

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046

# AUTOCLINICIAN: STRUCTURED CLINICAL GUIDELINE INTEGRATION FOR TRUSTWORTHY DIAGNOSTIC REASONING IN HEALTHCARE

Anonymous authors

Paper under double-blind review

## ABSTRACT

Recent advances in large language models (LLMs) have enabled clinical conversational systems with impressive diagnostic capabilities. However, existing approaches often lack alignment with real-world clinical workflows and fail to provide interpretable, evidence-grounded reasoning. In this work, we propose AutoClinician, a unified and training-free framework that integrates clinical guidelines to support stepwise and explainable diagnosis on real-world electronic health records (EHRs). AutoClinician first extracts and summarizes narrative guidelines into Clinical Evidence Graphs (CEGs). These graphs are further automatically verified and refined using a consistency-based strategy. To support trustworthy and patient-specific diagnosis, we utilize CEGs by conducting context-aware, evidence-grounded clinical reasoning on EHRs with Deterministic Finite Automaton (DFA). Our framework outperforms both general-purpose and clinically specialized LLMs, and exhibits stronger interpretability. Code is available here <sup>1</sup>.

## 1 INTRODUCTION

Recent advances in LLMs (Touvron et al., 2023; Wu et al., 2023) have created new opportunities to enhance diagnostic assistance and patient interactions in healthcare (Biswas, 2023; Li et al., 2023b; Shah, 2024; Singhal et al., 2023). Clinical conversational systems have demonstrated competitive performance in medical question-answering tasks (Wang et al., 2023a; Yang et al., 2024b). However, as shown in Figure 1, most existing efforts adopt a pure text-based setting that deviates significantly from real-world clinical workflows, focusing on tasks such as answering USMLE-style questions (Wang et al., 2023a; Labrak et al., 2024). These medical questions mainly address general inquiries but are limited in supporting clinical workflows such as test recommendation, result interpretation, and follow-up planning. In real-world clinical workflows, patient clinical events are documented in Electronic Health Records (EHRs) (Shickel et al., 2017), which have been largely overlooked in recent work. Moreover, high performance alone is insufficient for clinical deployment. Physicians require AI assistants that support evidence-grounded reasoning, where each decision adheres to the principles of *evidence-based medicine (EBM)* to ensure transparency, legal defensibility and clinical accountability (Subbiah, 2023).

Extending LLMs to work with EHRs is critical for aligning them to real-world medical workflows. However, the design of EHR-enhanced clinical conversational agents faces key limitations: (1) **Difficulty in interpreting medical records.** EHRs contain numerical and domain-specific data that require precise clinical interpretation, posing fundamental challenges for text-based LLMs which lack grounding in the diagnostic significance of various laboratory test values (Li et al., 2024). Recent studies show that GPT-4’s diagnostic accuracy drops

---

<sup>1</sup><https://anonymous.4open.science/r/Autoclinician-0AAD>

047

048

049

050

051

052

053

054

055

056

057

**Patient:** Hi doctor, I am getting one-sided headache. I do not get them often. They are behind my right eye and temple area. Now this time I am also fighting with heavy cold and cough. It came on late morning today and it hurts more when I cough or shake my head or bend over. Prior to this, I had the same headache about four months ago. At that time it came on midafternoon and lasted till next morning. Prior to that, I had the same headache exactly on the right side and lasted for a day. I am on CCB and Protonix. I am scared of brain tumor. Please help.

**Clinician:** Hi, Thank you for posting your query. First, I would like to reassure you that there is no need to worry about the brain tumor. Also, we should keep migraine apart as well. The most likely cause of your headaches could be sinusitis. This can be confirmed by doing an X-ray of the face. Treatment would consist of using decongestant tablets/nasal sprays. Best wishes, Chat Doctor.

(1) An example of medical question answering. (2) An example of Electronic Health Records (EHRs). (3) An example of an EHR-driven dialogue.

Name	Gender	Alert	Surfa-drugs	
Helen	Male			
Phone	Age			
371-482-9156	45			
Symptom				
Date	Code	System	Code	Lab Test
20250311	ICD-10		R10.12	Results
				Diagnosis
20250312	SNOMED		241373003	84.34 ml/min/(175 m2)
20250314	LOINC		48647-5	90.86 ml/min/(175 m2)
20250315	SNOMED		8462-4	93 mm[Hg]
	LOINC		271649006	86 mm[Hg]
	LOINC		29463-7	175 lb_a
20250317	ICD-10		Q65.89	
	ICD-10		M87.00	
Clinical Event				
ICD10R10.12				● Patient
SNOMED241373003				● Symptom
LOINC48647-5				● Laboratory Test
...				● Clinician
ICD10Q65.89				● +Recommend Laboratory Test
				● +Interprete Test Value
				● +Make Final Diagnosis
20250311				
20250312				
20250313				
20250317				

**Patient:** I'm a 41-year-old White female. I've been feeling nauseous, having abdominal discomfort, and experiencing pain while urinating. Could you recommend any lab tests?

**AutoClinician:** Based on your symptoms, I recommend testing for blood glucose (serum and capillary), urinalysis for glucose and ketones, and basic metabolic markers including creatinine and electrolytes.

**Patient:** Here are my Day 1 lab test ALT 32.0 U/L, AST 31.0 U/L, albumin 3.8 g/dL, calcium 8.1 mg/dL, amylase (pancreatic) 6.0 U/L, magnesium 1.8 mg/dL, total bilirubin 0.4 mg/dL, CRP 22.0 mg/L, and total CO<sub>2</sub> 25.0 mmol/L.

**AutoClinician:** Your glucose is within the normal range and ketones are negative. I recommend you perform a capillary glucose test to gather more information.

Figure 1: Comparison between classical medical QA and EHR-driven diagnosis. (1) A typical medical QA setting, which takes a textual question and directly returns a diagnosis result. (2) Electronic Health Records (EHRs), which document patient information such as symptoms and laboratory test results, and are routinely used in real clinical workflows. (3) AutoClinician performs guideline-grounded reasoning on EHRs to generate diagnoses with explicit rationales.

significantly when test results are not properly interpreted or integrated (Bhasuran et al., 2025). (2) **Lack of explicit medical rationale.** There are underlying rationales behind EHRs, such as why a specific test was requested or how the diagnosis is supported by test results. Capturing this reasoning evidence is essential for building trustworthy clinical assistants, yet it has been overlooked by existing work. To address these challenges, we pose a critical research question: how do we empower foundational LLMs with expert medical knowledge and evidence-grounded clinical reasoning abilities?

To overcome these limitations, we propose to incorporate *official clinical guidelines* into the reasoning of diagnostic agents. It is known that official clinical guidelines establish evidence-based diagnostic criteria, define validated thresholds for clinical interpretation, and provide legal justification for diagnosis. However, automating clinical guideline integration for diagnostic conversational systems faces three key challenges. (1) **Narrative complexity.** Clinical guidelines are often lengthy and comprehensive. It is important to extract concise and discriminative evidence to assist LLMs with diagnostic workflows. (2) **Lack of guideline-grounded validation mechanism.** The absence of concise guideline knowledge makes it difficult to evaluate the correctness of extracted information. Human evaluation or expert-annotated supervision is prohibitively costly for its scalability. (3) **Trustworthy reasoning for each patient.** Designing a transparent reasoning paradigm that is grounded in guideline-supported rationales is critical for a trustworthy diagnostic agent.

Concretely, we propose **AutoClinician**, a training-free framework that leverages clinical guidelines to reason over EHR-based dialogues. AutoClinician comprises two key components: (1) **Clinical Guideline Compression and Refinement.** We introduce Clinical Evidence Graphs (CEGs), which are distilled from clinical guidelines to encode structured diagnostic evidence. To ensure accuracy and scalability, we propose a consistency-based strategy that generates pseudo-patient examples to automatically refine CEGs quality at scale. (2) **Conversational Reasoning via Patient-Specific Deterministic Finite Automaton (DFA).** We model diagnostic interactions using DFA. This structured state tracker integrates an LLM with Retrieval-Augmented Generation (RAG), enabling an interpretable and stepwise reasoning process.

The contributions of our work are listed as follows: (1) We develop AutoClinician, the first framework to integrate clinical guidelines into EHR-based diagnostic conversations, enabling evidence-grounded reasoning. (2) We propose a scalable, unsupervised algorithm that automatically refines CEGs from clinical guidelines through a consistency-based strategy. (3) We introduce a DFA to track and manage patient-specific reasoning based on CEGs, enabling evidence-supported conversations. (4) AutoClinician outperforms both general-purpose and medical-specialized LLM baselines in laboratory test recommendation and diagnosis prediction, demonstrating superior reasoning accuracy, interpretability, and adherence to clinical guidelines.

094  
095  
096

## 2 RELATED WORKS

097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107  
108

**Diagnosis Modeling from Electronic Health Records.** EHRs encode diagnostic trajectories through structured data such as demographics, symptoms, and clinical test results, making them a key resource for diagnosis prediction. One research direction focuses on enriching medical concept representations using external knowledge sources like clinical ontologies and knowledge graphs (Choi et al., 2017; Panigutti et al., 2020; Jiang et al., 2023; Wang et al., 2023b), but these often suffer from limited disease coverage and poor generalizability. To address this, recent methods employ unsupervised learning and LLMs to derive code embeddings via self-supervised relational graphs (Yao et al., 2024), contrastive pretraining (Cai et al., 2022), or code generation (Ma et al., 2024), yet they frequently overlook EHR heterogeneity and omit numerical laboratory test values. A second line of work treats EHRs as tabular data, framing diagnosis prediction as a supervised classification task. Methods such as XGBoost (Chen & Guestrin, 2016), generalized additive models (Hastie, 2017), piecewise linear functions (Montomoli et al., 2021), and rule-based learners (Ren et al., 2024) aim to learn decision functions over laboratory test values. However, these models are prone to population bias and often lack alignment with clinically grounded criteria.

109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119

**Retrieval-Augmented Generation for Medical Dialogue Systems.** Integrating external knowledge sources is critical for dialogue systems to generate trustworthy responses. In biomedical domains, Self-BioRAG (Jeong et al., 2024) dynamically searches documents from knowledge corpus to improve the response quality. MRD-RAG (Chen et al., 2025) further explores multi-disease retrieval by analyzing semantic relationships and contrasts among candidate diagnoses. CLEAR (Lopez et al., 2025) extracts information from clinical notes to improve the understanding of EHR data. RAICL (Zhan et al., 2025) combines retrieval-augmented generation with in-context learning to select disease-relevant demonstrations to improve image-based diagnosis performance. MAVEN (Jadhav et al., 2024) applies RAG to identify and correct factual errors in clinical notes. For a comprehensive overview, we refer readers to the recent survey (He et al., 2025). In contrast, we retrieve from clinical guidelines and refine compressed guideline graphs to support evidence-based diagnosis.

120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130

**Learning from Clinical Guidelines.** Learning discriminative structures from clinical guidelines is crucial for clinical decision support (Zhu et al., 2024; He et al., 2024). Early works developed rule engines (Mei et al., 2011) to construct rules from medical literature. Previous work formalizes guideline-driven decision processes as an information extraction problem. For example, previous work (Wu, 2022) combines a BERT-style encoder (Cui et al., 2021) to extract relation triples and compose structured trees. PromptRE (Jiang et al., 2022) frames tree extraction as a multi-round conditional relation extraction task. Recent works (Zhu et al., 2024; Li et al., 2023a; He et al., 2024) further leverage LLMs to transform unstructured medical texts into machine-executable decision trees. Another line of works first generate if-else pseudocode and then refines the trees (Hou et al., 2025). However, these methods primarily extract textual information, while neglecting numerical data that are critical for interpreting laboratory test results. Moreover, they mainly rely on tree-structured representations, which impose a single-parent hierarchy and thus fail to capture the concurrent and interdependent nature of real-world diagnostic logic.

131  
132  
133  
134  
135  
136  
137  
138  
139  
140

## 3 METHODOLOGY

141 3.1 LEARNING CLINICAL EVIDENCE GRAPHS FROM GUIDELINES  
142143 In this section, we first introduce the construction process and formal definition of CEGs, and then present a  
144 consistency-based strategy that synthesizes pseudo-patient examples from source guidelines to iteratively  
145 refine these graphs.146  
147 3.1.1 CONSTRUCTION OF CLINICAL EVIDENCE GRAPHS  
148149 First, we extract and summarize concise and critical rationales from guidelines in the form of Clinical  
150 Evidence Graphs (CEGs). Each graph encodes the reasoning pathways for one guideline, where each node is  
151 annotated with a  $\langle \text{ID}, \text{context}, \text{trigger}, \text{state}, \text{termination} \rangle$  tuple. We define these node  
152 attributes as follows:153 

- **ID:** Denotes the index of current node in the graph.
- **Context:** Represents the current diagnostic status, (e.g., ‘Context = Confirm Testing’ after  
154 patient completing relevant laboratory tests). It links to the state of its parent node, ensuring consistent  
155 progression along the diagnostic pathway.
- **Trigger:** Specifies the clinical condition or threshold defined by guidelines. It is activated when patient’s  
156 laboratory results or clinical observations meet a specified criteria. For example: ‘Trigger = {A1C  $\geq$  6.5%,  
157 FPG  $\geq$  126 mg/dL, or random glucose  $\geq$  200 mg/dL}’ indicates suspicion of diabetes.
- **State:** Specifies the guideline-prescribed clinical action taken when the Trigger is activated. It reflects  
158 how to act upon the observed condition based on clinical guidelines. For example, ‘State = Suspect  
159 diabetes’ when the patient’s laboratory tests satisfy the defined Trigger.
- **Termination.** True or False. If true, this node is a leaf (final diagnosis/conclusion). If false, the engine  
160 will continue by locating the next step.

161 This structured 5-tuple formalizes the stepwise clinical reasoning encoded in guidelines, enabling systematic  
162 modeling of diagnostic pathways for downstream tasks. Concretely, we first collect and parse guidelines and  
163 use a state-of-the-art LLM to identify diagnosis-relevant sections. These identified sections are transformed  
164 into symptoms and corresponding diagnostic logic graphs, where symptoms serve as the key for CEG retrieval.  
165 Notably, each graph may lead to multiple diagnostic outcomes. For example, a diabetes-related guideline can  
166 terminate in different diagnoses, such as type 1 or type 2 diabetes<sup>2</sup>. Appendix B.2 provides the prompts for  
167 diagnosis-related section extraction, and Appendix B.3 provides the prompts for CEGs generation. A CEG  
168 derived from a diabetes-related guideline is shown in Appendix C.4.2.169  
170 3.1.2 AN AUTOMATIC VERIFICATION AND REFINEMENT STRATEGY FOR CEGS  
171172 CEGs abstract diagnostic rules from guidelines into structured decision paths, but their accuracy fundamentally  
173 depends on the capabilities of base LLMs. To ensure their correctness, a natural solution is to incorporate  
174 human-in-the-loop verification. However, given the extensive volume of guidelines and the cost of expert  
175 evaluation, this approach is inherently limited in scalability and susceptible to human bias and error.176 To address these limitations, we design an automatic verification and refinement strategy that synthesizes  
177 pseudo-patient examples from the source guidelines and employs them to iteratively refine the CEGs.  
178 Intuitively, each pseudo-patient example can be viewed as an instantiation of a specific CEG branch, serving  
179 as a concrete test case of the abstract rule. Because the CEG and the pseudo-patient examples share a common  
180 source, any inconsistency (for example, a pseudo-patient example that fails the CEG’s logic) pinpoints a  
181182 <sup>2</sup>[https://diabetesjournals.org/care/article/46/Supplement\\_1/S19/148056/2-Classification-and-Diagnosis-of-Diabetes](https://diabetesjournals.org/care/article/46/Supplement_1/S19/148056/2-Classification-and-Diagnosis-of-Diabetes)

misalignment that need to be corrected. To systematically enhance the effectiveness of CEGs, we first examine recurrent error modes and subsequently construct pseudo-patient examples to address them.

Based on our analysis with professional clinicians, we first identify four common error modes in CEGs: (1) **Incomplete Evidence** that misses diagnostic conditions, such as subtle symptoms or laboratory tests. (2) **Threshold Misalignment** that misassigns numerical thresholds to incorrect laboratory tests. (3) **Logic Errors** that misorders diagnostic test steps specified by guidelines. (4) **Conflict Resolution Failure** that handles conflicting symptoms or laboratory tests. Detailed examples can refer to Appendix C.4.1.

To systematically prevent and correct error modes, we design four types of pseudo-examples tailored to specific failure cases. Each pseudo-example is assessed by an LLM-as-a-judge, producing a binary alignment score with respect to the generated CEGs. If the alignment score falls below a predefined threshold (0.8 in our experiments), the LLM is prompted with failed pseudo-patient examples to refine the initial CEGs. These flagged examples are incorporated as in-context demonstrations, guiding the LLM to recover missing constraints, adjust threshold logic, and revise invalid graph transitions. This strategy enables unsupervised validation and correction of clinical reasoning structures without expert supervision. The prompt used for refinement is provided in Appendix B.4.

### 3.2 MODELING CONVERSATIONS WITH A DFA-GUIDED STATE TRACKER

In this section, we model the diagnostic workflow as a Deterministic Finite Automaton (DFA), which aligns multi-step reasoning with clinical guidelines. The DFA acts as a conversational state tracker, summarizing the patient’s current status, comparing EHR evidence with triggers of CEG nodes, and driving guideline-informed state transitions accordingly. This design provides a transparent and structured scaffold that ensures faithfulness between diagnostic reasoning and guideline logic.

#### 3.2.1 CONVERSATIONAL STATE REASONING TRACKED VIA DFA

DFA is used to model structured decision processes (Minsky, 1956). Formally, a DFA is defined as a tuple  $(\mathcal{S}, \Sigma, \delta, s_0, F)$ , where  $\mathcal{S}$  is a finite set of states,  $\Sigma$  denotes the input space,  $\delta : \mathcal{S} \times \Sigma \rightarrow \mathcal{S}$  is the transition function,  $s_0 \in \mathcal{S}$  is the start state, and  $F \subseteq \mathcal{S}$  is the set of terminal states. We capture the reasoning process by defining a DFA as the *Conversational State Tracker*:

- $\mathcal{S}$ : The set of conversational states, each representing the evolving diagnostic context defined by accumulated patient observations and reasoning progress. Concretely, we specify five distinct states: {Start, PartialResults, ConflictResults, ConfirmedDiagnosis, Unresolved}, where  $s_0 = \text{Start}$  and  $F = \{\text{ConfirmedDiagnosis}, \text{Unresolved}\}$  are terminal states indicating completion of the reasoning process.
- $\Sigma$ : The input space, which captures the output of diagnostic reasoning step after examining the EHR record with guideline CEGs. Specifically, we define five categories of inputs:
  - **ConditionSatisfied**: Given a patient’s input (laboratory test results or symptoms), the evidence is compared against the conditions specified in the trigger of the current CEG node. If the conditions are met, the input is labeled as **ConditionSatisfied**, and reasoning proceeds to the corresponding child node in the next step. For example, it will evaluate whether a laboratory value is abnormal according to a guideline-defined threshold, or whether multiple tests jointly satisfy a logical condition such as “A is abnormal *and* B is abnormal.”
  - **Value missingness**: If a guideline specifies required tests or symptoms that are not present in the patient record at the current turn, they are encoded as missing. For example, if the CEG requires an *HbA1c* test but the record does not contain it, this is represented as *HbA1c* is missing.
  - **Conflict**: A conflict arises when the patient’s laboratory test results or symptoms are inconsistent with the diagnostic requirements specified in the retrieved CEG. For example, diabetes diagnosis requires

235 both an abnormal A1C and an abnormal fasting plasma glucose (FPG). If a patient presents with an  
 236 abnormal A1C but a normal FPG, this inconsistency is represented as a conflict.  
 237

- 238 – *Sufficiency*: A patient’s symptoms and laboratory test results are considered *sufficient* when they  
 239 collectively satisfy all diagnostic requirements specified in the retrieved CEG, and the reasoning step  
 240 reaches its termination node with a diagnosis result.
- 241 – *Failure*: We transit to Unresolved when the number of explored CEGs has exceeded maximum limit  
 242 (which is set to 5 in this work). It represents that the target patient’ input does not meet any paths in  
 243 the retrieved candidate CEGs, and the model falls back to relying on itself for diagnosis.
- 244 •  $\delta$ : the transition function  $\delta : \mathcal{S} \times \Sigma \rightarrow \mathcal{S}$  encodes guideline-grounded reasoning rules that map a  
 245 conversational state and structured inputs to the next state. We define the transition types as follows:  
 246 (1) transition to PartialResults when the current query satisfies the trigger conditions but has not yet  
 247 reached a termination (True) state; (2) transition to ConflictResults when patient evidence is inconsistent  
 248 with CEG requirements, and then re-start the reasoning with the next candidate CEG; (3) transition to  
 249 ConfirmedDiagnosis when all requirements of the candidate CEG are satisfied, leading to a final diagnosis;  
 250 (4) when symptom or laboratory test missingness is detected, the DFA will continue the transition to  
 251 PartialResults during reasoning. (5) transition to Unresolved when diagnosis cannot be made with all the  
 252 five retrieved CEGs, and generate diagnosis by LLM itself.

### 253 3.2.2 CONVERSATION GENERATION VIA PATIENT-SPECIFIC DFA

254 This formalized DFA serves as a powerful tool to align reasoning steps with CEG paths. We prompt the LLM  
 255 to perform diagnosis using our defined DFA paradigm. The reasoning is an iterative process that includes  
 256 information seeking, retrieval of relevant CEGs, and state transitions until a final diagnosis is reached. Here  
 257 are the concrete steps:

258 (1) **Querying and Retrieving Relevant CEGs.** We first transform patient demographics and symptoms into  
 259 a retrieval query. We adopt the all-MiniLM-L6-v2 model (Reimers & Gurevych, 2019) to embed the patient’s  
 260 symptoms and each candidate CEG’s symptoms. We use cosine similarity to measure their relevance, and  
 261 select the top-5 candidates with the highest scores. During DFA navigation, these candidates are sequentially  
 262 selected to guide the reasoning until a termination state is reached.

263 (2) **DFA Input Generation.** Given a retrieved graph, we generate the input  $\Sigma$  for DFA navigation by  
 264 comparing the patient input with the triggers at the current node. The outcomes are abstracted into structured  
 265 indicators such as ConditionSatisfied, ValueMissingness, Conflict, and Sufficiency, which serve as inputs  
 266 driving state transitions.

267 (3) **Navigating DFA:** Based on  $\Sigma$  derived from the patient’s input and the retrieved graph, LLMs navigate the  
 268 DFA along a deterministic path to identify the corresponding state, as defined in Section 3.2.1.

269 (4) **LLM Response Generation:** This reasoning loop of step 2 and step 3 runs iteratively until a final state is  
 270 reached, after which we prompt the LLM to make a diagnostic decision based on the reasoning result.

## 271 4 EXPERIMENTS

272 We design extensive studies to examine the effectiveness of AutoClinician, covering both retrieval and  
 273 reasoning aspects. Q1) Can existing LLMs, both general-purpose and medical specialized, perform accurate  
 274 diagnostic reasoning over EHRs in the absence of clinical guidelines? Q2) Does AutoClinician achieve  
 275 superior diagnostic accuracy and lab test recommendation performance compared to baselines? Q3) Does  
 276 AutoClinician properly integrate structural information from clinical guidelines, and what factors influence the  
 277 effectiveness of guideline incorporation? Q4) Does modeling diagnostic reasoning as a patient-specific DFA  
 278 provide advantages over CoT prompting or unstructured retrieval? Q5) How does the number of retrieved  
 279 candidate CEGs affect diagnostic accuracy and the quality of reasoning? Q6) How do clinician evaluate the  
 280

282 quality, faithfulness, and clinical plausibility of the guideline-grounded reasoning produced by AutoClinician?

283

#### 284 4.1 DATASET AND METRIC DESCRIPTION

285

286 To evaluate the performance of LLMs on EHRs, we design datasets based on four principles: (I) *Reality*:

287 Sourced from real-world patient records to reflect authentic clinical workflows, following a pipeline of

288 ‘evidence acquisition, result interpretation, and diagnosis confirmation’. (II) *Unexposure*: Ensuring that the

289 corpora have not been previously used by most LLMs. (III) *Open-access*: We use datasets that are publicly

290 available or broadly accessible under standard data-use agreements. (IV) *Predictability*: Focusing on diseases

291 that can be predicted from EHR data. To satisfy these requirements, we adopt two prominent and challenging

292 EHR benchmarks that necessitate multi-turn dialogue reasoning: EHRSHOT (Wornow et al., 2023) and

293 TriNetX (Suarez Arbelaez et al., 2023). Detailed dataset descriptions can be found in Appendix C.1.

294

295 We evaluate diagnosis performance using accuracy (ACC) and mean reciprocal rank (MRR), and we also

296 report the laboratory test prediction accuracy (Lab Acc). Details of metrics are shown in Appendix C.2.

297 4.2 BASELINE DESCRIPTION

298

299 We compare AutoClinician with comprehensive baselines (1) Vanilla Zero-shot Prompting, including general-

300 purpose LLMs (Qwen3-32B Instruct (Yang et al., 2024a), QwQ-32B (Team, 2025), GPT-4o (Achiam et al.,

301 2023) and medical-specialized models ChatDoctor (Li et al., 2023b), BioMistral (Labrak et al., 2024), Yi-

302 34B (Young et al., 2024) and DiagnosisGPT-34B (Chen et al., 2024)). (2) Few-shot Demonstration with

303 Instructions. We prompt the LLM with a few exemplar dialogues and instructions illustrating a diabetes-

304 related diagnostic reasoning process. (3) Iterative RAG with Query Planning. LLMs will split the diagnostic

305 task into sub-queries through planning, and conduct RAG on guidelines during solving each sub-query.

306

307 Table 1: Performance comparison of AutoClinician against baselines on the EHRSHOT dataset, reporting

308 diagnosis accuracy (Acc), mean reciprocal rank (MRR), and laboratory test accuracy (Lab Acc).

309

Model	EHRSHOT-Metabolic			EHRSHOT-Respiratory			EHRSHOT-Circulatory		
	Acc	MRR	Lab Acc	Acc	MRR	Lab Acc	Acc	MRR	Lab Acc
<b>General-purpose LLMs</b>									
Qwen3-32B	19.73	17.81	16.61	22.31	16.15	18.08	18.47	19.40	18.73
QwQ-32B	17.38	16.59	14.64	21.25	15.45	18.97	18.36	17.28	19.70
GPT-4o	24.92	23.38	21.38	25.28	26.91	18.04	26.15	22.86	16.93
<b>Medical-specialized LLMs</b>									
ChatDoctor	8.54	11.26	5.09	4.62	5.73	7.09	9.93	13.65	8.17
BioMistral	9.93	12.45	8.14	7.88	8.96	8.12	9.26	12.43	8.28
Yi-34B	16.01	16.37	16.59	17.27	18.50	16.58	17.09	16.31	16.81
DiagnosisGPT-34B	17.02	18.07	18.27	19.95	18.35	18.51	18.17	18.58	18.01
<b>Few-shot Demo. w/ Instruction</b>									
Qwen3-32B	18.77	15.89	15.94	20.49	14.83	16.96	17.07	18.99	16.67
QwQ-32B	15.38	14.77	20.93	19.94	13.92	17.14	15.16	13.52	20.07
<b>Enhancing Reasoning with iterative RAG</b>									
RAG-Qwen3-32B									
w/ Query Planning	21.30	22.81	20.10	24.31	25.15	21.08	24.47	22.40	20.73
w/ <b>AutoClinician</b>	26.54	26.92	23.37	29.55	31.09	27.32	30.68	29.14	24.98
RAG-QwQ-32B									
w/ Query Planning	20.38	21.59	19.64	22.25	23.45	21.97	23.36	22.28	21.70
w/ <b>AutoClinician</b>	26.72	27.48	24.59	27.08	29.71	25.34	27.62	26.87	27.15

324 4.3 MAIN RESULTS (Q1 & Q2)

325

326 For space limitation, experimental results are presented in Table 1 and Table 5 in Appendix C.3. We can

327 make the following observations: (1) **Vanilla zero-shot prompting yields limited diagnostic performance**

328 **on EHRs reasoning**. Both general-purpose and medical-specialized LLMs achieve low accuracy and MRR,

revealing that they lack deep understanding of EHR data and have little domain knowledge to support diagnostic reasoning. (2) **Few-shot prompting with instructions does not improve diagnostic reasoning ability and can even degrade performance.** This result reflects both the limited reasoning ability of LLMs in complex diagnostic workflows and the insufficiency of few examples in providing generalizable diagnosis knowledge. (3) **Incorporating clinical guidelines substantially boosts performance.** RAG with query planning methods outperform other baselines with a clear margin, demonstrating that structured clinical evidence is essential for accurate diagnostic reasoning. Furthermore, AutoClinician demonstrates remarkable performance gain over others, underscoring the advantage of our proposed structured evidence integration.

#### 4.4 DETAILED ANALYSIS

**Importance of CEGs (Q3).** We investigate the contribution of guideline compression through a series of ablations, with results summarized in Figure 2 and Figure 3. Specifically, we compare with: (1) AutoClinician w/o Self-refinement, which removes the self-consistency refinement applied during CEG construction; and (2) AutoClinician w/o CEGs, which appends the original guideline text directly to input prompts.

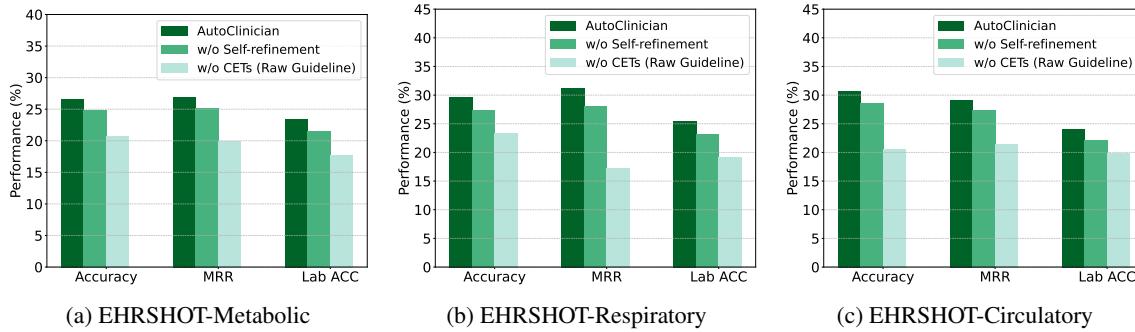


Figure 2: Ablation study of CEGs on the EHRSHOT dataset.

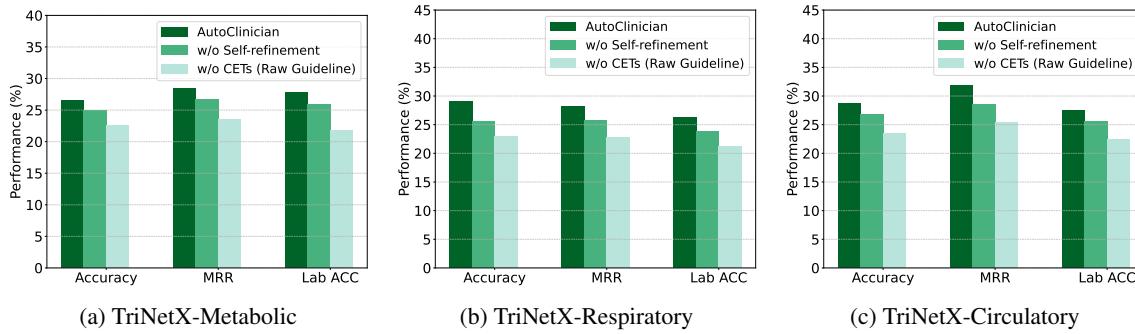


Figure 3: Ablation study of CEGs on the TriNetX dataset.

AutoClinician consistently achieves the highest performance, highlighting the critical role of leveraging structured knowledge from clinical guidelines for diagnostic reasoning. Eliminating self-refinement produces a moderate decline, indicating that refinement improves the qualities of CEGs. Removing CEGs performs better than zero-shot prompting but remain substantially worse than AutoClinician. This indicates that compressing guideline text help LLMs to focus on essential logical dependencies, thereby improving its capacity for multi-step diagnostic reasoning.

376 **Importance of DFA (Q4).** We conduct ablation studies to assess the effect of modeling diagnostic reasoning  
 377 as a DFA within AutoClinician, as shown in Table 2 and Table 3. Specifically, we compare against two  
 378 variants: (1) AutoClinician w/o DFA, which removes the state tracker and directly augments the prompts with  
 379 the retrieved CEG; (2) CoT Prompting, which replaces structured DFA with manually written reasoning cues.  
 380

381 Table 2: Ablation study of DFA on the EHRSHOT dataset.  
 382

Reasoning Method	EHRSHOT-Metabolic			EHRSHOT-Respiratory			EHRSHOT-Circulatory		
	Acc	MRR	Lab ACC	Acc	MRR	Lab ACC	Acc	MRR	Lab ACC
AutoClinician	26.54	26.92	23.37	29.55	31.09	27.32	30.68	29.14	24.98
w/o DFA	20.83	20.30	18.55	24.73	24.53	22.76	25.28	23.05	19.74
CoT	21.12	22.62	20.28	25.11	25.86	23.32	24.96	24.58	22.75

383 Table 3: Ablation Study of DFA on the TriNetX dataset.  
 384

Reasoning Method	TriNetX-Metabolic			TriNetX-Respiratory			TriNetX-Circulatory		
	Accuracy	MRR	Lab Acc	Accuracy	MRR	Lab Acc	Accuracy	MRR	Lab Acc
AutoClinician	26.50	28.50	27.80	29.00	28.20	26.33	28.70	31.80	27.40
w/o DFA	22.76	23.21	22.45	23.83	22.60	20.81	23.91	24.17	23.22
CoT	22.30	21.88	20.01	24.47	23.65	23.22	23.45	25.41	24.90

385 AutoClinician achieves the highest performance across all datasets, highlighting the importance of maintaining  
 386 and updating patient-specific reasoning states derived from both the EHR and CEGs. Accurate reasoning  
 387 requires conducting sufficient laboratory tests, interpreting value thresholds, and integrating multiple clinical  
 388 findings for joint evaluation, leading to AutoClinician’s improved performance over CoT.  
 389

390 **Effect of the Number of CEG Candidates (Q5).** We further analyze the model’s sensitivity to the number  
 391 of candidate CEGs. Experiments are conducted on the EHRSHOT dataset with Qwen3-32B as the backbone,  
 392 with results summarized in Table 4. It can be observed that increasing the candidate number leads to higher  
 393 accuracy score. We leave the exploration of more advanced retrieval mechanisms to future work.  
 394

402 Table 4: Accuracy of AutoClinician with different number of candidate CEGs across three EHRSHOT subsets.  
 403

Model	Metabolic			Respiratory			Circulatory					
	Acc@1	Acc@3	Acc@5	Acc@10	Acc@1	Acc@3	Acc@5	Acc@10	Acc@1	Acc@3	Acc@5	Acc@10
AutoClinician	19.47	23.31	26.54	27.95	21.12	26.29	29.55	31.61	18.86	22.15	30.68	30.76

408 **Human Evaluation and Case Study (Q6).** We perform a case study with the collaboration of clinicians to  
 409 review approximately 30 generated dialogues and evaluate the guideline-grounded reasoning from multiple  
 410 perspectives, including guideline adherence, laboratory test interpretation, conflict and ambiguity resolution,  
 411 and overall clinical plausibility. Detailed evaluation results are provided in the Appendix C.4.3.  
 412

## 5 CONCLUSIONS AND FUTURE WORKS

414 In this work, we introduced AutoClinician, a training-free framework that integrates structured clinical  
 415 guidelines into EHR-based diagnostic conversations. We first compress clinical guidelines into logical  
 416 evidence graphs and propose a self-consistency strategy to evaluate and refine the generated structures.  
 417 Diagnostic reasoning is further modeled as a Deterministic Finite Automaton (DFA) that tracks patient-  
 418 specific diagnostic states using sequential EHR-based inputs and guideline-derived graphs. Experiments  
 419 demonstrate that AutoClinician consistently outperforms both general-purpose and clinical-specialized LLMs.  
 420 Future work will extend AutoClinician beyond EHR-only prediction by incorporating multimodal input  
 421 and aligning guidelines with richer diagnostic inputs. Another direction is to develop AutoClinician into a  
 422 web-based agent that dynamically leverages public medical resources to support evidence-grounded patient  
 423 self-interpretation and physician decision-making.

## 423 ETHICAL STATEMENT

424

425 All data used in this study are fully de-identified. The datasets are available to researchers for academic  
 426 purposes, and access to certain datasets and documents requires institutional membership and appropriate  
 427 approvals. No personally identifiable information (PII) or sensitive clinical records were used. This study is  
 428 intended to model clinical reasoning processes and to use LLMs for response generation, without any attempt  
 429 to identify, profile, or make inferences about real individuals.

## 430

## 431 REPRODUCIBILITY STATEMENT

432

433 We are committed to ensuring the reproducibility of our work. All datasets used in this study are either  
 434 publicly available or accessible to academic researchers under appropriate agreements. We provide the full  
 435 implementation, including preprocessing scripts and modeling pipelines, via an anonymous link in this paper.  
 436 The prompts used in our experiments as well as detailed implementation descriptions are included in the  
 437 Appendix. These resources enable independent researchers to reproduce our results and extend our work.

## 438

## 439 REFERENCES

440

441 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo  
 442 Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*  
 443 *arXiv:2303.08774*, 2023.

444

445 Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan  
 446 Han, Zhiyong Lu, and Zhe He. Preliminary analysis of the impact of lab results on large language model  
 447 generated differential diagnoses. *npj Digital Medicine*, 8(1):166, 2025.

448

449 Som S Biswas. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869, 2023.

450

451 Derun Cai, Chenxi Sun, Moxian Song, Baofeng Zhang, Shenda Hong, and Hongyan Li. Hypergraph  
 452 contrastive learning for electronic health records. In *Proceedings of the 2022 SIAM International Conference*  
 453 on *Data Mining (SDM)*, pp. 127–135. SIAM, 2022.

454

455 Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards  
 456 an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*, 2024.

457

458 Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM*  
 459 *sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

460

461 Yixiang Chen, Penglei Sun, Xiang Li, and Xiaowen Chu. Mrd-rag: Enhancing medical diagnosis with  
 462 multi-round retrieval-augmented generation. *arXiv preprint arXiv:2504.07724*, 2025.

463

464 Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-  
 465 based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD*  
 466 *international conference on knowledge discovery and data mining*, pp. 787–795, 2017.

467

468 Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.  
 469 *Psychological bulletin*, 70(4):213, 1968.

470

471 Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for  
 472 chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.

473

474 Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge, 2017.

470 Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu,  
 471 Irene Li, and Douglas Teodoro. Retrieval-augmented generation in biomedicine: A survey of technologies,  
 472 datasets, and clinical applications. *arXiv preprint arXiv:2505.01146*, 2025.

473

474 Yuxin He, Buzhou Tang, and Xiaoling Wang. Generative models for automatic medical decision rule  
 475 extraction from text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language  
 476 Processing*, pp. 7034–7048, 2024.

477 Ruihui Hou, Xiaojun Wang, Weiyan Zhang, Zhexin Song, Kai Wang, Yifei Chen, Jingping Liu, and Tong  
 478 Ruan. Decision tree extraction for clinical decision support system with if-else pseudocode and planselect  
 479 strategy. *IEEE Journal of Biomedical and Health Informatics*, 2025.

480

481 Suramya Jadhav, Abhay Shanbhag, Sumedh Joshi, Atharva Date, and Sheetal Sonawane. Maven at mediqua-  
 482 corr 2024: Leveraging rag and medical llm for error detection and correction in medical notes. In  
 483 *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 374–381, 2024.

484 Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through  
 485 retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40  
 486 (Supplement\_1):i119–i129, 2024.

487

488 Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions  
 489 with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*, 2023.

490

491 Yiwen Jiang, Hao Yu, and Xingyue Fu. Medical decision tree extraction: A prompt based dual contrastive  
 492 learning method. In *China Health Information Processing Conference*, pp. 103–116. Springer, 2022.

493

494 Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349  
 (6245):255–260, 2015.

495

496 Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rovier, and Richard  
 497 Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains.  
 498 *arXiv preprint arXiv:2402.10373*, 2024.

499

500 Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. Meddm: Llm-executable clinical guidance  
 501 tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*, 2023a.

502

503 Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huiyi Yu, Loni Hagen, Yongfeng Zhang,  
 504 Themistocles L Assimes, Libby Hemphill, et al. A scoping review of using large language models (llms) to  
 505 investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*, 2024.

506

507 Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat  
 508 model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15  
 509 (6), 2023b.

510

511 Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P  
 512 Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. Clinical entity augmented  
 513 retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45, 2025.

514

515 Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy Chang, and Wei Wang.  
 516 Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision  
 517 making. *arXiv preprint arXiv:2406.09923*, 2024.

518

519 Jing Mei, Haifeng Liu, Guotong Xie, Shengping Liu, and Baoyao Zhou. An ocl-compliant gello engine. In  
 520 *User Centred Networked Health Care*, pp. 130–134. IOS Press, 2011.

517 Marvin L Minsky. Some universal elements for finite automata. *Automata studies*, 34:117–128, 1956.

518

519 Jonathan Montomoli, Luca Romeo, Sara Moccia, Michele Bernardini, Lucia Migliorelli, Daniele Berardini,  
520 Abele Donati, Andrea Carsetti, Maria Grazia Bocci, Pedro David Wendel Garcia, et al. Machine learning  
521 using the extreme gradient boosting (xgboost) algorithm predicts 5-day delta of sofa score at icu admission  
522 in covid-19 patients. *Journal of Intensive Medicine*, 1(02):110–116, 2021.

523

524 Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-  
525 box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness,  
accountability, and transparency*, pp. 629–639, 2020.

526

527 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv  
preprint arXiv:1908.10084*, 2019.

528

529 Weijieying Ren, Xiaoting Li, Huiyuan Chen, Vineeth Rakesh, Zhuoyi Wang, Mahashweta Das, and Vasant G  
530 Honavar. Tablog: Test-time adaptation for tabular data using logic rules. In *Forty-first International  
531 Conference on Machine Learning*, 2024.

532

533 Savyasachi V Shah. Accuracy, consistency, and hallucination of large language models when analyzing  
534 unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953–e2425953,  
535 2024.

536

537 Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent  
538 advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical  
and health informatics*, 22(5):1589–1604, 2017.

539

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

543

544 Maria Camila Suarez Arbelaez, Sirpi Nackeeren, Khushi Shah, Ruben Blachman-Braun, Isaac Bronson,  
545 Maxwell Towe, Abhishek Bhat, Robert Marcovich, Ranjith Ramasamy, and Hemendra N Shah. Association  
546 between body mass index, metabolic syndrome and common urologic conditions: a cross-sectional study  
547 using a large multi-institutional database from the united states. *Annals of medicine*, 55(1):2197293, 2023.

548

549 Vivek Subbiah. The next generation of evidence-based medicine. *Nature medicine*, 29(1):49–58, 2023.

550

551 QwQ Team. Qwq-32b: Embracing the power of reinforcement learning, 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.

552

553 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay  
554 Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and  
fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

555

556 Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning  
llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023a.

557

558 Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and  
559 Rishita Anubhai. Biobridge: Bridging biomedical foundation models via knowledge graph. *arXiv preprint  
arXiv:2310.03320*, 2023b.

560

561 Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark  
562 for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:  
563 67125–67137, 2023.

564 Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview  
565 of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica  
566 Sinica*, 10(5):1122–1136, 2023.

567 Zihong Wu. Research on decision tree method of medical text based on information extraction. In *China  
568 Health Information Processing Conference*, pp. 127–133. Springer, 2022.

569 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu,  
570 Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

571 Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan.  
572 *Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback  
573 and real-world multi-turn dialogue*. In *Proceedings of the AAAI Conference on Artificial Intelligence*,  
574 volume 38, pp. 19368–19376, 2024b.

575 Hao-Ren Yao, Nairen Cao, Katina Russell, Der-Chen Chang, Ophir Frieder, and Jeremy T Fineman. Self-  
576 supervised representation learning on electronic health records with graph kernel infomax. *ACM Transac-  
577 tions on Computing for Healthcare*, 5(2):1–28, 2024.

578 Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng  
579 Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint  
580 arXiv:2403.04652*, 2024.

581 Zaifu Zhan, Shuang Zhou, Xiaoshan Zhou, Yongkang Xiao, Jun Wang, Jiawen Deng, He Zhu, Yu Hou, and  
582 Rui Zhang. Retrieval-augmented in-context learning for multimodal large language models in disease  
583 classification. *arXiv preprint arXiv:2505.02087*, 2025.

584 Wei Zhu, Wenfeng Li, Xing Tian, Pengfei Wang, Xiaoling Wang, Jin Chen, Yuanbin Wu, Yuan Ni, and Guo-  
585 tong Xie. Text2mdt: extracting medical decision trees from medical texts. *arXiv preprint arXiv:2401.02034*,  
586 2024.

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

---

611	APPENDIX	
612		
613		
614	<b>A The Use of Large Language Models</b>	<b>15</b>
615		
616	<b>B Implementation</b>	<b>15</b>
617		
618	B.1 Implementation Details . . . . .	15
619	B.2 Example Prompt for Extracting Text from Clinical Guidelines . . . . .	16
620	B.3 Example Prompt for Clinical Evidence Graph Generation . . . . .	17
621	B.4 Example Prompt for Clinical Evidence Graph Refinement . . . . .	17
622	B.5 Prompts for Dialogue Generation using DFA . . . . .	18
623	B.6 CoT prompts . . . . .	18
624		
625		
626		
627	<b>C Additional Experiments</b>	<b>18</b>
628		
629	C.1 Dataset Description . . . . .	18
630	C.2 Evaluation Metrics . . . . .	19
631	C.3 Main Results . . . . .	19
632	C.4 Detailed Analysis . . . . .	19
633	C.4.1 Examples of the Generated Pseudo-patient Samples . . . . .	19
634	C.4.2 Human Evaluation on CEGs and Case Study . . . . .	19
635	C.4.3 Human Evaluation on Dialogues and Case Study . . . . .	25
636	C.5 Sensitivity Analysis on Different Judge threshold . . . . .	26
637	C.6 Ablation Studies on different pseudo-patient types. . . . .	26
638	C.7 Human Inter-rater Reliability . . . . .	27
639	C.8 Sensitivity to Different Judge Models. . . . .	27
640	C.9 Analysis on Fallback Rate . . . . .	27
641	C.10 Results of Different open-source Base Models for CEG Generation and Refinement . . . . .	28
642	C.11 Statistical Significance . . . . .	28
643		
644		
645		
646		
647		
648		
649		
650		
651		
652		
653		
654		
655		
656		
657		

658 **A THE USE OF LARGE LANGUAGE MODELS**  
659660 Large Language Models (LLMs) were employed in this project in a limited, assistive capacity. For manuscript  
661 preparation, all sections were drafted by the authors, with LLMs occasionally used to improve grammar,  
662 enhance clarity, and refine academic tone. During the implementation phase, LLMs served as coding  
663 assistants, providing code completions, debugging suggestions, and formatting support. However, all final  
664 code, experimental design, analyses, and validations were developed and verified solely by the authors.  
665 Importantly, LLMs were not used for generating research ideas, designing experiments, or conducting the  
666 literature review. All conceptual contributions, methodological innovations, and scientific insights originated  
667 exclusively from the authors.  
668669 **B IMPLEMENTATION**  
670671 **B.1 IMPLEMENTATION DETAILS**  
672673 In our experiments, we first extract diagnosis-relevant text from clinical guidelines and then summarize  
674 this content into Clinical Evidence Graphs (CEGs). Each CEG consists of a set of symptoms serving as  
675 query targets and a graph that represents the diagnostic logic derived from the guideline. We utilize the  
676 all-MiniLM-L6-v2 model (Reimers & Gurevych, 2019) as the embedding model to index both the CEGs and  
677 input queries. At test time, each input query is encoded and compared against all embeddings in the corpus,  
678 and the top-5 nearest neighbors are retrieved for inference.  
679680 For HTML-based guidelines<sup>3</sup>, we use the Python library `lxml` to traverse the DOM structure and extract  
681 section headers and corresponding content blocks. Tags such as `<h2>`, `<p>`, and `<li>` are used to identify  
682 semantic structure. For PDF-based guidelines, we convert documents to structured HTML using the Python  
683 library PyMuPDF, preserving layout features including heading fonts, indentation, and section boundaries.  
684 We then apply regular expression and keyword-based filters to identify diagnosis-relevant sections. Candidate  
685 sections are selected based on the presence of headings or sentence patterns containing terms such as  
686 “diagnosis,” “screening,” “classification,” “introduction,” or “criteria.” The extracted content is treated as  
687 diagnosis-related context and serves as input for CEGs generation.  
688689 For Clinical Evidence Graph (CEG) generation, we utilize GPT-4o for more accurate guideline parsing. We  
690 first extract diagnosis-related sections of clinical guidelines with the prompt in Figure 4, and the extracted text  
691 is then passed to the CEG generation prompt (see Figure 5). The model is instructed to output a structured  
692 JSON object consisting of verbatim symptoms as query key and step-wise diagnostic reasoning represented  
693 as graph nodes. Each node links context, trigger, and state, with deterministic transitions ensuring that every  
694 non-terminal state matches the context of a subsequent node. In our preliminary experiments, we identified  
695 four types of potential errors in CEGs. To address these, we generated pseudo-patient samples directly from  
696 the diagnosis-related content (as shown in Figure 6), and then used these samples to refine the CEGs (as  
697 shown in Figure 7). We exclude clinical tests with a frequency of less than 10 across the entire dataset.  
698 Additionally, we filter out tests for which no test results are available for each patient.  
699700 In our experiments, LLMs are instructed to output answers in descending order of confidence, forming a  
701 ranked list. We adopt the all-MiniLM-L6-v2 model (Reimers & Gurevych, 2019) to embed both the LLMs’  
702 responses and the ground-truth responses for diagnoses or laboratory tests. When the base model is QwQ-32B,  
703 we additionally employ Qwen3-14B to extract and summarize reasoning results, which are then compared  
704 against the ground truth for evaluation.  
705706 

---

707 <sup>3</sup>[https://diabetesjournals.org/care/article/48/Supplement\\_1/S27/157566/2-Diagnosis-and-Classification-of-Diabetes](https://diabetesjournals.org/care/article/48/Supplement_1/S27/157566/2-Diagnosis-and-Classification-of-Diabetes)

705 To enable semantic processing of structured EHR data, we mapped coded medical concepts to their tex-  
 706 tual descriptions. Specifically, we downloaded three standardized clinical vocabularies: (1) International  
 707 Classification of Diseases (ICD) codes from the World Health Organization (WHO, <https://www.who.int/standards/classifications/classification-of-diseases>), (2) the SNOMED CT  
 708 International Release (20250801) from the SNOMED International Member Licensing and Distribution  
 709 Service (MLDS, <https://mlds.ihtsdotools.org>), and (3) RxNorm (2025AA) from the U.S. Na-  
 710 tional Library of Medicine (NLM, <https://www.nlm.nih.gov/research/umls/rxnorm>). Each  
 711 resource provides machine-readable mappings from coded identifiers to human-readable terms, which we use  
 712 to convert diagnosis, procedure, and medication codes into natural language representations. For example,  
 713 the ICD-10 code E11.9 corresponds to the description “Type 2 diabetes mellitus without complications,”  
 714 the SNOMED CT concept 44054006 maps to “Diabetes mellitus type 2 (disorder),” and the RxNorm  
 715 code 860975 represents “Metformin 500 MG Oral Tablet.” These mappings allow EHR data to be directly  
 716 integrated into text.  
 717

## 718 B.2 EXAMPLE PROMPT FOR EXTRACTING TEXT FROM CLINICAL GUIDELINES

### 720 **Prompt for Extracting Diagnosis-related Text from Clinical Guidelines**

721 You are a clinical information extractor. You read guideline pages and extract verbatim diagnosis-relevant  
 722 content.

723 Input: <An Guideline file> describing <Guideline Title>.

724 Task: From the guideline file, localize and extract only the sections that directly support diagnosis prediction,  
 725 including (but not limited to):

- 726 1. Symptoms/signs/demographics explicitly mentioned (verbatim phrases only)
- 727 2. Diagnostic lab tests (test names, specimen, timing, test procedures) 3. Cut-points / thresholds and normal/ab-  
 728 normal ranges (all units included)
- 729 4. Conditions/contraindications/limitations where a test should not be used or is unreliable
- 730 5. Confirmation rules (e.g., repeat testing, discordant results handling)
- 731 6. Risk-based screening criteria only if they contain thresholds or explicit test instructions relevant to diagnosis

732 Inclusions and Exclusions:

- 733 1. Include only content that is directly diagnosable signals (symptoms/signs), test definitions, numeric threshold-  
 734 s/ranges, confirmation rules, and test limitations.
- 735 2. Exclude epidemiology, pathophysiology discussions, economic or implementation commentary, treatment/-  
 736 management recommendations, and non-diagnostic background unless it contains explicit diagnostic thresholds  
 737 or rules.

738 Requirements: Follow the original document order. For each extracted block, add the original section heading  
 739 (verbatim) as a json header before the content. Quality and Formatting requirements (must-follow):

- 740 1. Output one json file only.
- 741 2. Use original text verbatim; no rephrasing.
- 742 3. For tables, use Markdown tables; do not insert long sentences into tables-only the cells as in source.
- 743 4. Preserve units, symbols (e.g., >=,<), and footnote markers.
- 744 5. If a referenced table/figure is outside the text but the table is present, include it. If a table is an image, write a  
 745 short line: Table image present.
- 746 6. If a section references another table (e.g., "Table 2.2"), and that table exists in the guideline, include it next.
- 747 7. If any segment cannot be confidently localized, add a one line placeholder Section not found in guideline and  
 748 continue.

749 Figure 4: Prompt for Extracting Text from Clinical Guidelines. This prompt instructs a LLM to transform  
 750 clinical guideline into a structured json file that preserves verbatim diagnosis-relevant evidence (symptoms,  
 751 tests, thresholds, confirmation rules, and limitations) while excluding all non-diagnostic material.

752 Example prompt is shown in Figure 4.  
 753

754  
 755 **B.3 EXAMPLE PROMPT FOR CLINICAL EVIDENCE GRAPH GENERATION**  
 756

757  
 758 **Example Prompt for Generating CEGs**

759  
 760 You are an clinical graduates trained to undertand clinical guideline texts and extract structured diagnostic logic.  
 761 Given a clinical guideline text focused on diagnosis, your task is to output a JSON object that contains:

- 762 1. Symptom: A list of clinical symptoms, patient demographics, or signs explicitly mentioned in the text that are  
 763 associated with the diagnosis. Extract what is written and do not infer or paraphrase. Use the exact wording from  
 764 the text.
- 765 2. Graphs: A sequence of diagnostic reasoning steps represented in the following format:  
 766   - ID: Denotes the index of current node in the graph.  
 767   - Context\*: The current diagnostic state or clinical situation. This links to the ““state”“ of the previous step.  
 768   - Trigger: Specifies the clinical condition or threshold defined by guidelines. A threshold, condition, lab result,  
 769 or observed symptom that triggers diagnosis progression.  
 770   - State: Specifies the guideline-prescribed clinical action taken when the Trigger is activated. It reflects how to  
 771 act upon the observed condition based on clinical guidelines. For example, ‘State = Suspect diabetes’ when the  
 772 patient’s laboratory tests satisfy the defined Trigger.  
 773   - Termination: true or false. If true, this node is a leaf (final diagnosis or conclusion). If false, the engine will  
 774 continue by locating the next node where context equals to this node state.  
 775 For every node with “terminal”: false, the value of “state” MUST exactly match the “context” of a subsequent  
 776 node. Do NOT invent text.

777 Instructions:

- 778   - Use only information that is explicitly stated in the input text. Maintain the chain of reasoning by linking  
 779 each step ‘context’ to the ‘state’ of the prior step.
- 780   - Output the result strictly in JSON format.
- 781   - Do not include any commentary, explanation, or extra text, only the JSON object.

782 Input (diagnosis-related sections): <INSERT TEXT HERE>

783 Output JSON format only. Output format:

784   “Symptom”: [“<verbatim symptom/sign/demographic>”, “...”],  
 785   “Graphs”: [ “Context”: “<where we are now>”, “Trigger”: [“<verbatim trigger 1>”, “<verbatim trigger  
 786 2>”], “State”: “<resulting status>”, “Terminal”: False , “Context”: “<must equal previous state if previous  
 787 Terminal=false>”, “Trigger”: [“<verbatim trigger>”], “State”: “<final diagnosis or next status>”, “Terminal”:  
 788 True ]

789  
 790 Figure 5: Example Prompt for Generating CEGs. This prompt instructs the model to convert diagnosis-related  
 791 guidelines into a structured JSON-based Clinical Evidence Graph, capturing verbatim symptoms and  
 792 stepwise diagnostic reasoning logic.

793 Example prompt is shown in Figure 5.

794  
 795 **B.4 EXAMPLE PROMPT FOR CLINICAL EVIDENCE GRAPH REFINEMENT**

796 Prompt for pseudo-patient example generation is shown in Figure 6. If the alignment score between  
 797 pseudo-patient examples and CEGs falls below 0.8, we prompt LLM with failed pseudo-patient examples to  
 798 refine the initial CEGs. Example prompt for CEGs refinement is shown in Figure 7.

799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  

### Example Prompt for Generating Pseudo-patient Examples

You are a clinical reasoning expert. You are given a diagnosis-related clinical guideline passage `INPUT` `DIAGBISTIC RELATED SECTIONS`. Your task is to generate synthetic patient cases to test the robustness of a diagnostic logic graph automatically compressed from that guideline.

Follow these instructions:

1. `**Read the guideline**` carefully and identify key diagnostic criteria: symptoms, lab thresholds, age/BMI ranges, and test sequences.

2. `**Generate pseudo-patient example samples**` that intentionally challenge the logic tree understanding. Each sample should fall into one of the following categories:

- Incomplete Evidence: A required symptom, test, or condition is missing.
- Threshold Misalignment: The input includes a value just above or below a diagnostic threshold.
- Sequential Logic Error: The order of diagnosis steps is violated (e.g., insulin started before testing).
- Contradictory Evidence Injection: The patient presents with mixed evidence that makes diagnosis ambiguous.

3. For each sample, return the following fields in JSON format:
  - Patient Input: Include age, BMI, symptoms, lab results, relevant family or autoimmune history.
  - Type: One of: "Incomplete Evidence", "Threshold Misalignment", "Sequential Logic Error", or "Contradictory Evidence Injection".
  - Expected Output: What the correct diagnosis should be according to the guideline.
  - Guideline Evidence: A quoted or paraphrased sentence from the guideline that justifies the expected output.

Please generate 15 samples for each of the four types (total around 50 examples), and ensure all reasoning strictly follows the logic stated in the guideline.

819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
Figure 6: Prompt for Generating Pseudo-patient Examples. This prompt is used to generate structured pseudo-patient examples that deliberately challenge diagnostic logic graphs, ensuring robustness against incomplete evidence, borderline thresholds, sequential errors, and contradictory clinical findings.

## B.5 PROMPTS FOR DIALOGUE GENERATION USING DFA

Example Prompt for using DFA as a dialogue tracker can be found in Figure 8.

## B.6 COT PROMPTS

Example prompt for CoT used in baselines can be found in Figure 9.

## C ADDITIONAL EXPERIMENTS

### C.1 DATASET DESCRIPTION

**EHRSHOT** is a de-identified, public large-scale benchmark of electronic health records from 6,739 adult patients at Stanford Medicine (Wornow et al., 2023). It contains 41.6 million structured clinical events across 921,499 encounters, spanning demographics, diagnoses, procedures, laboratory results, and medications.

**TriNetX Dataset**<sup>4</sup> is a global health research network providing access to large-scale de-identified patient EHR data (Suarez Arbelaez et al., 2023). TriNetX includes records from 33,105 de-identified patients, spanning data from 1982 to 2023, collected across more than 100 community hospitals and 500 outpatient clinics.

<sup>4</sup><https://trinetx.com/>.

## Prompt for Refining CEGs using Failed Pseudo-patient Examples.

You are given a Clinical Evidence Graph (CEG) represented as nodes with attributes `<ID, Context, Trigger, State, Termination>`. You are also provided with a set of pseudo-patient examples, each synthesized directly from the guideline text and containing verbatim guideline conditions (e.g., laboratory thresholds, diagnostic criteria, symptom descriptions).

Here is the CEGs <CEGs input> and pseudo-example <pseudo-examples> with the type of <error type>.

### Instructions:

For each pseudo-example, follow the reasoning path in the CEG by matching the Trigger  $\rightarrow$  State  $\rightarrow$  Context transitions. If a pseudo-example cannot be mapped consistently, identify the misaligned nodes.

Refine the CEG by:

Pruning inconsistent or redundant nodes.

Adjusting node attributes only using conditions and wording explicitly stated in the guideline text.

Ensuring that all pseudo-examples can be faithfully explained by the refined CEG.

## Output format:

Return the refined CEG in the same 5-tuple format <ID, Context, Trigger, State, Termination>. Additionally, include a concise note summarizing which nodes were modified or pruned, and cite the exact guideline phrase used for the refinement.

Figure 7: Prompt for Refining CEGs using failed pseudo-patient examples.

## C.2 EVALUATION METRICS

In our experiments, LLMs are instructed to output answers in descending order of their confidence, forming a prioritized ranked list. The Mean Reciprocal Rank (MRR) (Jordan & Mitchell, 2015) is a metric used to evaluate agents that return a ranked list of answers to queries, focusing on the position of the first relevant answer. A higher MRR indicates that relevant items tend to appear higher in the ranked list of results. It is defined as the average of the reciprocal ranks of the first relevant answer for a set of queries.

### C.3 MAIN RESULTS

Our main results on TrinexT dataset is shown in Table 5.

## C.4 DETAILED ANALYSIS

#### C.4.1 EXAMPLES OF THE GENERATED PSEUDO-PATIENT SAMPLES

We generate four types of pseudo-patient examples to automatically validate and refine CEGs: (1) ***Evidence Modification***. For example, we randomly remove a symptom or a required lab test. Examples are shown in Figure 10. (2) ***Boundary Threshold Perturbation***. For example, given a guideline threshold of 100 mg/dL, a perturbed case at 98 mg/dL should be rejected by the tree to ensure threshold adherence. Examples are shown in Figure 11. (3) ***Step Reordering*** that reverses the order of diagnostic steps. Examples are shown in Figure 12. (4) ***Contradictory Evidence Injection*** that introduces inconsistency between test results. Examples are shown in Figure 13.

#### C.4.2 HUMAN EVALUATION ON CEGs AND CASE STUDY

893  
894  
895  
896897 **Example Prompt for Dialogue Reasoning using DFA**

898

899 You operate a Deterministic Finite Automaton (DFA) defined as  $(S, \Sigma, \delta, s_0, F)$ .900  $S$ : a finite set of conversational states with start state  $s_0$  and terminal states  $F$ .901  $\Sigma$ : input categories derived from CEG-guided reasoning over the patient's dialogue (based on Retrieved CEG  
and patient Input).902  $\delta$ : a deterministic transition function (provided in DFA specification).

903

**Requirements:**904 1) Read <Retrieved CEG> and the user's dialogue to determine which triggers are satisfied, which values  
905 are missing, and whether results are discordant.906 2) From those conclusions, derive  $\Sigma$  at each step (choose one of: ConditionSatisfied, Missingness,  
907 Conflict, Sufficiency, Failure).908 3) Apply  $\delta$  deterministically to move from the current state in  $S$  to the next, repeating step by step for up to  
909 MAX\_STEPS=15, or stop early at a terminal state.910 4) Perform all intermediate reasoning internally; **DO NOT** reveal intermediate steps.

911

**DFA specification:**

912

913  $S = \{\text{Start, PartialResults, ConflictResults, ConfirmedDiagnosis,}$ 

914

Unresolved, AnotherCandidate}

915

 $\Sigma = \{\text{ConditionSatisfied, Missingness, Conflict, Sufficiency, Failure}\}$ 

916

 $\delta(s, \text{Missingness}) \rightarrow \text{PartialResults}$ 

917

 $\delta(s, \text{ConditionSatisfied}) \rightarrow \text{PartialResults (until a terminal True node is reached)}$ 

918

 $\delta(s, \text{Conflict}) \rightarrow \text{ConflictResults}$ 

919

 $\delta(s, \text{Sufficiency}) \rightarrow \text{ConfirmedDiagnosis}$ 

920

 $\delta(s, \text{Failure}) \rightarrow \text{AnotherCandidate}$ 

921

Terminal states = {ConfirmedDiagnosis, Unresolved}

922

**Retrieved CEG:**

923

&lt;Input Retrieved CEGs&gt;

924

FORMAT: Return ONE sentence only.

925

&lt;Current Patient Input&gt;

926

**Output Requirement:**

927

Output **EXACTLY ONE** sentence in plain English, ending with a period. The sentence MUST include:

928

(a) a concise, guideline-based explanation of the patient's current status (from CEG + inputs), and  
929 (b) the current state and the most appropriate next step implied by your DFA traversal (e.g., which test is missing,  
whether results conflict, or that diagnosis is confirmed).

930

No bullet points, no lists, no JSON, no headings, no code fences, no extra lines and ONE sentence only.

931

932

933 Figure 8: Example prompt used for dialogue reasoning with a DFA, specifying the state set, input categories,  
934 and transition rules, along with requirements for processing retrieved CEGs and patient inputs to generate a  
935 guideline-grounded sentence summarizing the patient's status and next step.

936

937

938

939

940  
941  
942  
943**CoT Prompts.**944 You are a clinical expert assistant. Given a patient's EHR, reason step by step to reach a diagnosis. First, examine  
945 the symptoms and lab results. Then, check if diagnostic thresholds are met. Finally, determine the disease  
946 subtype and provide a diagnosis based on medical guidelines.

947 Patient EHR: &lt;insert input&gt;

948 Step-by-step reasoning:

- 949 1. What symptoms does the patient report? Are they consistent with any common disease?
- 950 2. What are the clinical semantic meaning of lab test results, e.g., abnormal or normal? Are any lab test results  
above diagnostic thresholds?
- 951 3. Based on symptoms and lab data, can we confirm a diagnosis?
- 952 4. What is the likely disease subtype based on age, BMI, and symptom onset?
- 953 5. What is the final diagnosis?

954 Answer:

955

956 Figure 9: Example of Cot Prompt used in Baselines.  
957  
958  
959  
960  
961**Examples for Incomplete Evidence.**

962 Example 1:

963 "Patient Input": "age": 29, "BMI": 27.2, "symptoms": [], "A1C": 6.6 %, "pregnant": true, "trimester": 2,  
964 "Type": "Incomplete Evidence.",  
965 "Expected Output": "Not diagnostic; Do not use A1C to diagnose in pregnancy; apply plasma glucose  
966 criteria.",  
967 "Guideline Evidence": "In conditions... pregnancy (second and third trimesters...) only plasma blood  
968 glucose criteria should be used to diagnose diabetes."

969 Example 2:

970 "Patient Input": "age": 46, "BMI": 31.4, "symptoms": ["fatigue"], "FPG": 128 mg dL, "fasting\_hours": 5,  
971 "A1C": 6.3 %,  
972 "Type": "Incomplete Evidence.",  
973 "Expected Output": "Not diagnostic; fasting status invalid (<8 h). Repeat FPG with  $\geq$  8 h fast or use another  
974 test.", "Guideline Evidence": "FPG  $\geq$  126 mg/dL... Fasting is defined as no caloric intake for at least 8 h."  
975 Example 3:

976 "Patient Input": "age": 41, "BMI": 28.0, "symptoms": [], "two hour OGTT": 202 mg/dL, "glucose": 75 g,  
977 "carb prep days": null,  
978 "Type": "Incomplete Evidence.",  
979 "Expected Output": "Not diagnostic; Potentially invalid OGTT; ensure  $\geq$  150 g/day carbohydrate for 3 days  
prior, then retest.",  
980 "Guideline Evidence": "Adequate carbohydrate intake (at least 150 g/day) should be assured for 3 days prior  
981 to oral glucose tolerance testing..."

982  
983  
984  
985  
986

Figure 10: Pseudo-patient Examples for Incomplete Evidence.

987  
 988 **Pseudo-patient Examples for Threshold Misalignment.**  
 989  
 990 Example 1:  
 991 "Patient Input": "age": 60, "BMI": 30.1, "symptoms": [], "FPG": 125 mg/dL, "fasting": 9 hours,  
 992 "Type": "Threshold Misalignment",  
 993 "Expected Output": "Does not meet diabetes threshold; near threshold—repeat or perform OGTT; consider  
 994 retesting in 3–6 months if near margins.",  
 995 "Guideline Evidence": "FPG  $\geq$  126 mg/dL" and "If patients have test results near the margins... repeat the  
 996 test in 3–6 months."  
 997 Example 2:  
 998 "Patient Input": "age": 39, "BMI": 27.6, "symptoms": [], "A1C": 6.4 %,  
 999 "Type": "Threshold Misalignment",  
 1000 "Expected Output": "Below diagnostic threshold; not diabetes; repeat/alternative testing if near margins.",  
 1001 "Guideline Evidence": "A1C  $\geq$  6.5%..." and "If... near the margins... repeat... in 3–6 months."  
 1002 Example 3:  
 1003 "Patient Input": "age": 50, "BMI": 31.8, "symptoms": ["polyuria", "polydipsia"], "random glucose": 199  
 1004 mg/dL  
 1005 "Type": "Threshold Misalignment",  
 1006 "Expected Output": "Not diagnostic; random value is below 200 mg/dL even with symptoms; perform/confirm  
 1007 with FPG, 2-h OGTT, or A1C.", "Guideline Evidence": "In a patient with classic symptoms... random plasma  
 1008 glucose  $\geq$  200 mg/dL."  
 1009

Figure 11: Pseudo-patient Examples for Threshold Misalignment.

1010  
 1011 **Pseudo-patient Examples for Logic Error.**  
 1012 <Example 1>:  
 1013 "Patient Input": "age": 56, "BMI": 29.7, "symptoms": [], "random glucose": 205 mg/dL, "diagnosis declared  
 1014 before confirmation": true,  
 1015 "Type": "Sequential Logic Error",  
 1016 "Expected Output": "Do not diagnose yet; without classic symptoms, obtain confirmatory abnormal testing.",  
 1017 "Guideline Evidence": "In the absence of unequivocal hyperglycemia, diagnosis requires two abnormal test  
 1018 results..."  
 1019 <Example 2>:  
 1020 "Patient Input": "age": 64, "BMI": 27.9, "symptoms": [], "A1C %": 6.6, "FPG": 99 mg/dL, "repeat\_timing":  
 1021 "delayed months",  
 1022 "Type": "Sequential Logic Error",  
 1023 "Expected Output": "Repeat the above-threshold test without delay; do not postpone confirmation for months.",  
 1024 "Guideline Evidence": "It is recommended that the second test... be performed without delay."  
 1025 <Example 3>:  
 1026 "Patient Input": "age": 30, "BMI": 28.3, "symptoms": [], "pregnant": true, "A1C": 6.7%, "plasma\_tests": "not  
 1027 ordered",  
 1028 "Type": "Sequential Logic Error",  
 1029 "Expected Output": "Incorrect sequence; in pregnancy, use plasma glucose criteria (order glucose testing)  
 1030 rather than A1C.",  
 1031 "Guideline Evidence": "In conditions... pregnancy... only plasma blood glucose criteria should be used to  
 1032 diagnose diabetes."  
 1033

Figure 12: Pseudo-patient Examples for Logic Error.

1034  
1035 Table 5: Performance comparison of AutoClinician against baselines on the TriNetX dataset, reporting  
1036 diagnosis accuracy (Acc), mean reciprocal rank (MRR), and laboratory test accuracy (Lab Acc).

Model	TriNetX-Metabolic			TriNetX-Respiratory			TriNetX-Circulatory		
	Accuracy	MRR	Lab ACC	Accuracy	MRR	Lab ACC	Accuracy	MRR	Lab ACC
Vanilla Zero-shot Prompting ( <i>Baselines w/o Retrieval</i> )									
<b>General-purpose LLMs</b>									
Qwen3-32B	19.63	18.34	21.34	20.21	18.17	17.83	20.59	19.86	21.16
QwQ-32B	23.96	20.45	23.47	18.56	15.41	14.35	17.34	20.16	19.21
GPT-4o	26.28	21.56	20.01	24.83	21.52	23.73	25.84	23.96	24.82
<b>Medical-specialized LLMs</b>									
Chatdoctor	6.87	18.92	1.77	2.89	9.44	0.95	16.24	22.08	1.18
BioMistral	9.37	16.62	2.33	2.67	3.98	1.70	4.44	6.56	5.79
Yi-4B	14.06	13.28	11.21	18.89	14.67	16.79	12.22	21.90	11.29
DiagnosisGPT-34B	16.22	21.38	2.90	17.33	24.32	1.37	18.22	17.26	9.91
<b>Few-shot Demo. w/ Instruction</b>									
Qwen3-32B	20.57	16.82	19.75	23.45	17.23	18.44	21.59	18.06	20.36
QwQ-32B	21.57	18.91	13.52	19.89	16.06	14.81	22.51	23.31	18.20
<b>Enhancing Reasoning with iterative RAG</b>									
RAG-Qwen3-32B w/ Query Planning	22.85	23.81	22.10	24.31	23.15	22.08	22.47	26.40	24.73
<b>w/ AutoClinician</b>	26.50	28.50	27.80	29.00	28.20	26.33	28.70	31.80	27.40
RAG-QwQ-32B w/ Query Planning	25.37	26.59	18.64	23.25	25.45	24.97	23.36	25.28	25.33
<b>w/ AutoClinician</b>	31.50	31.20	26.37	27.10	26.41	29.50	32.79	33.02	30.28

### 1056 Pseudo-patient Examples for Contradictory Evidence.

1057 <Example 1>:

1058 "Patient Input": "age": 28, "BMI": 23.8, "symptoms": [], "A1C": 6.9 %, "FPG": 92 mg/dL, "sickle cell": true,  
1059 "Type": "Contradictory Evidence Injection",  
1060 "Expected Output": "Do not use A1C in sickle cell disease; diagnose only via plasma glucose criteria.",  
1061 "Guideline Evidence": "... hemoglobinopathies... only plasma blood glucose criteria should be used..."  
1062 and "Marked discordance... should raise the possibility of A1C assay interference."

1063 <Example 2>:

1064 "Patient Input": "age": 52, "BMI": 31.2, "symptoms": [], "random glucose": 210 mg/dL, "A1C": 6.4%,  
1065 "Type": "Contradictory Evidence Injection",  
1066 "Expected Output": "Not diagnostic without classic symptoms; proceed with FPG/OGTT/A1C confirmation.",  
1067 "Guideline Evidence": "In a patient with classic symptoms... random plasma glucose  $\geq$  200 mg/dL."

1068 <Example 3>:

1069 "Patient Input": "age": 70, "BMI": 24.6, "symptoms": [], "hemodialysis": true, "A1C": 6.8%, "FPG": 102  
mg/dL,  
1070 "Type": "Contradictory Evidence Injection",  
1071 "Expected Output": "Do not diagnose by A1C in hemodialysis; require plasma glucose criteria.",  
1072 "Guideline Evidence": "... hemodialysis... only plasma blood glucose criteria should be used...""

1073  
1074 Figure 13: Pseudo-patient Examples for Contradictory Evidence.  
1075  
1076

1077  
1078 We collaborated with clinicians to review metabolic-related CEGs and selected one diabetes-related case  
1079 as a representative example. Human experts then evaluated the CEGs in Table 14 along three dimensions:  
1080 guideline faithfulness, Logic coherence, and clinical utility.

1081  
 1082 **One example of Generated Diabete-related CEGs**  
 1083  
 1084  
 1085 {  
 1086 "Symptom": {  
 1087 "classic symptoms of hyperglycemia",  
 1088 "hyperglycemic crisis",  
 1089 "pregnancy (second and third trimesters and the postpartum period)",  
 1090 "hemoglobinopathies including sickle cell disease",  
 1091 "glucose-6-phosphate dehydrogenase deficiency",  
 1092 "HIV",  
 1093 "hemodialysis",  
 1094 "recent blood loss or transfusion",  
 1095 "erythropoietin therapy",  
 1096 "symptomatic patients" },  
 1097  
 1098 "Graphs": {  
 1099 "Nodes": {  
 1100 {  
 1101 "ID": "S1",  
 1102 "Context": "Diagnostic Tests for Diabetes",  
 1103 "Trigger": "Diabetes may be diagnosed based on plasma glucose criteria, either the fasting plasma glucose (FPG) value or the 2-h plasma glucose (2-h PG) value during a 75-g oral glucose tolerance test (OGTT) or A1C criteria",  
 1104 "State": "Criteria for the diagnosis of diabetes",  
 1105 "Terminal": false  
 1106 },  
 1107 {  
 1108 "ID": "S2",  
 1109 "Context": "Criteria for the diagnosis of diabetes",  
 1110 "Trigger": "FPG 126 mg/dL (7.0 mmol/L). Fasting is defined as no caloric intake for at least 8 h.",  
 1111 "State": "Confirming the Diagnosis",  
 1112 "Terminal": false  
 1113 },  
 1114 ...  
 1115 },  
 1116 "ID": "S14",  
 1117 "Context": "Multiple confirmed islet autoantibodies is a risk factor for clinical diabetes",  
 1118 "Trigger": "When multiple islet autoantibodies are identified, referral to a specialized center for further evaluation and/or consideration of a clinical trial or approved therapy to potentially delay development of clinical diabetes should be considered.",  
 1119 "State": "High risk of type 1 diabetes (pre-clinical stage)",  
 1120 "Terminal": true  
 1121 }  
 1122 } } }

1123 Figure 14: This result presents a JSON output of the generated CEGs for diabetes diagnosis, including a list  
 1124 of relevant symptoms used for query embedding alongside a sequence of numbered logic nodes (S1–S14).  
 1125 Each node specifies its clinical context, trigger condition, diagnostic state, and termination flag, outlining the  
 1126 structured reasoning path from symptoms to confirmed diagnosis..  
 1127

**Faithfulness:** Accurately encodes ADA diagnostic criteria, including FPG ( $\geq 126$  mg/dL), OGTT 2h-PG ( $\geq 200$  mg/dL), A1C ( $\geq 6.5\%$ ), and random plasma glucose ( $\geq 200$  mg/dL) with symptoms or hyperglycemic crisis. Confirmatory testing logic (Nodes 6–9) is preserved. Minor redundancy exists in OGTT-related triggers (Nodes 3 and 12).

**Logic Coherence:** Node transitions are logically consistent and lead to appropriate terminal outcomes. Terminal states include both diagnostic conclusions (Nodes 5, 7, 9) and clinical actions (Node 14). Pre-test carbohydrate intake branches (Nodes 10–12) introduce unnecessary complexity.

**Clinical Utility:** Covers all major diagnostic pathways, including glucose thresholds, symptom-based criteria, and autoantibody testing for type 1 diabetes. Handles indeterminate or discordant results through non-terminal states (e.g., `ConflictResults`, `Unresolved`). High-risk clinical conditions (e.g., HIV, dialysis, transfusion) appear in the symptom set but are not explicitly integrated into decision pathways.

**Assessment:** The CEG exhibits high fidelity to guidelines and strong logic coherence. Recommended refinements include reducing redundancy, simplifying OGTT-related pathways, and integrating high-risk conditions into decision logic.

#### C.4.3 HUMAN EVALUATION ON DIALOGUES AND CASE STUDY

Turn	Dialogue (EHR + DFA Reasoning)
$t = 1$	<p><b>Patient:</b> I'm a woman, not Hispanic. Lately I've been urinating a lot, feeling very thirsty, having blurry vision, and losing weight without trying</p> <p><b>AutoClinician:</b> Based on your reported symptoms, the current state is EvidenceGathering, and the next step is to conduct diagnostic tests such as FPG, OGTT, or A1C to confirm the diagnosis of diabetes..</p>
$t = 2$	<p><b>Patient:</b> Here are my Lab results: Fasting plasma glucose=104.0 mg/dL; Glucose [Mass/volume] in Serum or Plasma (random)=180.0 mg/dL; Hemoglobin A1c/Hemoglobin.total in Blood=6.7 %;</p> <p><b>AutoClinician:</b> Based on the lab results, the patient has conflicting indicators for diabetes, with a fasting plasma glucose within normal range but elevated random glucose and hemoglobin A1c. Further testing such as a repeat A1C or an OGTT is needed to resolve the diagnostic uncertainty.</p>
$t = 3$	<p><b>Patient:</b> My A1C is 5.4%.</p> <p><b>AutoClinician:</b> Your A1C is below threshold, which conflicts with your FPG. Let's repeat the A1c or OGTT.</p>
$t = 4$	<p><b>Patient:</b> My OGTT fasting is 118 mg/dL.</p> <p><b>AutoClinician:</b> Repeat testing confirms abnormal OGTT. I have collect enough evidence and confirm you the E11 Type 2 diabetes mellitus</p>

Table 6: An EHR-driven Dialogue based on DFA.

We collaborated with clinicians to review approximately 30 generated dialogues and selected one diabetes-related case as a representative example. Human experts then evaluated the dialogue in Table 6 along three dimensions: guideline faithfulness, logic coherence, and clinical utility.

**Faithfulness to Guidelines:** Autoclinician correctly mapped patient-reported classic symptoms (polyuria, polydipsia, blurry vision, unintentional weight loss) to the EvidenceGathering state and recommended guideline-concordant diagnostic tests (FPG, OGTT, A1C). It appropriately interpreted numerical values: for example, FPG of 104 mg/dL was recognized as within the non-diabetic range ( $< 126$  mg/dL), random glucose of 180 mg/dL as elevated but below the diagnostic threshold (200 mg/dL), and A1C of 6.7% as above the diagnostic cutoff (6.5%). The model advised repeat testing when discordant results were observed, consistent with ADA recommendations. A minor inaccuracy occurred at Turn 3, where Autoclinician attributed the

1175 conflict to FPG instead of the discrepancy between the initial elevated A1C (6.7%) and the subsequent normal  
 1176 A1C (5.4%).  
 1177

1178 **Logic Coherence:** Dialogue progression followed the deterministic DFA policy. Each transition was grounded  
 1179 in the observed inputs, and the terminal state ConfirmedDiagnosis (Type 2 diabetes) was reached only after  
 1180 confirmatory evidence had been obtained.

1181 **Clinical Utility:** Autoclinician provided clear next steps at each turn, such as recommending repeat A1C  
 1182 or OGTT to resolve uncertainty, reflecting good alignment with guideline-directed clinical workflows. By  
 1183 explicitly interpreting laboratory values relative to diagnostic thresholds, the dialogue enhanced transparency  
 1184 and patient understanding. The final diagnosis of type 2 diabetes was appropriate. The explicit statement of  
 1185 diagnostic thresholds (e.g., OGTT is abnormal) further highlighted the model’s interpretability and clinical  
 1186 credibility by clarifying normal and abnormal ranges

1187 **Overall Assessment.** Autoclinician demonstrated high fidelity to guidelines, accurate interpretation of  
 1188 numerical test results, and appropriate use of guideline-based recommendations for further laboratory testing.  
 1189

### 1190 C.5 SENSITIVITY ANALYSIS ON DIFFERENT JUDGE THRESHOLD

1192 Table 7: Performance on Varying the LLM-as-a-judge Alignment Thresholds  
 1193

1194 <b>Dataset</b>	0.5	0.6	0.7	0.8	0.9
1195 EHRSHOT-Metabolic	25.33	26.35	26.12	26.54	26.32
1196 EHRSHOT-Respiratory	27.95	29.10	30.42	29.55	30.47
1198 EHRSHOT-Circulatory	27.06	28.54	29.37	30.68	30.51

1200 Table 7 Results report performance when varying the LLM-judge alignment threshold from 0.5 to 0.9. Across  
 1201 all disease settings, accuracy improves notably when increasing the threshold from 0.5 to 0.7, indicating  
 1202 that pseudo-patient examples effectively correct CEG errors and strengthen reasoning quality. Beyond 0.7,  
 1203 performance stabilizes and changes become marginal, demonstrating robustness to threshold choice. In our  
 1204 experiment, we adopt a slightly conservative threshold of 0.8.

### 1206 C.6 ABLATION STUDIES ON DIFFERENT PSEUDO-PATIENT TYPES.

1208 Table 8: Ablation Study on Removing Different Refinement Types.  
 1209

1210 <b>Dataset</b>	1211 <b>w/o Incomplete Evidence</b>	1212 <b>w/o Threshold Misalignment</b>	1213 <b>w/o Logic Errors</b>	1214 <b>w/o Conflict Resolution Failure</b>
1213 EHRSHOT-Metabolic	24.31	25.67	26.05	26.13
1214 EHRSHOT-Respiratory	27.39	28.35	28.79	27.67
1215 EHRSHOT-Circulatory	28.76	29.58	28.17	29.34

1217 Table 8 presents ablations isolating each refinement component. Removing Incomplete Evidence correction  
 1218 leads to the largest degradation, reflecting the importance of capturing missing diagnostic signals. Threshold  
 1219 Misalignment removal also produces a measurable drop, especially in circulatory disorders where diagnostic  
 1220 thresholds are critical. Removing Logic Ordering or Conflict Resolution yields smaller but consistent declines,  
 1221 suggesting their contribution mainly affects ambiguous cases.

1222 C.7 HUMAN INTER-RATER RELIABILITY  
12231224 Table 9: Clinician Evaluation Cohen’s Scores.  
1225

1226 Dimension	1227 Mean Score (Clinician A)	1228 Mean Score (Clinician B)	1229 Cohen’s $\kappa$
1228 Guideline Faithfulness	1229 4.6	1230 4.8	0.74
1229 Logical Coherence	1230 4.4	4.5	0.69
1230 Clinical Utility	3.3	3.1	0.65

1231 We incorporated quantitative blinded human evaluation results. Two licensed clinicians independently  
1232 reviewed 30 randomly sampled model-generated diagnostic dialogues. Each dialogue was rated using  
1233 a 5-point Likert scale across three dimensions: guideline faithfulness, reasoning coherence, and clinical  
1234 utility, where 1 indicates clinically incorrect reasoning and 5 represents excellent guideline concordance and  
1235 interpretability. Experimental results are shown in Table 9.

1236 To assess the consistency of the evaluation, we computed inter-rater agreement using Cohen’s Cohen (1968).  
1237 The resulting values indicate moderate to substantial agreement, demonstrating that the evaluation is reliable  
1238 and not dependent on a single reviewer’s subjective judgement. Importantly, clinicians noted that when  
1239 evidence was insufficient, our model appropriately declined to make a diagnosis and switch to the base model  
1240 rather than hallucinating or generating unsafe recommendations.

1241 C.8 SENSITIVITY TO DIFFERENT JUDGE MODELS.  
12421243 Table 10: Performance Comparison on Different Judge Models.  
1244

1245 Model	1246 EHRSHOT-Metabolic			1247 EHRSHOT-Respiratory			1248 EHRSHOT-Circulatory		
	1249 Acc	1250 MRR	1251 Lab Acc	1252 Acc	1253 MRR	1254 Lab Acc	1255 Acc	1256 MRR	1257 Lab Acc
1258 QwQ-32B	25.45	26.18	22.48	28.62	30.05	25.16	29.41	28.95	23.87
1259 DiagnosisGPT-34B	26.10	26.75	23.01	29.20	30.66	26.81	30.21	29.03	24.65
1260 Qwen-32B	26.54	26.92	23.37	29.55	31.09	27.32	30.68	29.14	24.98

1261 To assess whether the judge model affects refinement outcomes, we evaluated AutoClinician with multiple  
1262 LLM judges (Qwen-32B (we used in the paper), QwQ-32B, DiagnosisGPT-34B). As shown in Table 10,  
1263 results remain highly consistent across judge model choices.

1264 C.9 ANALYSIS ON FALBACK RATE  
12651266 Table 11: Fallback Rate across Different Disease Subtypes.  
1267

1268 Category	1269 Subtypes & Fallback Rates			
1270 <b>Metabolic</b>	Thyroid (34%)	1271 Diabetes (21%)	1272 Metabolism (46%)	1273 Endocrine (24%)
1272 <b>Circulatory</b>	Hypertension (17%)	1273 Coronary (26%)	1274 Cerebrovascular (32%)	1275 Aortic (43%)
1274 <b>Respiratory</b>	Lower Respiratory Infection (12%)	1275 Chronic Airway (19%)	1276 Lung Injury (31%)	

1277 To quantify how often fallback occurs, we measured fallback frequency across disease subdomains, with  
1278 results shown in Table 11. The fallback rate ranges from 12% to 46%, and this variability is clinically  
1279 interpretable. The primary contributing factor is missingness in real EHR data. In real-world clinical settings,

1269 patient observations (e.g., laboratory values) are often absent, either due to record errors or because tests  
 1270 were not ordered for clinical or cost-based reasons. In this work, we preserved data missingness rather  
 1271 than filtering incomplete samples, ensuring that evaluation reflects realistic deployment conditions. Our  
 1272 results demonstrate that guideline-grounded reasoning can be automated and applied to real EHRs. However,  
 1273 handling incomplete evidence and enabling trustworthy clinical reasoning under uncertainty remains an open  
 1274 and important direction for future research.

1275

### 1276 C.10 RESULTS OF DIFFERENT OPEN-SOURCE BASE MODELS FOR CEG GENERATION AND 1277 REFINEMENT

1279 Table 12: Results of Different Open-source Base model for CEG generation and Refinement.

1281 <b>Model</b>	1282 <b>EHRSHOT-Metabolic</b>			1283 <b>EHRSHOT-Respiratory</b>			1284 <b>EHRSHOT-Circulatory</b>		
	1285 Acc	MRR	LabAcc	Acc	MRR	LabAcc	Acc	MRR	LabAcc
1286 Yi-34B	25.60	26.75	23.19	28.71	31.31	26.53	29.75	29.25	24.28
1287 GPT-OSS-120B	27.11	27.25	24.56	30.22	32.87	28.11	30.50	30.47	25.50
1288 Qwen-32B	26.54	26.92	23.37	29.55	31.09	27.32	30.68	29.14	24.98

1289 To assess scalability and accessibility, we further tested refinement using Yi-34B and GPT-OSS-120B. As  
 1290 shown in Table 12, the resulting performance remains consistent across base models. We observed that  
 1291 GPT-OSS-120B tends to generate longer responses and higher inference latency. In contrast, Qwen-32B and  
 1292 Yi-34B produce cleaner structured reasoning and more accurate extraction of guideline logic, making them  
 1293 well-suited for CEG refinement and reasoning.

1294

### 1295 C.11 STATISTICAL SIGNIFICANCE

1296 Table 13: Statistically Significance.

1297 <b>Dataset</b>	1298 <b>Mean Accuracy (Baseline)</b>	1299 <b>Mean Accuracy (AutoClinician)</b>	1300 <b>Mean <math>\Delta</math></b>	1301 <b>p-value</b>
1302 EHRSHOT-Metabolic	1303 $22.88 \pm 0.12$	1304 $26.50 \pm 0.07$	1305 $+3.62$	1306 0.0022
1307 EHRSHOT-Respiratory	1308 $24.31 \pm 0.11$	1309 $29.00 \pm 0.13$	1310 $+4.75$	1311 0.0009
1312 EHRSHOT-Circulatory	1313 $22.47 \pm 0.14$	1314 $28.70 \pm 0.17$	1315 $+6.21$	0.0004

1316 To evaluate whether the improvements introduced by AutoClinician are statistically significant rather than  
 1317 random variation, we conducted a paired t-test comparing our method against the strongest baseline (RAG-  
 1318 Qwen-32B w/ Query Planning) across the 5 runs. These results indicate that the improvement is highly  
 1319 statistically significant. The summary is Table 13.

1320 AutoClinician consistently improves accuracy across all three datasets, with p-values  $< 0.01$ , indicating the  
 1321 gains are statistically significant rather than random variation. This confirms that AutoClinician’s gains are  
 1322 robust, reliable, and not attributable to random noise.