96.2 ± 1.1	64.6 ± 2.2	90.8 ± 2.3	85.2 ± 1.9
CCR	35.0 ± 1.1	**46.9** ± 1.8	31.0 ± 1.0	37.4 ± 1.6	34.1 ± 1.4	27.7 ± 0.8	32.8 ± 0.7	40.4 ± 0.0	32.1 ± 1.2	28.8 ± 0.9	41.5 ± 1.1
RTR*	70.7 ± 0.5	**85.2** ± 1.8	64.8 ± 0.5	68.9 ± 2.1	66.2 ± 3.3	41.0 ± 1.0	64.3 ± 1.3	67.8 ± 0.6	44.4 ± 0.5	63.6 ± 1.1	57.0 ± 1.6
RC*	32.5 ± 2.5	**50.1** ± 2.8	27.7 ± 1.3	34.0 ± 2.2	29.8 ± 1.4	8.9 ± 1.9	21.7 ± 2.5	32.4 ± 3.6	13.0 ± 0.8	24.2 ± 3.1	20.2 ± 2.1

Table 13: Final-timestep results of cumulative-context prompting on RLR across three tasks. Each eligible case contributes only its own final timestamp; an incomplete final response is not replaced by an earlier response. Best and second-best means are **bold** and shaded. * denotes an LLM judge. RC: reasoning comprehensiveness; RTR: reasoning trace recall; CCR: clinical concept recall. All numbers are percentages, reported as mean [95% CI]. Action planning and hypothesis tracking tasks are not applicable at the final timestep.

Metric	GPT†	Claude	Grok	Kimi	DeepSeek	Llama	Qwen9B	Qwen27B	Lingshu	Huatuo	MedGemma
TASK 1: PROGRESSIVE DIAGNOSTIC REASONING											
Top-1 DDx Acc (ICD-10)	49.1 [41.7,56.5]	45.9 [38.7,53.3]	51.9 [44.8,59.1]	47.2 [40.0,54.6]	43.9 [36.7,51.1]	50.8 [43.6,58.0]	**53.9** [46.6,61.1]	49.2 [41.9,56.6]	37.2 [30.3,44.4]	48.6 [41.3,55.9]	42.0 [35.0,49.4]
Top-1 DDx Acc*	**93.8** [89.8,97.2]	87.0 [82.1,91.9]	87.0 [82.1,91.9]	84.8 [79.4,89.7]	90.2 [85.9,94.0]	78.8 [72.8,84.8]	85.9 [81.0,90.8]	88.6 [83.7,92.9]	57.6 [50.5,64.7]	84.2 [78.8,89.1]	66.9 [59.8,73.4]
Top-3 DDx Acc*	**98.3** [96.0,100.0]	92.4 [88.6,96.2]	92.4 [88.6,95.7]	92.9 [89.1,96.2]	94.0 [90.2,97.3]	89.1 [84.2,93.5]	92.9 [89.1,96.2]	94.0 [90.2,97.3]	66.9 [59.8,73.4]	89.7 [85.3,94.0]	77.2 [71.2,83.2]
Top-5 DDx Acc*	**98.3** [96.0,100.0]	92.9 [89.1,96.2]	92.4 [88.6,95.7]	93.5 [89.7,96.7]	94.6 [91.3,97.8]	91.9 [87.5,95.7]	94.0 [90.2,97.3]	95.1 [91.9,97.8]	69.6 [63.0,76.1]	90.8 [86.4,94.6]	82.1 [76.6,87.5]
CCR	35.5 [33.2,37.8]	**44.6** [41.8,47.4]	31.3 [29.0,33.8]	34.6 [32.1,37.2]	44.0 [41.2,46.8]	29.7 [27.4,32.2]	32.2 [29.8,34.5]	41.1 [38.9,43.5]	32.6 [29.8,35.4]	29.8 [27.6,32.1]	39.9 [37.1,42.7]
RTR*	72.2 [68.8,75.6]	**81.5** [77.6,85.1]	66.1 [62.2,69.9]	68.6 [64.5,72.7]	71.8 [68.1,75.3]	42.8 [39.0,46.5]	63.0 [59.1,66.8]	71.6 [67.8,75.2]	49.8 [45.8,53.8]	65.4 [61.3,69.6]	59.1 [54.9,63.3]
RC*	36.6 [32.8,40.4]	**48.3** [44.0,52.7]	28.8 [25.3,32.4]	33.7 [30.0,37.4]	38.1 [34.2,42.1]	9.1 [7.1,11.3]	22.4 [19.3,25.7]	36.1 [32.3,40.0]	14.2 [11.4,17.1]	27.2 [24.0,30.6]	21.6 [18.2,25.1]

B.4 TIMESTEP-WISE PERFORMANCE

As shown in Figure 14, we report timestep-wise trajectories over 11 LLMs on progressive diagnostic reasoning task.

Table 14: Results of cumulative-context prompting on RLR over all timesteps across five downstream tasks. For each metric, the best result is in **bold** and the second best is shaded. * denotes evaluation by an LLM judge. All numbers are percentages, reported as mean [95% CI]. RC: reasoning comprehensiveness; RTR: reasoning trace recall; CCR: clinical concept recall.

Metric	GPT†	Claude	Grok	Kimi	DeepSeek	Llama	Qwen9B	Qwen27B	Lingshu	Huatuo	MedGemma
Task 1: Progressive Diagnostic Reasoning											
Top-1 DDx Acc (ICD-10)	**24.3** [20.6,28.2]	23.9 [20.1,27.9]	22.5 [19.1,25.9]	22.1 [18.9,25.3]	19.9 [16.5,23.4]	16.1 [13.4,18.9]	20.0 [17.1,23.0]	22.0 [18.8,25.4]	11.0 [8.9,13.2]	17.5 [14.5,20.7]	13.6 [11.2,16.1]
Top-1 DDx Acc*	**50.5** [46.5,54.5]	47.4 [43.4,51.6]	43.4 [39.7,47.2]	42.6 [38.8,46.6]	44.4 [40.9,48.1]	33.8 [30.2,37.6]	40.9 [37.1,44.9]	42.3 [38.9,45.9]	27.4 [23.9,31.0]	37.0 [33.3,40.8]	32.5 [28.8,36.2]
Top-3 DDx Acc*	63.5 [59.2,67.6]	**64.1** [59.9,68.4]	56.7 [52.4,61.0]	58.0 [54.0,62.1]	56.5 [52.4,60.6]	47.0 [42.5,51.5]	54.2 [49.5,58.9]	55.8 [51.5,60.1]	38.4 [34.3,42.8]	46.4 [42.1,50.8]	45.0 [40.5,49.4]
Top-5 DDx Acc*	69.1 [64.9,73.2]	**70.0** [65.8,74.2]	62.8 [58.5,67.2]	64.4 [60.3,68.5]	61.7 [57.7,65.9]	54.2 [49.5,58.9]	60.1 [55.5,64.7]	62.8 [58.5,67.1]	44.1 [39.8,48.5]	51.0 [46.3,55.8]	52.7 [48.1,57.3]
CCR*	25.8 [24.6,27.0]	**36.5** [35.0,37.9]	22.8 [21.8,23.9]	26.0 [24.7,27.2]	30.5 [29.2,31.9]	19.4 [18.3,20.5]	23.0 [22.0,24.1]	29.2 [28.1,30.4]	20.3 [19.4,21.3]	22.5 [21.5,23.5]	27.1 [25.9,28.3]
RTR*	59.5 [57.7,61.2]	**75.8** [73.6,77.9]	50.1 [48.4,51.8]	55.3 [53.4,57.0]	56.9 [55.0,58.8]	31.4 [29.9,32.9]	49.2 [47.6,50.8]	54.6 [53.0,56.3]	32.6 [31.0,34.2]	48.9 [47.1,50.8]	42.7 [40.9,44.5]
RC*	20.0 [18.6,21.4]	**33.1** [31.3,34.9]	15.7 [14.6,16.9]	18.6 [17.4,19.9]	20.1 [18.7,21.6]	4.0 [3.3,4.6]	11.6 [10.6,12.7]	18.6 [17.2,20.1]	5.2 [4.5,6.0]	13.2 [12.1,14.4]	9.5 [8.4,10.5]
Task 2: Action Planning											
Clinical concept recall	23.2 [20.5,25.9]	**30.3** [27.6,33.1]	11.1 [9.3,13.0]	13.2 [11.0,15.4]	13.7 [11.7,15.7]	10.1 [8.5,11.8]	7.7 [6.1,9.3]	15.8 [13.5,18.2]	6.1 [4.7,7.5]	12.2 [10.1,14.3]	5.9 [4.6,7.2]
Action recall*	56.7 [53.9,59.4]	**69.9** [67.5,72.3]	31.0 [28.5,33.5]	36.3 [33.6,39.1]	38.6 [36.1,41.2]	16.6 [14.7,18.5]	26.4 [24.1,28.7]	37.0 [34.4,39.7]	11.9 [10.0,13.9]	32.6 [30.0,35.3]	11.0 [9.2,12.9]
Task 3: Hypothesis Tracking											
Added Hypothesis recall*	23.8 [21.3,26.2]	**42.2** [40.1,44.2]	17.9 [16.3,19.4]	18.2 [16.6,19.8]	14.4 [13.0,15.8]	13.4 [11.8,15.1]	9.0 [7.8,10.3]	13.3 [11.9,14.7]	10.6 [9.2,12.1]	19.9 [18.3,21.6]	11.8 [10.2,13.4]
Ruled-out hypothesis recall	58.8 [53.1,64.4]	59.0 [55.1,63.0]	**69.7** [65.9,73.4]	69.6 [66.1,73.1]	63.3 [59.2,67.3]	45.6 [41.7,49.3]	69.4 [65.5,73.3]	64.9 [61.0,68.8]	33.3 [29.3,37.3]	68.5 [65.0,71.9]	27.9 [23.9,32.0]
Task 4: Clinical Evidence and Reasoning Typing											
Evidence type macro-F_1	**66.5** [64.8,67.8]	66.3 [64.4,67.6]	65.3 [63.4,66.6]	60.5 [58.6,65.7]	59.4 [57.7,65.4]	19.4 [18.7,26.9]	59.1 [57.2,60.5]	59.9 [58.2,65.5]	52.6 [50.7,54.0]	62.4 [60.6,63.6]	21.1 [20.3,29.4]
Step type macro-F_1	**41.0** [40.3,50.8]	38.8 [38.3,47.9]	38.9 [38.2,48.1]	35.3 [34.8,47.2]	36.0 [35.4,47.3]	13.5 [13.0,14.6]	35.1 [34.6,46.1]	32.7 [32.1,38.5]	16.4 [15.9,17.7]	33.0 [32.3,37.2]	9.1 [8.8,10.5]
Reasoning attributes macro-F_1	**44.2** [43.7,49.1]	41.8 [41.4,46.3]	43.7 [43.2,48.6]	38.5 [38.0,42.6]	37.4 [36.8,42.1]	11.4 [10.4,12.8]	31.9 [31.4,36.6]	36.6 [36.0,41.0]	9.5 [8.6,10.9]	29.1 [28.8,33.3]	4.3 [3.1,5.7]
Task 5: Evidence–Hypothesis Relation Prediction											
Relation macro-F_1	87.4 [85.9,88.8]	**88.2** [86.9,89.5]	58.4 [57.5,59.4]	57.3 [56.4,58.3]	43.0 [42.4,58.0]	64.3 [62.0,66.5]	85.5 [84.1,86.9]	86.3 [84.8,87.7]	67.1 [65.0,69.1]	81.9 [80.3,83.5]	32.4 [31.4,66.0]

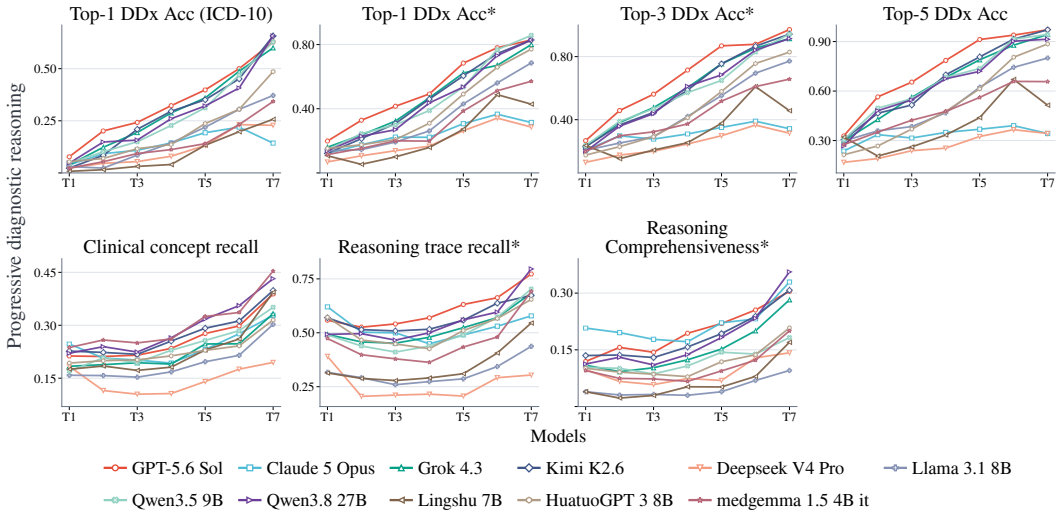


Figure 14: Per-timestep performance trajectories for all LLMs on progressive diagnostic reasoning task. Along the bottom, T1 to T7 represents the timestep points in the case where new information arrives. Y-axis represents the averaged score of each metric. DDx: Differential diagnosis; *: LLM judge.

B.5 Supervised Fine-tuning Results

As we reported in Sec. 5, the supervised fine-tuning is performed under one random seed of VMR data over cumulative-context prompting setting on both `Llama` and `Qwen9B` models. As shown in Table 15, we report the complete evaluation metrics over all five tasks over all timestep setting.

B.6 LLM-as-a-Judge Validation

On 44 randomly sampled VMR cases, one physician independently scored outputs evaluated by both LLM judges. Each judge agreed with the physician on 97.6% of action-recall decisions and 90.9% of top-1 diagnosis decisions (Table 16). Agreement was lower for reasoning-trace recall (75.4–76.9%) than for reasoning comprehensiveness (88.0–90.4%). The physician assigned equal or higher mean scores across all reported metrics; for example, both judges obtained a mean top-3

Table 15: Cumulative context prompting results of supervised fine-tuning (SFT) on VMR over all timesteps (one seed). Each cell reports percentage mean [95% CI]. * indicates evaluation by an LLM judge.

	Llama		Qwen9B	
Metric	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**
A. Pooled timestamps (all timesteps)				
Task 1: Progressive Diagnostic Reasoning				
Top-1 DDx Acc (ICD-10)	**13.4** [10.6, 16.4]	11.3 [8.9, 13.9]	**22.7** [18.8, 26.8]	17.4 [14.6, 20.4]
Top-1 DDx accuracy*	**29.1** [25.1, 33.3]	28.9 [25.3, 32.6]	**40.6** [36.1, 44.7]	37.4 [32.1, 40.7]
Top-3 DDx accuracy*	40.7 [36.3, 45.2]	**44.5** [39.9, 49.0]	53.2 [48.3, 58.1]	**54.6** [49.8, 59.4]
Top-5 DDx accuracy*	48.2 [43.3, 53.2]	**52.5** [47.7, 57.4]	61.7 [56.9, 66.4]	**64.5** [50.8, 67.8]
Clinical concept recall	18.1 [16.9, 19.2]	**24.3** [23.2, 25.4]	22.7 [21.7, 23.7]	**24.8** [21.3, 26.0]
Reasoning trace recall*	30.4 [28.8, 32.0]	**42.7** [40.8, 44.7]	48.5 [46.9, 50.1]	**52.2** [47.0, 56.1]
Reasoning comprehensiveness*	4.2 [3.5, 5.0]	**8.3** [7.2, 9.4]	11.8 [10.8, 12.8]	**12.6** [8.4, 14.0]
Task 2: Action Planning				
Clinical concept recall	9.2 [7.3, 11.3]	**12.5** [10.3, 14.8]	6.3 [4.7, 8.1]	**9.3** [7.7, 11.1]
Action recall*	16.6 [14.3, 18.9]	**20.8** [18.4, 23.3]	**27.5** [24.7, 30.4]	19.8 [17.4, 22.4]
Task 3: Hypothesis Tracking				
Added-hypothesis recall*	13.3 [11.7, 15.0]	**28.6** [26.8, 30.5]	10.1 [8.8, 11.5]	**21.2** [19.5, 22.9]
Ruled-out-hypothesis recall	30.8 [27.8, 34.7]	**37.3** [33.3, 41.4]	**55.8** [50.7, 60.8]	46.7 [42.2, 51.3]
Task 4: Clinical Evidence and Reasoning Typing				
Evidence type macro-F1	16.0 [14.1, 22.6]	**62.2** [61.4, 74.0]	58.4 [57.2, 59.6]	**66.8** [62.0, 71.0]
Step type macro-F1	13.6 [13.1, 15.6]	**36.9** [36.2, 45.1]	33.1 [32.7, 43.4]	**46.6** [40.5, 54.7]
Reasoning attributes macro-F1	6.8 [4.6, 9.4]	**30.3** [27.4, 37.0]	29.9 [29.6, 36.5]	**30.6** [30.3, 37.3]
Task 5: Evidence–hypothesis Relation Prediction				
Relation macro-F1	**64.4** [62.4, 66.3]	52.8 [51.6, 80.6]	**84.8** [83.0, 86.4]	52.5 [51.4, 80.0]

Table 16: LLM-judge validation against a physician on 44 randomly sampled VMR cases. All values are percentages.

	Mean score		
Metric	**GPT-5.6 Luna**	**Claude Sonnet 5**	**Physician**
RC	13.4	10.7	14.0
RTR	63.7	54.9	68.4
Top-1 DDx	72.7	72.7	81.8
Top-3 DDx	81.8	81.8	100.0
Top-5 DDx	100.0	100.0	100.0
AR	29.5	29.5	31.4
	Pairwise agreement		
Metric	**GPT–Claude**	**GPT–Physician**	**Claude–Physician**
Reasoning comprehensiveness	95.2	90.4	88.0
Reasoning trace recall	86.2	76.9	75.4
Top-1	100.0	90.9	90.9
Top-3	100.0	81.8	81.8
Top-5	100.0	100.0	100.0
Action recall	100.0	97.6	97.6

diagnosis score of 81.8%, compared with 100.0% from the physician. These results suggest that the LLM judges are generally conservative and align well with the physician on diagnosis and action evaluation, while reasoning coverage remains more evaluator-sensitive. Because the comparison

uses a single physician, we interpret it as a calibration check rather than a definitive estimate of inter-rater reliability.

C CASE STUDY

As shown in Table 17, we randomly select a VMR case with expert clinical discussion and `Claude` model's generation.

Table 17: A case study of CLINTRACE (VMR test split). Light green marks content aligned with the reference. Evidence types: demo: demographic, sym: clinical symptom, hist: patient history, pex: physical evaluation, lab: laboratory testing, dx-test: diagnostic testing.

key evidence | evidence | hypothesis | aligned with reference | + added | — ~~dropped~~ | [reasoning step type] | [action]

Case presentation revealed at t	Reference (expert discussion)	Model prediction (`Claude Opus 5`)
t_1 A 78-year-old female demo presented with weakness sym… There was 7 kg of unintentional weight loss sym over 4 weeks, with episodes of subjective fever sym.	[problem representation] "Weight loss over several weeks with subjective fevers… suggests chronic infection… or a malignancy." weight loss + fever —supports→ chronic infection, subacute bacterial endocarditis … Hypothesis: + chronic infection, subacute bacterial endocarditis	Clinical reasoning: "Chronic/subacute infection remains a key competing hypothesis" ✓; "night-sweat-equivalent symptoms" ✗ (not reported). Relation: weight loss + fever —supports→ chronic infection, subacute bacterial endocarditis ✓ Hypothesis: Non-Hodgkin lymphoma ✗
t_2 Past medical history included anemia under investigation hist, diabetes hist, Alzheimer's disease, previous ischemic stroke hist… alcohol use unknown hist…	[hypothesis generation] "Anemia could be a paraphenomenon of another inflammatory process, a bone marrow process, or many other causes." anemia under investigation —supports→ bone marrow process, inflammatory process, malignancy Hypothesis: + bone marrow process, urinary catheter infection, aspiration, sacral decubitus ulcer …	Clinial reasoning: "…anemia antedating the current illness raises the possibility of a primary bone marrow disorder …" ✓; Hypothesis: aspiration, sacral decubitus ulcer, amiodarone-induced thyroid dysfunction ✗ (no exposure reported). Evidence typing: alcohol use unknown → other ✗ Hypothesis: urinary catheter infection, aspiration, sacral decubitus ulcer … ✓
t_3 Vital signs were… oxygen saturation 90% on room air pex, temperature 38 degrees Celsius pex… Cardiac examination showed… a diastolic murmur pex in the apical area, intensity 2/6.	[hypothesis generation] "A murmur in the apical area sounds like mitral valve involvement, one of the more frequently involved valves for bacterial endocarditis." [laboratory testing] "blood cultures before giving antibiotics; look for splinter hemorrhages, Osler's nodes; getting a chest X-ray …" diastolic murmur —supports→ mitral stenosis Hypothesis: + mitral stenosis; — ~~active rheumatic fever~~	Clinical reasoning: fever/murmur "is highly suggestive of subacute infective endocarditis" ✓; "iron-deficiency-pattern anemia" ✗ (not established at t_3). Actions: all recovered ✓ — among 19 proposed Hypothesis: mitral stenosis, keep active rheumatic fever ✗
t_4 Laboratory studies showed hemoglobin 6.6 lab… WBC 18,000 lab with 7% neutrophil lab and 8% basophil lab… CRP 70 lab, lactate 5.2 lab.	[hypothesis evaluation] "Strong support for endocarditis… and now a neutrophilia, but the basophilia may be a hidden clue that this is not endocarditis at all." 7% neutrophil —supports→ endocarditis; 8% basophil —supports→ CML; Hb 6.6 + WBC 18k —supports→ anemia of chronic disease [laboratory testing] "get a smear" Hypothesis: + anemia of chronic disease, CML (basophilia).	Clinical reasoning: basophilia "raises concern for a myeloproliferative neoplasm…" ✓ 7% neutrophil —refutes→ endocarditis ✗ *"absence of the neutrophilic leukocytosis typically seen with bacterial"* Hb 6.6 + WBC 18k —refutes→ anemia of chronic disease ✗ Actions "repeat… blood smear" ✓ among 13 Hypothesis: CML; fail to add anemia of chronic disease ✗
t_5 … Echocardiography showed aortic valve vegetation dx-test. Blood culture was positive for Enterococcus sp lab.	[diagnostic confirmation] "Aortic valve vegetation… plus blood culture positive for Enterococcus sp establishes enterococcal endocarditis." vegetation + Enterococcus —supports→ enterococcal endocarditis; —refutes→ mitral stenosis Hypothesis: — ~~AI (aortic insufficiency)…~~ Final diagnosis: *infectious endocarditis* due to *Enterococcus* sp	Clinical reasoning: vegetation plus positive culture, "establishing a definite diagnosis of infective endocarditis" ✓. Relation: 6/6 labels aligned ✓. Final diagnosis: **enterococcal** (likely *E. faecalis*) native **aortic valve infective endocarditis** with sepsis ✓ (core diagnosis aligned) Hypothesis: fail to rule out aortic insufficiency ✗

D PROMPT TEMPLATES

D.1 DOWNSTREAM TASKS PROMPTS

Task 1: Progressive Diagnostic Reasoning

Prompt 1: Progressive Diagnostic Reasoning

```
Clinical information is revealed sequentially over time. Use only
↪  the information available through the current timestep, as a
↪  clinician would when reasoning through an unfolding case.

You will be given the case id, the current timestep, and all
↪  patient's case presentations available up to that timestep.
↪  Provide the following together in one JSON object:

1. reasoning_traces: a list of professional and comprehensive
↪  clinical reasoning statements that are justified by the
↪  available information. Each statement should stand alone.
2. ddx: the diagnoses or diagnostic hypotheses that remain
↪  plausible at this point, ordered from most to least probable,
↪  giving the most likely 5 differential diagnoses.
3. final_dx: the single most likely final diagnosis. The final
↪  diagnosis must be identical to the first entry in the ddx list
↪  (the most probable diagnosis).

Case id: {case_id}

Current timestep: {timestep}

Available case presentation through the current timestep: {context}

Output schema:
{"reasoning_traces":["professional and comprehensive clinical
↪  reasoning statement", "..."],
"ddx":["diagnosis 1 (equal to final_dx)", "diagnosis 2", ...,
↪  "diagnosis 5"],
"final_dx":"the most likely final diagnosis"
}
Return only valid JSON matching this schema. Do not include
↪  markdown, commentary, or explanatory text outside the JSON.
```

Task 2: Action planning

Prompt 2: Action Planning

```
Clinical workup proceeds by choosing useful next actions based on
↪  the information available at the current point in the case.

You will be given the case id, the current timestep, and all
↪  patient's case presentations available up to that timestep.
↪  Recommend the next useful diagnostic, monitoring, consult, or
↪  management actions. For each recommended action, return the
↪  action type, the action atom (the specific action to take), and
↪  the reason for recommending it. Use an empty actions list if no
↪  action is needed or if the presentation does not justify a
↪  specific next action.

Allowed action types: {allowed_types}

Case id: {case_id}

Current timestep: {timestep}
```

```
Available case presentation through the current timestep: {context}

Output schema:
{"actions":
  [
    {"action_type":"type",
     "action atom":"action",
     "purpose":"reason for recommending the action"}
  ]
}

Return only valid JSON matching this schema. Do not include
↪  markdown, commentary, or explanatory text outside the JSON.
```

Task 3: Hypothesis tracking

Prompt 3: Hypothesis Tracking

```
Diagnostic hypotheses should be maintained, revised, added, or
↪  deprioritized as new information becomes available.

You will be given the longitudinal case history of a patient
↪  through the previous timestep, then the list of hypotheses
↪  active at that point, then the next timestep's new case
↪  presentation. Predict the active hypotheses after incorporating
↪  this new presentation.

Apply all three operations to the prior list and report each one
↪  separately: "kept" for the prior hypotheses the new
↪  presentation leaves in play, "dropped" for the prior hypotheses
↪  it rules out, and "added" for hypotheses it newly raises that
↪  were not in the prior list. Every prior hypothesis must appear
↪  in exactly one of "kept" or "dropped". The prior list is a
↪  starting point, not a closed set -- new information can
↪  introduce a hypothesis that was not previously under
↪  consideration, and such a hypothesis belongs in "added" when
↪  the new presentation justifies it. Then output
↪  "active_hypotheses" as the complete resulting set, which is
↪  "kept" followed by "added".

Case id: {case_id}
Previous timestep: {previous}
Target timestep: {target}

History through previous timestep:
{history}

Prior active hypotheses:
{prior_active}

New target timestep presentation [{target_timstamp}]:
{new_presentation}

Output schema:
{"kept":["prior hypothesis the new presentation leaves in play"],
 "dropped":["prior hypothesis the new presentation rules out"],
 "added":["hypothesis the new presentation newly raises"],
 "active_hypotheses":["the complete resulting set, kept followed by
  ↪  added"]}

Return only valid JSON matching this schema. Do not include
↪  markdown, commentary, or explanatory text outside the JSON.
```

Task 4a: Clinical evidence and reasoning typing – Evidence span

Prompt 4a: Clinical evidence and reasoning typing — Evidence Span

```
Case presentations contain evidence spans that play different roles
↪  in clinical reasoning.

You will be given one evidence span from a case presentation.
↪  Classify it into exactly one allowed evidence type.

Allowed evidence types: {evidence_types}

Evidence span:
{evidence_text}

Output schema:
{"target":"evidence","evidence_type":"one allowed evidence type"}

Return only valid JSON matching this schema. Do not include
↪  markdown, commentary, or explanatory text outside the JSON.
```

Task 4b: Clinical evidence and reasoning typing – Reasoning step

Prompt 4b: Clinical evidence and reasoning typing — Reasoning Step

```
Clinical reasoning traces can be categorized by the type of
↪  reasoning step and by finer-grained reasoning attributes.

You will be given one clinical reasoning statement. Classify the
↪  statement into exactly one allowed step type and all applicable
↪  allowed attributes.

Allowed step types: {step_types}
Allowed attributes: {attributes}

Clinical reasoning statement:
{reasoning_statement}

Output schema:
{"target":"reasoning_step","step_type":"one allowed step
↪  type","attributes":["allowed tag"]}

Return only valid JSON matching this schema. Do not include
↪  markdown, commentary, or explanatory text outside the JSON.
```

Task 5: Evidence-hypothesis relation prediction This is a binary classification task.

Prompt 5: Evidence–Hypothesis Relation Prediction

```
Clinical reasoning often requires deciding how newly observed
↪  evidence changes support for a diagnostic hypothesis.

You will be given one candidate hypothesis and the current evidence
↪  spans associated with that hypothesis. Choose the relation
↪  label that best describes whether the evidence supports or
↪  refutes related to the candidate hypothesis, using only the
↪  allowed relation labels.

Allowed relation labels: {allowed}

Case id: {case_id}
Timestep: {timestep}
Candidate hypothesis: {candidate}
```

```
Current evidence:
{evidence_list}

Output schema:
{"relation":"one allowed relation"}

Return only valid JSON matching this schema. Do not include
↪  markdown, commentary, or explanatory text outside the JSON.
```

D.2 LLM-AS-A-JUDGE PROMPTS

We use the consistent LLM-as-a-judge(i.e., `GPT-5.6 Luna`[6]) with the default parameter setting for all LLJ-based evaluation.

Clinical Reasoning Evaluation Prompts Referring to MedCaseReasoning (Wu et al., 2025), we apply the same clinical reasoning recall evaluation prompt to evaluate the reasoning trace recall between gold and predicted traces. The recall is calculated on trace-level.

Clinical Reasoning Recall

```
You are a medical expert aligning individual clinical reasoning
↪  statements.

This task measures semantic alignment between ground-truth and
↪  predicted reasoning traces. It does not evaluate whether the
↪  predicted overall reasons are comprehensive.

For each numbered Groundtruth Diagnostic Reasons, identify any
↪  Predicted Diagnostic Reasons that shows similar clinical
↪  proposition.

Matching criteria:
- Match paraphrases that assert the similar clinical finding,
↪  inference, causal relationship, or evidence-to-diagnosis
↪  relationship.
- A predicted statement may contain additional compatible detail.
- Do not require the prediction to reproduce every secondary detail
↪  of the ground-truth statement.
- Do not match based only on shared keywords, symptoms, or
↪  diagnosis names.
- Do not match statements that reverse the clinical relationship.
- Do not judge the completeness or overall quality of the full
↪  predicted reasons.

Example:
- Ground truth: "The patient has a fever"
- Match: "The patient has a fever due to a viral infection"
- Not a match: "The patient has a sore throat"

Output instructions:
1. Return one JSON object with the single top-level key
↪  "matching_dict".
2. Each key must be a ground-truth statement number represented as
↪  a string.
3. Each value must be a list of matching predicted statements.
4. Copy every matched statement verbatim from Predicted Reasoning
↪  Statements.
5. Use an empty list when no predicted statement matches.
```

[6] https://developers.openai.com/api/docs/models/gpt-5.6-luna

```
6. Include exactly one key for every ground-truth statement.
7. Return only the JSON object wrapped in ```json tags.

Required structure:
```json
{{
  "matching_dict": {{
    "1": [],
    "2": ["Verbatim matching predicted statement"],
    "3": ["First verbatim match", "Second verbatim match"]
  }}
}}
```
Groundtruth Diagnostic Reasons:
{gold_block}

Predicted Diagnostic Reasons:
{pred_block}
```

To further evaluate the comprehensiveness of model-generated clinical reasoning traces, we apply the same prompt used in MedR-Bench (Qiu et al., 2025) to evaluate the alignment between the gold and predicted reasoning traces. This metric is calculated by taking the reasoning traces as a whole instead of trace-level, which is more strict than reasoning recall in the above prompt.

Clinical Reasoning Quality Evaluation

```
You are an experienced medical expert evaluating the quality of a
↪  model's predicted rationales relative to ground-truth
↪  rationales and the clinical information available at the
↪  current point in time.

You will be given the cumulative case information available through
↪  the current timestep, a set of ground-truth rationales, and the
↪  model's predicted reasoning traces. Evaluate the prediction in
↪  two ways.

1. Comprehensiveness

Given the set of ground-truth rationales and the model's
↪  prediction, evaluate whether the prediction comprehensively
↪  mentions each ground-truth rationale. Grade in a point-wise
↪  manner:

- A ground-truth rationale is considered mentioned only if its main
↪  content is explicitly stated in the model's prediction.
- Do not award the rationale if any key information of the
↪  ground-truth rationale is missing from the model's prediction.
- Do not award the rationale if any information in the model's
↪  prediction conflicts with the ground-truth rationale.
- Do not award the rationale if only a small part of the
↪  ground-truth rationale is mentioned in the model's prediction.
- Judge only what is explicitly stated in the model's prediction.
↪  Do not infer missing reasoning from the case information.

For each ground-truth rationale, output a Boolean indicating
↪  whether the prediction comprehensively mentions it.

Cumulative case information:
{question}

Ground-truth rationales:
{gold_block}
```

```
Model-predicted reasoning traces:
{predicted_block}

Return only a JSON object with exactly this key structure,
↪  replacing every <true or false> placeholder with a JSON
↪  Boolean:
{response_template}

- "evaluations" must contain exactly one key per ground-truth
↪  rationale, numbered in order starting at 1.
- "reasoning_review" must contain exactly one key per predicted
↪  reasoning trace, numbered in order starting at 1.
- Every value in "evaluations" must be a Boolean.
- Use an empty object when the corresponding list is empty.
```

Diagnosis Accuracy Evaluation Given the complexity of clinical diagnosis evaluation, traditional ICD-10-based metrics are unable to match more implicit overlap between two diagnoses. Therefore, we follow the prior work to evaluate the diagnosis accuracy using LLM-as-a-judge (Gaber et al., 2025).

Diagnosis Accuracy Evaluation

```
You are an experienced healthcare professional with expertise in
↪  medical and clinical domains. I will provide a list of real
↪  diagnoses for a patient and {n} predicted diagnoses. For each
↪  predicted diagnosis, determine if it has the same meaning as
↪  one of the real diagnoses or if the prediction falls under a
↪  broader category of one of the real diagnoses (e.g., a specific
↪  condition falling under a general diagnosis category).

If it matches, return 'True'; otherwise, return 'False'.

Output format:
- Return exactly {n} lines, one verdict per predicted diagnosis, in
↪  the same order as listed below.
- Each line must be exactly True or False (capitalized).
- Wrap all verdict lines inside <evaluation> tags.
- Do not include numbering, labels, explanations, markdown, or any
↪  text outside the <evaluation> block.

Example for 2 predicted diagnoses:
<evaluation>
False
True
</evaluation>

Real Diagnoses: {real_block}

{pred_lines}
```

Action Planning Evaluation To evaluate the model-predicted actions compared with the gold actions in our dataset, we design a two-way LLM-as-a-judge prompt inspired by the Clinical Reasoning Quality Evaluation prompt.

Prompt 6f: Clinical Action Matching

```
You are an experienced medical expert evaluating the next clinical
↪  actions a model proposed for a patient case at a given point in
↪  time.
```

```
You will be given:
- Case information available at this timestep
- Reference actions: the next actions the treating clinician
 ↪  actually requested at this point (the coverage target)
- Proposed actions: the model's recommended next actions

Evaluate the proposed actions in two ways.

1. Matching (reference coverage). For each numbered reference
 ↪  action, find the numbered proposed actions that serve the same
 ↪  clinical intent. For instance, if reference action 1 is "CT
 ↪  angiography of the chest" and proposed action 2 is "CTA chest
 ↪  to evaluate for pulmonary embolism", then proposed action 2
 ↪  matches reference action 1. If reference action 1 is "send
 ↪  blood cultures" and proposed action 1 is "order labs", then
 ↪  that is not a match.
- Do not match actions that differ in modality, anatomical site,
 ↪  specialty, drug class, or route.
- Do not match a broader proposed action unless carrying it out
 ↪  necessarily includes the reference action.
- Match on the action itself, not on the type label or the stated
 ↪  purpose.
- One proposed action may match several reference actions, and
 ↪  several proposed actions may match the same reference action.

Case information:
{case_context}

Reference actions (use these numbers in matching_dict keys):
{reference_block}

Proposed actions (use these numbers in matching_dict values and
 ↪  action_review keys):
{predicted_block}

Wrap your JSON output in ```json tags.

Example of the required JSON structure:
```json
{{
  "matching_dict": {{
    "1": [2],
    "2": []
  }}
}}
```

JSON numbering rules:

- "matching_dict" keys are reference-action numbers from the
 ↪ Reference actions list above.
- Each "matching_dict" value is a list of proposed-action numbers
 ↪ from the Proposed actions list above. In the example, reference
 ↪ action 1 matches proposed action 2.
- Use an empty list when a reference action has no matching
 ↪ proposed action.
- Include every reference-action number even when the list is
 ↪ empty. Use {{}} only when there are zero reference actions.
- "action_review" keys are proposed-action numbers from the
 ↪ Proposed actions list above.
- Refer to actions only by these list numbers. Do not invent new
 ↪ numbers.
```