CEDAR: Constraint-based Evidence-driven Diagnostic Evaluation and Reasoning

Anonymous ACL submission

Abstract

Clinical diagnostic evaluation requires determining whether the clinical evidence documented in electronic medical records (EMRs) sufficiently supports diagnoses under guideline-based constraints.In practice, this process involves not only verifying diagnostic validity, but also assessing whether the required diagnostic evidence has been adequately documented. However, existing approaches mainly rely on end-to-end large language model (LLM) reasoning, making them sensitive to incomplete evidence and complex guideline constraints.To address this issue, we propose CEDAR(**C**onstraint-based **E**vidence-driven **D**iagnostic **A**uditing and **R**easoning), a structured diagnostic evaluation framework grounded in clinical guideline constraints.CEDAR explicitly separates evidential sufficiency assessment from diagnostic validity evaluation, and combines evidence–guideline relation graphs with deterministic constraint execution to perform interpretable diagnostic evaluation and critical missing-evidence attribution.We evaluate CEDAR on 3,079 real-world electronic medical records spanning 18 diseases. Experimental results show that CEDAR consistently outperforms existing baselines in evidence sufficiency assessment, module-level auditing, and missing-evidence identification.Code and datasets are available at `https://anonymous.4open.science/r/CEDAR-22A1`

1 Introduction

Clinical diagnostic evaluation is a critical task in medical quality control (Singhal et al., 2023; Rao et al., 2026). In real-world clinical practice, reviewers assess whether diagnoses recorded in electronic medical records (EMRs) are supported by guideline-required clinical evidence, including symptoms, medical history, laboratory tests, and imaging findings. Unlike conventional diagnostic reasoning, clinical diagnostic evaluation requires not only determining whether a diagnosis is clinically valid, but also assessing whether the corresponding EMR sufficiently documents the required diagnostic evidence. Diagnostic validity and evidential sufficiency are therefore not equivalent. A patient may satisfy the diagnostic criteria for a disease, while the corresponding EMR still fails to document the guideline-required evidence necessary for clinical auditing. Figure 1 illustrates this distinction.

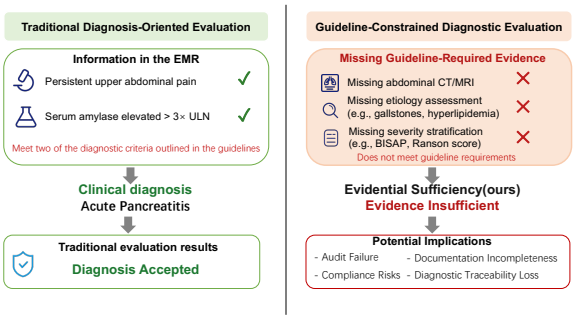

Figure 1: Diagnostic validity does not necessarily imply evidential sufficiency. CEDAR extends traditional diagnosis-oriented evaluation by further assessing whether guideline-required evidence has been sufficiently documented in the EMR.

As the volume of EMRs continues to grow, manual diagnostic review becomes increasingly difficult to scale. Recent advances in large language models (LLMs) have demonstrated strong capability in medical reasoning, medical question answering, and clinical decision support (Singhal et al., 2023; Thirunavukarasu et al., 2023; Long et al., 2026). Existing LLM-based clinical reasoning approaches mainly focus on diagnosis prediction or medical question answering, where the primary objective is to determine whether a diagnosis can be inferred from the available evidence. However, clinical diagnostic auditing requires a fundamentally different evaluation objective. In practical medical quality control, reviewers must additionally determine whether the EMR sufficiently documents the guideline-required evidence supporting

the diagnosis.

This process introduces several unique challenges. Clinical guidelines often contain complex logical constraints, including conjunction conditions, exclusion criteria, optional evidence, and threshold-based rules. Meanwhile, clinical evidence is distributed across heterogeneous EMR modules with substantial semantic variation, making reliable evidence–guideline alignment difficult. Moreover, missing evidence identification requires determining whether absent evidence would alter the final sufficiency outcome under guideline constraints. Under such settings, end-to-end LLM reasoning often produces unstable and weakly interpretable evaluation results.

To address these challenges, we propose CEDAR (**C**onstraint-based **E**vidence-driven **D**iagnostic **E**valuation and **R**easoning), a structured clinical diagnostic evaluation framework grounded in guideline constraints. CEDAR combines semantic evidence–guideline alignment with deterministic constraint execution to jointly evaluate evidence sufficiency (Bai et al., 2025), diagnostic validity, and critical missing evidence. Specifically, LLMs are used for semantic relation induction, while deterministic execution is used for logical constraint reasoning over the aligned evidence.

To evaluate CEDAR, we construct a clinical diagnostic evaluation benchmark consisting of 3,079 real-world EMRs across 18 diseases and multiple diagnostic reasoning types. Experimental results demonstrate that CEDAR achieves more stable and interpretable clinical diagnostic evaluation than direct end-to-end generation under complex guideline constraints (Ji et al., 2023; Wang et al., 2024).

In summary, the main contributions of this paper are as follows:

- We propose CEDAR(**C**onstraint-based **E**vidence-driven **D**iagnostic **E**valuation and **R**easoning), a framework that formulates clinical diagnostic evaluation as a guideline-constrained evidence assessment problem, and achieves stable and interpretable structured evaluation through a hybrid reasoning mechanism combining LLMs with deterministic execution.
- We construct a structured guideline knowledge base for clinical diagnostic evaluation, converting unstructured clinical guidelines into an executable representation that incorporates logical constraint relationships, thereby supporting subsequent structured reasoning.
- Systematic experiments are conducted on a clinical diagnostic evaluation benchmark covering multiple diagnostic reasoning patterns, and the results show that CEDAR achieves better evaluation performance and explainability in scenarios involving complex logical constraints.

2 Related Work

2.1 Clinical Diagnostic Reasoning

Clinical diagnostic reasoning has been widely studied in medical natural language processing (Tu et al., 2024; Rao et al., 2026), particularly in disease prediction, medical question answering, and differential diagnosis (Zhu et al., 2025). Recent advances in large language models (LLMs) have further improved the ability of models to interpret clinical narratives and perform complex medical reasoning. Existing studies mainly evaluate whether a model can generate the correct diagnosis given the available clinical information (McCoy et al., 2025). However, in real-world clinical quality control, diagnostic evaluation involves not only determining diagnostic correctness, but also assessing whether the medical record sufficiently documents the evidence required by clinical guidelines. As a result, methods designed primarily for diagnosis prediction are insufficient for structured diagnostic evaluation under incomplete evidence conditions.

2.2 Guideline-grounded Clinical Evaluation

To improve the reliability and interpretability of medical reasoning, prior work has incorporated clinical guidelines, rule-based systems, symbolic reasoning (Pan et al., 2024; Wu et al., 2025), and structured medical knowledge into clinical decision-making tasks (Kresevic et al., 2024). Existing approaches have explored knowledge graphs, clinical pathway modeling, and executable medical rules for structured clinical analysis (Rezaei et al., 2025; Zhu et al., 2026). These methods improve the representation of medical knowledge and partially enhance reasoning interpretability. Nevertheless, most existing studies focus on disease reasoning or knowledge representation itself, rather than explicitly evaluating evidential sufficiency under guideline constraints. In particular, complex diagnostic rules involving conjunction, exclusion conditions, optional evidence, and score-based accumulation remain difficult to model consistently in practical clinical auditing scenarios (Long et al., 2026; Wang

et al., 2026).

2.3 LLM-based Medical Evaluation and Reasoning

Recent studies have increasingly explored LLM-based medical evaluation frameworks (Liu et al., 2024; Thirunavukarasu et al., 2023; Singhal et al., 2025), including retrieval-augmented generation (RAG), multi-stage reasoning, and agent-based clinical systems (Lewis et al., 2020; Mialon et al., 2023; Yao et al., 2022). By introducing external medical knowledge and guideline information (Chen et al., 2025; Kadusabe et al., 2025; Wong and Wong, 2026), these methods improve the performance of LLMs on complex clinical tasks. However, most existing approaches still rely primarily on free-form semantic reasoning, where logical constraints are implicitly handled within the generation process. Such approaches often lack explicit constraint modeling and deterministic execution mechanisms (Wei et al., 2022; Wang et al., 2022; Huang et al., 2025), making them vulnerable to unstable reasoning when evidence is incomplete or guideline structures become complex. In contrast, our work formulates clinical diagnostic evaluation as a guideline-constrained evidence assessment problem and combines semantic alignment with deterministic constraint execution to achieve more stable and interpretable diagnostic evaluation.

3 Method

We propose CEDAR (**C**onstraint-based **E**vidence-driven **D**iagnostic **E**valuation and **R**easoning), a structured diagnostic evaluation framework grounded in clinical guideline constraints. Given an electronic medical record (EMR) and its corresponding clinical guideline, CEDAR jointly evaluates diagnostic validity, evidential sufficiency, and critical missing evidence through guideline-constrained evidence assessment.

As illustrated in Figure 2, CEDAR consists of five stages. First, structured guideline knowledge construction transforms clinical guidelines into executable constraint representations (Li et al., 2025). Second, clinical evidence normalization converts heterogeneous EMR contents into unified evidence units. Third, evidence–guideline relation graph construction employs large language models to infer semantic relations between clinical evidence and guideline items. Fourth, deterministic constraint execution performs rule-based reasoning over the relation graph to evaluate diagnostic validity and evidential sufficiency under predefined logical constraints. Finally, counterfactual missing evidence attribution identifies critical missing evidence through intervention-based re-execution.

3.1 Structured Guideline Constraint Modeling

Clinical guidelines are first transformed into structured executable constraint representations. Specifically, hierarchical parsing and text normalization are performed to recover the structural organization of diagnostic requirements from the original guideline documents (Kadi et al., 2025). To reduce long-context interference during guideline understanding, retrieval-augmented parsing is further adopted to retrieve guideline content associated with the target disease and diagnostic modules (Asai et al., 2024).

Based on the retrieved guideline context, diagnostic items, examination requirements, and their logical relationships are extracted through multi-stage constraint analysis. The extracted guideline items are subsequently categorized into structured clinical modules, including symptoms, laboratory examinations, imaging examinations, auxiliary examinations, and exclusion conditions. Formally, a clinical guideline is represented as:

$$G = \{(M_i, C_i, L_i)\}_{i=1}^{N} \tag{1}$$

where M_i denotes the clinical module, C_i denotes the corresponding guideline items, and L_i denotes the logical constraint type.

To support executable diagnostic evaluation, multiple constraint structures are further defined, including `AND`, `OR`, `COREAND+OPTIONAL`, `COREOR+OPTIONAL`, `AT_LEAST_N`, and `OPTIONAL`. These structures provide unified executable semantics for subsequent evidence reasoning. Detailed definitions and structured examples are provided in Appendix A.

3.2 Clinical Evidence Normalization

Clinical evidence in EMRs is distributed across heterogeneous clinical modules and expressed with diverse documentation styles (Yang et al., 2022; Chung et al., 2025), making it difficult to directly perform structured constraint execution. To obtain a unified representation for reasoning, the EMR contents are normalized into standardized evidence units. Formally, given an electronic medical record

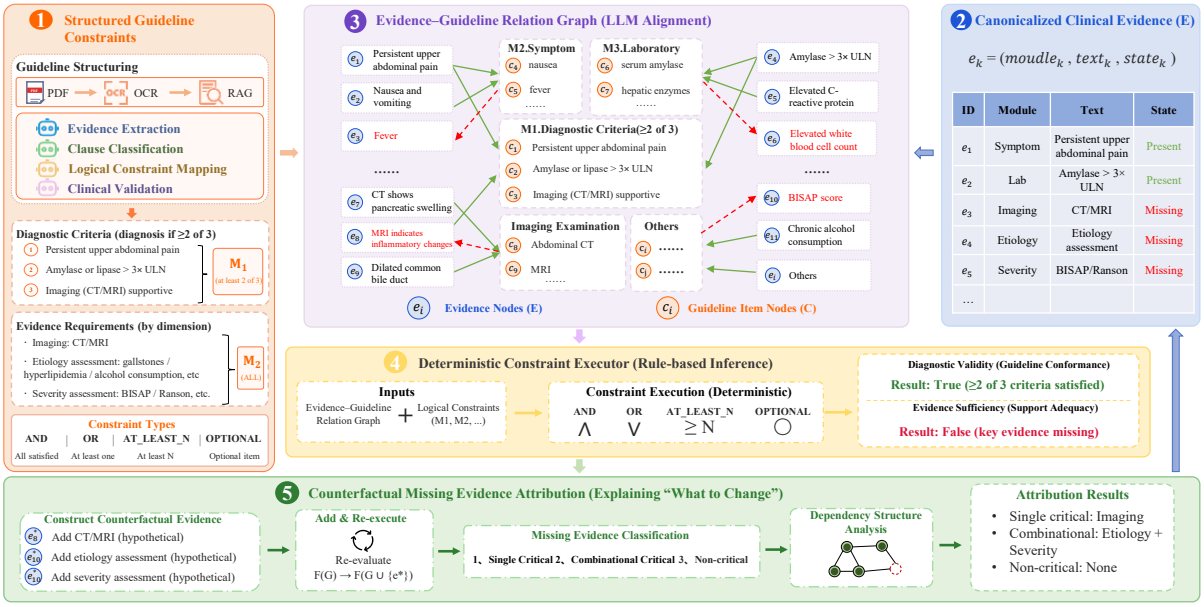

Figure 2: Overview of CEDAR using acute pancreatitis as an example. The framework first converts clinical guidelines into executable structured constraints, then constructs evidence–guideline relation graphs from normalized clinical evidence in EMRs. Based on deterministic constraint execution, CEDAR performs structured evaluation of diagnostic validity and evidential sufficiency, while further identifying critical missing evidence through counterfactual intervention.

E, the normalized evidence set is represented as:

$$E = \{e_1, e_2, \ldots, e_k\} \tag{2}$$

where each evidence item preserves its source module, normalized evidence text, and corresponding clinical state. Each evidence item is further categorized into one of three states: *positive*, *negative/normal*, or *uncertain*. Unlike conventional diagnostic reasoning settings that mainly focus on positive findings, negative or normal findings are retained as valid clinical evidence rather than treated as missing information. This distinction is important for exclusion-based diagnostic conditions and normality-dependent clinical evaluations.

3.3 Evidence–Guideline Relation Graph Construction

Clinical evidence and guideline items often exhibit substantial semantic variation across different documentation styles and clinical expressions, making direct evidence matching unreliable. To model fine-grained semantic alignment between EMR evidence and guideline constraints, an evidence–guideline relation graph is constructed (Edge et al., 2024). Let C denote the set of all guideline items. For each evidence–guideline pair $(e, c) \in E \times C$, a large language model is employed to infer their semantic relation. The relation set is defined as:

$$R = \{\texttt{Support}, \texttt{Performed}, \texttt{Contradiction}\} \tag{3}$$

where `Support` denotes that the evidence satisfies the corresponding guideline item, `Performed` denotes that the examination has been completed with normal or negative findings, and `Contradiction` denotes semantic conflict between the evidence and the guideline item. Based on the inferred relations, a bipartite relation graph is constructed as:

$$\mathcal{G} = (E, C, R) \tag{4}$$

where each edge (e, c, r) represents the semantic relation r between evidence item e and guideline item c. Guideline items without any `Support` or `Performed` relation are treated as unsatisfied during constraint execution.Detailed relation definitions and examples are provided in Appendix B.

To reduce invalid semantic matching, relation inference is performed primarily within the same clinical module, while cross-module alignment is allowed only when no valid intra-module relation can be established.

3.4 Deterministic Constraint Execution

Final diagnostic evaluation relies on deterministic constraint execution over the evidence–guideline

relation graph (Zhang et al., 2026). Large language models are used only for semantic relation inference, while all evaluation outcomes are determined exclusively by predefined logical constraints and graph relations.

For each guideline item, the corresponding state is determined according to its associated relations. Items with `Support` or `Performed` relations are treated as satisfied, items containing only `Contradiction` relations are treated as contradicted, and all remaining items are treated as unsupported. Detailed execution rules are provided in Appendix C.

Evidence sufficiency and diagnostic validity are evaluated separately during constraint execution. Evidence sufficiency is formulated as a module-level constraint satisfaction problem, while diagnostic validity is evaluated using an independent set of diagnostic criteria.

3.5 Counterfactual Missing Evidence Attribution

Critical missing evidence is identified through counterfactual intervention over the evidence–guideline relation graph. For any unsatisfied guideline item c, a corresponding counterfactual supporting evidence item e^* is introduced into the relation graph through a `Support` relation, followed by re-execution of deterministic constraint reasoning:

$$F(\mathcal{G}) \rightarrow F(\mathcal{G} \cup \{e^*\}) \tag{5}$$

where $F(\cdot)$ denotes the evidence sufficiency prediction obtained through constraint execution.

Missing evidence is categorized according to the intervention outcome. *Single critical missing evidence* changes the sufficiency result independently, *combinational critical missing evidence* changes the result only under joint intervention with other missing items, and *non-critical missing evidence* does not affect the final sufficiency decision. Unlike free-form explanation generation, counterfactual intervention produces executable and verifiable missing evidence attribution directly grounded in the constraint execution process.

4 Dataset Construction

4.1 Data Source

The dataset consists of 3,079 real-world electronic medical records spanning 18 disease categories and diverse diagnostic constraint structures. The medical records are collected from the publicly available MedChain benchmark (Liu et al., 2026). All records were de-identified prior to annotation to protect patient privacy (Sousa et al., 2025).re de-identified prior to annotation to protect patient privacy.

Category	Counts
Multi-Evidence Conjunctive Type	921
Staging-Dependent Chronic Disease Type	566
Exclusion Diagnosis Type	99
Imaging-Light Dependent Type	1092
Score-Based Accumulation Type	206
Symptomology-Based Diagnostic Type	195

Table 1: Sample distribution of six types of diagnostic reasoning

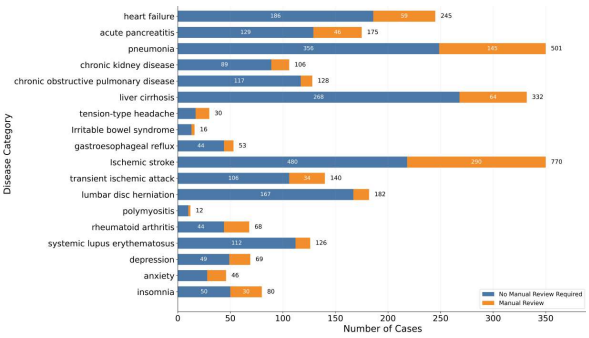

Figure 3: Detailed statistics of the clinical evaluation dataset across different disease categories, including sample size,and expert review distribution.

4.2 Diagnostic Taxonomy

The disease categories are organized according to their underlying diagnostic constraint structures in order to evaluate clinical diagnostic evaluation under diverse reasoning settings. The resulting taxonomy covers six representative diagnostic reasoning types, including combinational constraints, exclusion-based diagnosis, temporal-dependent diagnosis, and threshold-based accumulation.Table 1 summarizes the distribution across different diagnostic taxonomy categories. Representative diseases and detailed diagnostic taxonomy definitions are provided in Appendix D.

4.3 Annotation Pipeline

Annotations are obtained through a dual-model annotation pipeline with clinical expert adjudication (Liu et al., 2026). Qwen-3.5-Plus and Gemini-2.5-Pro independently produce structured diagnostic evaluations, and inconsistent cases are resolved by three clinical experts. The two mod-

els achieve complete agreement on 73.6% of the samples.Figure 3 further shows the distribution of model agreement and expert adjudication across different disease categories.

5 Experiments and Results

5.1 Experimental Setup

Experiments are conducted on a clinical diagnostic evaluation benchmark consisting of 3,079 real-world electronic medical records spanning 18 disease categories. Each case is annotated according to its corresponding clinical guideline, including evidence sufficiency, diagnostic validity, module-level support status, and missing evidence labels. To evaluate clinical diagnostic evaluation under diverse constraint structures, the dataset further covers six representative diagnostic taxonomy categories.

We evaluate the framework on three tasks: evidence sufficiency assessment, diagnostic validity assessment, and missing evidence identification. Accuracy, Precision, Recall, and F1-score are adopted as evaluation metrics.

CEDAR is compared against Direct-LLM, CoT-based, and RAG-based (Lewis et al., 2020) baselines. Qwen3.5-Plus is used as the default relation induction model with the temperature set to 0.01. For retrieval, structured clinical guidelines are used as the document corpus. The guidelines are segmented into semantically coherent chunks with an average length of approximately 512 tokens. Dense retrieval is implemented using text2vec-base-chinese together with FAISS, while a Cross-Encoder reranker is further introduced to improve retrieval accuracy under long-guideline settings.

5.2 Main Results

Table 2 reports the performance of CEDAR and baseline methods on evidence sufficiency assessment (**Sufficiency**), diagnostic validity assessment (**Validity**), and module-level evaluation (**Module-level Evaluation**).

CEDAR consistently achieves the strongest overall performance across different foundation models, with the largest improvements observed on evidence sufficiency assessment and module-level evaluation. These results indicate that explicit constraint execution is more effective than direct end-to-end generation under complex guideline constraints.

The improvement on evidence sufficiency assessment is particularly notable. Existing LLM-based methods mainly focus on diagnostic validity, but often fail to determine whether diagnostic evidence has been comprehensively documented. By explicitly modeling evidence–guideline constraints, CEDAR achieves substantially better sufficiency evaluation performance across most model settings.

CEDAR also achieves superior Mod.Acc and Mod.Macro-F1 in module-level evaluation, demonstrating the effectiveness of relation-graph-based constraint execution for modeling structured dependencies across clinical modules.

5.3 Analysis Across Diagnostic Reasoning Types

Table 3 reports the performance of CEDAR across different diagnostic reasoning types. The results show clear variation across reasoning categories, indicating that clinical diagnostic evaluation difficulty is strongly related to the underlying diagnostic constraint structure.

CEDAR achieves the strongest performance on the *Multi-Evidence Conjunctive Type*, where diagnostic requirements are typically defined through explicit conjunction constraints across multiple clinical modules. Strong performance is also observed on the *Exclusion Diagnosis Type* and *Score-Based Accumulation Type*, suggesting that explicit constraint execution effectively handles exclusion conditions, negative evidence, and threshold-based diagnostic criteria.

Performance decreases on the *Symptomology-Based Diagnostic Type*, which relies more heavily on subjective symptom descriptions and implicit clinical judgment, making evidence alignment and sufficiency evaluation more challenging.

5.4 Analysis of Missing Evidence Identification

Table 4 reports the performance on the missing evidence identification task. Conventional LLM-based methods achieve relatively limited performance, while CEDAR consistently improves Miss.P, Miss.R, and Miss.F1.

Missing evidence identification requires determining whether absent evidence changes the final sufficiency outcome under guideline constraints. Existing generative methods often identify superficially missing information, but fail to distinguish evidence that genuinely affects the final evaluation result.

Model	Method	Suff. Acc	Val. Acc	Val. F1	Mod. Acc	Mod. Macro-F1
Qwen3.5-Plus	Direct-LLM	68.7	84.3	74.7	56.7	55.3
	CoT-LLM	83.0	74.5	69.5	62.4	63.4
	RAG-LLM	71.5	80.5	73.6	60.0	60.5
	CEDAR(ours)	**91.6**	**85.3**	**77.6**	**93.6**	**85.8**
DeepSeek-V3.2	Direct-LLM	57.3	79.7	72.3	57.3	55.3
	CoT-LLM	87.1	**82.5**	72.7	59.2	58.6
	RAG-LLM	87.4	72.4	68.5	59.6	57.9
	CEDAR(ours)	**93.6**	79.6	**73.6**	**95.0**	**87.3**
gpt-4.1-mini	Direct-LLM	50.2	66.6	61.8	60.8	59.2
	CoT-LLM	71.2	72.9	63.1	51.3	49.6
	RAG-LLM	72.1	70.6	**65.1**	53.0	51.5
	CEDAR(ours)	**90.7**	**73.1**	51.9	**93.9**	**86.5**
Qwen3.5-27B	Direct-LLM	67.4	83.9	74.2	57.3	56.3
	CoT-LLM	87.0	79.1	72.1	59.8	59.6
	RAG-LLM	72.7	79.9	72.8	63.7	64.8
	CEDAR(ours)	**92.4**	**84.3**	**74.5**	**94.7**	**87.4**
Qwen3.5-35B-A3B	Direct-LLM	66.7	82.4	74.9	58.2	57.2
	CoT-LLM	80.5	75.9	70.0	65.8	66.6
	RAG-LLM	63.2	79.2	73.1	62.6	63.2
	CEDAR(ours)	**91.3**	**83.1**	**74.9**	**93.5**	**86.1**
Baichuan-M2-32B	Direct-LLM	82.8	81.1	68.7	56.7	55.1
	CoT-LLM	89.6	78.7	70.7	56.7	55.9
	RAG-LLM	83.4	**81.5**	60.8	59.7	57.9
	CEDAR(ours)	**89.9**	73.8	**73.1**	**91.6**	**83.6**
Gemma-4-E4B	Direct-LLM	65.6	75.0	62.7	63.9	63.0
	CoT-LLM	87.2	63.0	21.2	33.1	15.9
	RAG-LLM	88.0	72.6	64.0	64.9	63.8
	CEDAR(ours)	**92.9**	**78.5**	**65.5**	**94.4**	**87.0**

Table 2: Experimental results of different models and methods.

Reasoning Type	**Suff. Acc**	**Val. Acc**	**Miss. P**
Multi-Evidence	**0.926**	0.901	0.548
Staging-Dependent	0.869	0.923	0.609
Exclusion-Based	0.654	**0.929**	0.611
Imaging-Light	0.840	0.796	0.522
Score-Based	0.909	0.922	**0.627**
Symptom-Based	0.646	0.636	0.471

Table 3: Performance of CEDAR across different diagnostic reasoning types.

Method	**Miss. P**	**Miss. R**	**Miss. F1**
Direct-LLM	0.173	0.171	0.171
CoT-LLM	0.321	0.316	0.303
RAG-LLM	0.269	0.238	0.297
CEDAR	**0.597**	**0.535**	**0.547**

Table 4: Performance comparison on the missing evidence identification task.

By contrast, CEDAR performs missing evidence attribution through counterfactual intervention over the evidence–guideline relation graph. Missing evidence is identified according to whether the intervention changes the final sufficiency prediction under deterministic constraint execution. This intervention-based attribution mechanism enables more accurate identification of critical missing evidence.

5.5 Module-Level Analysis

Figure 4 reports the performance of CEDAR across different clinical modules. The symptom, physical examination, and auxiliary examination modules achieve comparatively strong performance, indicating that CEDAR can effectively model clinical evidence with relatively standardized documentation patterns and explicit diagnostic requirements.

Performance is lower on the imaging examination and laboratory examination modules. Although the laboratory examination module maintains relatively high Accuracy, its F1-score decreases noticeably, reflecting the difficulty of handling complex combinational constraints, negative result interpretation, and cross-module evidence interaction. Imaging examination evidence remains particularly challenging due to lengthy descriptions and implicit clinical expressions, which increase the difficulty of semantic evidence alignment.

Method	Suff. Acc	Val. Acc	Val. F1	Mod. Acc	Mod. Macro-F1	Miss. P	Miss. R	Miss. F1
CEDAR w/o Cross-Module Evidence Sharing	81.2	70.3	62.4	82.8	78.3	39.6	35.5	35.8
CEDAR w/o Explicit Evidence–Guideline Alignment	69.4	74.1	55.7	57.9	50.9	10.2	9.4	9.6
CEDAR w/o Intervention	87.9	76.5	70.1	86.6	80.8	44.1	39.1	39.4
CEDAR	**91.6**	**85.3**	**77.6**	**93.6**	**85.8**	**55.2**	**52.5**	**51.7**

Table 5: Ablation study of different components in CEDAR.

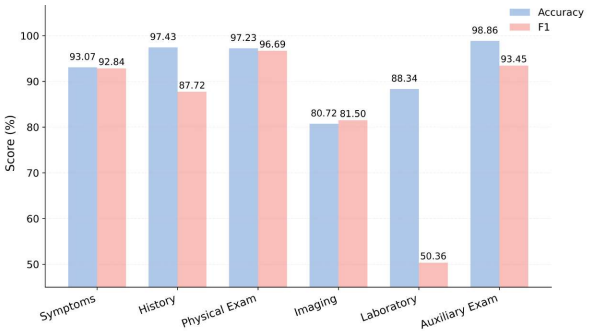

Figure 4: Module-level evaluation performance of CEDAR across different clinical modules.

Metric	Agreement
Sufficiency Agreement	0.88
Validity Agreement	0.91
Module-level Agreement	0.84
Missing Evidence Agreement	0.79
Cohen's Kappa	0.81

Table 6: Agreement between CEDAR and human expert annotations.

5.6 Ablation Study

To analyze the contribution of different components, we conduct three groups of ablation experiments, with results reported in Table 5.

Removing cross-module evidence sharing (**CEDAR w/o Cross-Module Evidence Sharing**) leads to noticeable degradation on Sufficiency and module-level evaluation metrics. This result indicates that cross-module evidence interaction helps alleviate evidence misalignment introduced by structured extraction.

Removing explicit evidence–guideline alignment and deterministic constraint execution (**CEDAR w/o Explicit Evidence–Guideline Alignment**) causes the largest overall performance drop, particularly on module-level evaluation and missing evidence identification. The results show that direct end-to-end generation is insufficient for stable clinical diagnostic evaluation under complex guideline constraints (Nagar et al., 2025).

Removing counterfactual intervention (**CEDAR w/o Intervention**) only moderately affects overall evaluation performance, but substantially reduces missing evidence identification performance. This result demonstrates that intervention-based attribution plays a critical role in identifying evidence that genuinely changes the final sufficiency outcome.

5.7 Comparison with Human Reviewers

To evaluate the consistency of CEDAR with real-world clinical evaluation practice, we compare its predictions with manual reviews conducted by clinically experienced reviewers based on the corresponding clinical guidelines.Table 6 reports the agreement between CEDAR and human reviewers across evidence sufficiency, diagnostic validity, and module-level evaluation.CEDAR also achieves strong agreement on missing evidence identification, indicating that the framework can effectively identify critical missing evidence affecting the final sufficiency outcome.

6 Conclusion

Clinical diagnostic evaluation requires assessing not only diagnostic validity, but also whether diagnostic evidence is sufficiently documented under guideline constraints. We present CEDAR, a structured clinical diagnostic evaluation framework that combines evidence–guideline relation graphs, deterministic constraint execution, and counterfactual intervention for evidence sufficiency assessment, diagnostic validity evaluation, and critical missing evidence identification.

Experimental results demonstrate that explicit guideline constraint modeling provides more stable and interpretable clinical diagnostic evaluation than direct end-to-end generation under complex diagnostic constraint structures.

Limitations

CEDAR relies on structured guideline representations and predefined constraint schemas, which may require additional adaptation across different medical domains and guideline styles. In addition,

the evidence–guideline relation graph is partially constructed through LLM-based semantic relation induction. Although final evaluation results are determined through deterministic constraint execution, errors in semantic relation identification may still affect downstream reasoning in cases involving ambiguous clinical descriptions or implicit evidence.

Furthermore, the current framework mainly focuses on disease-level clinical diagnostic evaluation and does not explicitly model longitudinal clinical progression or multi-visit treatment trajectories. More comprehensive clinical evaluation may require integrating temporal patient history and dynamic clinical evidence.

Ethical Consideration

This work involves clinical diagnostic evaluation based on electronic medical records and clinical guidelines. All electronic medical records used in this study were fully de-identified prior to annotation and experimentation, with personally identifiable information removed in accordance with data privacy protection requirements. No real patient identities, contact information, or sensitive personal attributes were retained in the dataset.

The clinical guidelines used in this work were obtained from publicly available medical guideline documents and were used solely for academic research purposes. In addition, the proposed framework is designed for research on clinical diagnostic auditing and evidence evaluation, rather than direct clinical decision-making. We are not aware of additional ethical concerns arising from this work.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Yanan Bai, Wanying Wang, Jiajia Duan, Xue Wang, Mingjing Jiang, Han Ni, Lu Sun, Chuanxin Liu, and Ding Li. 2025. The evolution and application of the medical evidence chain. *Frontiers in Medicine*, 12:1690197.

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. 2025. Towards omni-rag: Comprehensive retrieval-augmented generation for large language models in medical applications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15285–15309.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1):AIdbp2500418.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Hocine Kadi, Alaa Abdellatif, Daniel Isaac Kemajou Njamen, and Florian Pereme. 2025. Automatic extraction of smpc document for idmp data model construction using foundation llm and rag: a preliminary experiment for pharmaceutical regulatory affairs. *Frontiers in Medicine*, 12:1598979.

Provia Kadusabe, Abhishek Kaushik, and Fiona Lawless. 2025. Dkitnlp at archehr-qa 2025: A retrieval augmented llm pipeline for evidence-based patient question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 165–170.

Simone Kresevic, Mauro Giuffrè, Milos Ajcevic, Agostino Accardo, Lory S Crocè, and Dennis L Shung. 2024. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ digital medicine*, 7(1):102.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dubai Li, Nan Jiang, Kangping Huang, Ruiqi Tu, Shuyu Ouyang, Huayu Yu, Lin Qiao, Chen Yu, Tianshu Zhou, Danyang Tong, and 1 others. 2025. Streamlining evidence based clinical recommendations with large language models. *npj Digital Medicine*.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen,

Michael R Lyu, and Wenting Chen. 2026. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. *Advances in Neural Information Processing Systems*, 38.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Zhuohan Long, Zhijie Bao, and Zhongyu Wei. 2026. Strong reasoning isn't enough: Evaluating evidence elicitation in interactive diagnosis. *arXiv preprint arXiv:2601.19773*.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, Peter Brodeur, and 1 others. 2025. Assessment of large language models in clinical reasoning: a novel benchmarking study. *NEJM AI*, 2(10):AIdbp2500120.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, and 1 others. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Aishik Nagar, Viktor Schlegel, Thanh-Tung Nguyen, Hao Li, Yuping Wu, Kuluhan Binici, and Stefan Winkler. 2025. Llms are not zero-shot reasoners for biomedical information extraction. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 106–120.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Arya S Rao, Kaiz P Esmail, Richard S Lee, Sharon Jiang, Bianca Arraiza Carlo, Jasleen Gill, Praneet Khanna, Ezra Kalmowitz, Basile Montagnese, Kimia Heydari, and 1 others. 2026. Large language model performance and clinical reasoning tasks. *JAMA Network Open*, 9(4):e264003.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson L Parker, Rahul G Krishnan, and Milad Lankarany. 2025. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge. *arXiv preprint arXiv:2502.13010*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

Samuel Sousa, Michael Jantscher, Mark Kröll, and Roman Kern. 2025. Large language models for electronic health record de-identification in english and german. *Information*, 16(2):112.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

C. Wang and 1 others. 2026. Evaluating large language models for evidence-based medicine. *Cell Reports Medicine*.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *NPJ digital medicine*, 7(1):41.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Hang Sheung Wong and Tsz Kwan Wong. 2026. Multi-evidence clinical reasoning with retrieval-augmented generation for emergency triage: Retrospective evaluation study. *JMIR Medical Informatics*, 14(1):e82026.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, and 1 others. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Delvin Ce Zhang, Suhan Cui, Zhelin Chu, Xianren Zhang, and Dongwon Lee. 2026. Mever: Multimodal and explainable claim verification with graph-based evidence retrieval. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5236–5255.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

Zulun Zhu, Tiancheng Huang, Kai Wang, Junda Ye, Xinghe Chen, and Siqiang Luo. 2026. Graph-based approaches and functionalities in retrieval-augmented generation: A comprehensive survey. *ACM Computing Surveys*, 58(10):1–38.

A Structured Guideline Representation Details

A.1 Guideline Structuring Example

Figure A.1 presents examples of the structured guideline representation used in CEDAR. The examples illustrate how original clinical guideline content is transformed into executable structured constraints, including diagnostic modules, logical constraint types, and corresponding guideline items. The upper part shows the Chinese version, while the lower part presents the corresponding English version.

A.2 Executable Constraint Examples

Figure A.2 presents representative executable constraint structures used in CEDAR for deterministic diagnostic reasoning.

Executable Constraint Types	
AND	The module is satisfied only when all required items are satisfied.
OR	The module is satisfied when at least one item is satisfied.
AT_LEAST_N	The module is satisfied when at least N items are satisfied.
OPTIONAL	The module is not treated as a mandatory requirement and does not affect the final sufficiency decision.
CORE AND + OPTIONAL	The module is satisfied only when all core items are satisfied; optional items are recorded but not required.
CORE OR + OPTIONAL	The module is satisfied when at least one core item is satisfied; optional items are recorded but not required.

Figure A.2: Representative executable constraint structures in CEDAR.

A.3 Retrieval-Augmented Parsing Settings

To reduce long-context interference during clinical guideline understanding, CEDAR adopts a retrieval-augmented guideline parsing workflow to construct localized semantic contexts for structured constraint extraction.

Specifically, original clinical guideline PDFs are first parsed into structured document elements using OCR-based document parsing models. Each element contains textual content, element type, page information, and layout metadata. Several post-processing operations are then applied, including element type normalization, cross-page sentence merging, hierarchical paragraph reconstruction based on title levels, and table-title binding. After preprocessing, the guideline document is reorganized into semantically coherent text chunks with associated metadata.

All chunk texts are subsequently encoded into dense vector representations using the BGE embedding model, and a FAISS vector index is constructed for semantic retrieval. For each predefined clinical module (e.g., symptoms, signs, laboratory examinations, and imaging examinations), CEDAR retrieves the most relevant guideline chunks through vector retrieval combined with cross-encoder reranking, thereby constructing localized guideline contexts for downstream structured parsing.

Based on the retrieved guideline context, CEDAR further employs a multi-stage LLM workflow for structured constraint extraction. In the first stage, an extraction agent identifies candidate diagnostic items together with their relevance levels and source positions from the retrieved guideline chunks. In the second stage, a classification agent categorizes each extracted item into different constraint roles, including core conditions, optional conditions, exclusion conditions, and uncertain conditions, while simultaneously identifying their logical relationships. In the third stage, a rule-oriented structuring agent maps the extracted logical relations into executable constraint structures, such as `AND`, `OR`, and `COREAND+OPTIONAL`. Finally, a verification agent performs global consistency checking and structural correction over the generated constraint representation to produce the final structured guideline JSON.

The resulting structured guideline representation is subsequently used for evidence–guideline relation graph construction and deterministic constraint execution.

B Evidence–Guideline Relation Types

Figure A.3 illustrates the detailed construction process of the evidence–guideline relation graph in CEDAR.

C Module-level Constraint Execution Rules

For any clinical module M_i, CEDAR performs module-level reasoning according to its corresponding logical constraint type L_i. The semantic definitions of different logical types are described in Section 3.3, while the corresponding execution rules

are defined as follows.
For modules of type AND:

$$\forall c \in C_i,\ \texttt{satisfied}(c) = \texttt{True} \tag{6}$$

For modules of type OR:

$$\exists c \in C_i,\ \texttt{satisfied}(c) = \texttt{True} \tag{7}$$

For modules of type AT_LEAST_N:

$$\sum_{c \in C_i} \mathbb{I}\big(satisfied(c)\big) \geq N \tag{8}$$

For modules of type OPTIONAL, the module is regarded as satisfied by default and does not participate in the final sufficiency constraint.For modules of type COREAND+OPTIONAL and COREOR+OPTIONAL, the corresponding logical constraints are applied only to the core item subset, while optional items are excluded from mandatory constraint evaluation.

D Diagnostic Reasoning Type Definitions

Multi-Evidence Conjunctive Type This disease type requires multiple evidence dimensions to be jointly satisfied, where no single piece of evidence alone is sufficient to support the final diagnosis. Representative diseases include pneumonia, heart failure, and acute pancreatitis.

Staging-Dependent Chronic Disease Type The diagnosis of this disease type relies on long-term evidence accumulation or disease staging information rather than results from a single examination. Representative diseases include chronic obstructive pulmonary disease (COPD), chronic kidney disease, and liver cirrhosis.

Exclusion Diagnosis Type This disease type relies on systematically ruling out other etiologies rather than depending solely on positive evidence. Representative diseases include irritable bowel syndrome, gastroesophageal reflux disease, and tension-type headache.

Imaging-Light Dependent Type In this disease type, imaging findings alone cannot determine the final diagnosis and need to be combined with clinical symptoms and disease course information for comprehensive judgment. Representative diseases include ischemic stroke, transient ischemic attack, and lumbar disc herniation.

Score-Based Accumulation Type The diagnosis of this type is determined by the accumulation of multiple discrete conditions and requires meeting predefined thresholds or classification criteria. Representative diseases include systemic lupus erythematosus, rheumatoid arthritis, and polymyositis.

Symptomology-Based Diagnostic Type This type lacks a clear gold standard, and diagnosis primarily relies on symptom patterns and disease course information. Representative diseases include major depressive disorder, anxiety disorders, and insomnia disorder.

E Structured Prompting Details

The prompting workflow in CEDAR is divided into two stages. The first stage focuses on structured evidence extraction from heterogeneous electronic medical record modules, including symptoms, medical history, physical examinations, laboratory examinations, imaging examinations, auxiliary examinations, pathological examinations, and diagnosis extraction. The second stage focuses on evidence–guideline relation graph construction for deterministic constraint reasoning.

F Human Annotation

We involved 15 clinically experienced medical experts for annotation adjudication and quality verification. All experts had professional clinical knowledge and experience with diagnostic guideline interpretation. The total cost for expert annotation and review was approximately $2,000.

Part of the Rheumatoid Arthritis Guideline

The typical manifestation is arthritis, with varying degrees of pain and swelling, which may be accompanied by limited movement and morning stiffness lasting more than 1 hour. The most common involvement is the proximal interphalangeal joints, metacarpophalangeal joints, wrists, elbows, shoulders, knees, ankles, and toes, usually symmetrical; It can also involve the cervical spine, temporomandibular joint, sternoclavicle, and acromioclavicular joints. Patients with long disease courses may develop joint deformities, such as wrist ankylosis, limited elbow extension, ulnar deviation of the metacarpophalangeal joints, "swan neck" and "button flower" deformities of the fingers, etc. In severe cases, gradual atrophy of the muscles around the joints leads to further loss of function and inability to care for oneself. ("page": 2,"side": "left")

(1) Blood routine: Common chronic diseases in active phase RA patients include anemia, increased platelet count, and elevated acute phase reactants such as erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP).

(2) Autoantibodies: The most commonly used and most valuable autoantibodies for RA diagnosis are rheumatoid factor (RF) and anticyclic citrulline peptide (CCP) antibodies. Meta-analysis showed that for RA diagnosis, RF had a sensitivity of 69% and specificity of 85%; The sensitivity of anti-CCP antibodies is 50%~80%, but their specificity is > 90%. RA patients with high titers of anti-CCP antibodies or RF-positive are more prone to bone destruction, often accompanied by extra-articular manifestations such as cardiovascular disease and interstitial lung lesions, which are poor prognosis factors for RA. Additionally, anti-keratin antibodies (AKA), anti-perinuclear factor (APF), anti-mutant citrulline vibrate protein (MCV) antibodies, anticarbasylated protein (CarP) antibodies, and anti-glucose-6-phosphate isomerase (GPI) antibodies also have certain value in diagnosing RA.

(3) Imaging Examinations 1. X-ray: The most classic imaging method for detecting joint bone destruction in RA patients, but it shows poor sensitivity in early bone destruction and cannot reveal inflammatory lesions such as synovitis or myeloedema. Therefore, it has significant limitations in early diagnosis and short-term efficacy assessment of RA. Currently, it is mainly used in follow-up to assess the progression of bone structural changes. Changes in RA on X-ray can be divided into four stages: Stage I shows osteoporosis without bone destruction; Stage II shows osteoporosis, which may have mild subchondral bone destruction but no joint deformities; Stage III shows osteoporosis, bone destruction, and joint deformities, with no fibrous or bony ankylosi; Fibrous or bony ankylosis appears in stage IV. 2. Ultrasound: It can detect various common RA lesions such as synovial hyperplasia/synovitis, joint cavity effusion, tendinitis/tenosynovitis, bursitis, and bone erosion. Ultrasound has a high detection rate for synovial inflammation, allowing for very early detection of bone erosion, so ultrasound can significantly improve the diagnostic rate of RA. 3. Magnetic Resonance Imaging (MRI): MRI not only identifies typical lesions such as synovitis and bone erosion, but can also detect myeloedema, which represents early inflammation, providing important evidence for early diagnosis, disease activity, and prognosis assessment of RA. MRI is comparable in value to auxiliary diagnosis and monitoring efficacy compared to ultrasound, but due to high cost, time-consuming nature, and equipment limitations, it is rarely used in RA patient follow-up, and its use nationwide, especially in primary hospitals, is also greatly restricted. ("page": 3,"side": "left")

RA classification standards established by the American College of Rheumatology (ARA) in 1987: see Table 1. 2. RA classification criteria established by ACR/EULAR in 2010: (1) At least one joint presents with clinical synovitis; (2) Synovitis cannot be explained by other diseases; (3) X-rays did not show typical bone erosion changes. If the above three conditions are met, four scoring items are performed (Table 2), with a maximum score of 10 points. When the total score is ≥ 6, RA can be diagnosed. 3. The RA classification criteria established by ARA in 1987 can effectively distinguish between inflammatory and non-inflammatory arthritis, but cannot sensitively identify early RA. ("page": 3,"side": "right")


```
{
  "Diagnosis": "Rheumatoid arthritis",
  "Diagnostic criteria": { "Symptoms": { "Logic": "OR", "Items": [
{"Content": "Arthritis","Source": {"page": 2,"side": "left","path": ""}},
{"Content": "Varying degrees of pain and swelling","Source": {"page": 2,"side": "left","path": "}},
{"Content": "Pain and swelling of varying degrees, morning stiffness lasting more than 1 hour",
                "Source": {"page": 2,"side": "left","path": ""}},
{"Content": "Joint Deformity","Source": {"page": 2,"side": "left","path": "}},
{"Content": "Gradually atrophy of muscles around the joint","Source": {"page": 2,"side": "left","path":"}}]},
  "Signs": {}, "Medical History": {},
  "Laboratory Test": {"Logic": "COREAND+OPTIONAL", "Core Item": [
{"Content": "Complete Blood Routine","Source": {"page": 3,"side": "left","path": "}},
{"Content": "Autoantibody, Anti-citrulline protein autoantibody (ACPA)and rheumatoid factor (RF) are
                important serological markers for RA.""Source": {"page": 3,"side": "left","path": ""}},
  "Optional": [
{"Content": "Antikeratin antibodies (AKA) have certain value in diagnosing RA", "Source":
                {"page": 3,"side": "left","path": ""}},
{"Content": "Perinuclear factor (APF) has some value in diagnosing RA", "Source": {"page":
                3,"side": "left","path": ""}},
{"Content": "Anti-mutant citrulline vibroprotein (MCV) antibodies have certain value in
                diagnosing RA","Source": {"page": 3,"side": "left","path": "}},
{"Content": "Anticaramylated protein (CarP) antibodies have certain value in diagnosing RA",
                "Source": {"page": 3,"side": "left","path": ""}},
{"Content": "Anti-glucose-6-phosphate isomerase (GPI) antibodies have certain value in
                diagnosing RA", "Source": {"page": 3,"side": "left","path": "}}]},
  "Physical Examination": {},
  "Imaging Examination": {"Logic": "OR", "Core Item": [
{"Content": "X-ray","Source": {"page": 3,"side": "left","path": ""}},
{"Content": "Ultrasound","Source": {"page": 3,"side": "left","path": ""}},
{"Content": "Magnetic Resonance Imaging (MRI)","Source": {"page": 3,"side": "left","path": "}}]},
  "Auxiliary Examination": {}, "Pathological Examination": {},"Classification and Staging
    Criteria": {}, "Differential Diagnosis": {},
  "Diagnostic criteria": {"Logic": "OR", "Items": [
{"Content": " The 1987 American College of Rheumatology classification criteria for rheumatoid arthritis include
objective synovitis lasting no less than 6 weeks. Inflammation is a prerequisite, requiring at least 1 hour of morning
stiffness, polyarthritis, hand arthritis, and symmetrical joint involvement. The seven criteria also include rheumatoid
nodules, positive rheumatoid factor, and wrist X-ray showing bone erosions. Meeting 4 or more of these 7 indicators is
used to distinguish between inflammatory and non-inflammatory arthritis.","Source": {"page": 3,"side": "right","path":
"}},
{"Content": "At least one joint has clinical synovitis manifestations that cannot be explained by other diseases.
Moreover, patients have not yet shown typical bone erosion changes on radiographic X-rays. After meeting the above
prerequisites, the scoring system is applied, which is divided into four domains: the number of affected joints,
autoantibody status, levels of acute phase reactants, and duration of synovitis. Quantitative scoring is conducted for
each domain. For the number of affected joints, scores range from 0 to 0.5 points based on involvement of major and
small joints. For autoantibodies (rheumatoid factor and anti-cyclic citrullinated peptide antibodies), negative status,
low-titer positive, or high-titer positive yield scores from 0 to 3 points. For acute phase reactants, elevation of C-
reactive protein or erythrocyte sedimentation rate scores 0 or 1 point. For duration of synovitis, reaching 6 weeks
scores 0 or 1 point. The maximum total score for these four items is 10 points. When the total score reaches or exceeds
6 points, the patient can be classified as having rheumatoid arthritis.","Source": {"page": 3,"side": "right","path":
""}}]}},
  "Source of Knowledge": "Rheumatoid Arthritis Diagnosis and Treatment Standards"
}
```


Structured Guideline for Rheumatoid Arthritis

典型表现为关节炎，不同程度的疼痛肿胀，可 以伴活动受限，晨僵长达 1 h 以上。以近端指间关 节、掌指关节、腕、肘、肩、膝、踝和足趾关节受累最 为多见，通常呈对称性；亦可累及颈椎、颞颌关节、 胸锁和肩锁关节。长病程患者可以发生关节畸形， 如腕关节强直、肘关节伸直受限、掌指关节尺侧偏 斜、手指的“天鹅颈”和“纽扣花”畸形等等，严重者 关节周围肌肉逐渐萎缩导致功能进一步丧失、生活 不能自理。("page": 2,"side": "left")

（一）血常规 活动期 RA 患者常见慢性病贫血、血小板计数 增多 ，以及红细胞沉降率（ESR）和 C 反应蛋白 （CRP）等急性期反应物升高。

（二）自身抗体 临床上普遍使用且对RA诊断价值最高的自身 抗体是类风湿因子（RF）和抗环瓜氨酸多肽（CCP） 抗体。Meta 分析显示，对 RA 诊断，RF 的敏感度为 69%，特异度为85%；抗CCP抗体的敏感度为 50%~ 80%，但特异度>90%。高滴度抗 CCP抗体或 RF阳性的RA患者更易发生骨破坏，合并心血管疾病、肺 间质病变等关节外表现，是 RA 预后不良因素。另 外，抗角蛋白抗体（AKA）、抗核周因子（APF）、抗突 变型瓜氨酸波形蛋白（MCV）抗体、抗氨甲酰化蛋白 （CarP）抗体和抗葡萄糖-6-磷酸异构酶（GPI）抗体 对诊断RA亦有一定价值。

（三）影像学检查 1. X 线：是检测 RA 患者关节骨破坏最经典的 影像学方法，但发现早期骨破坏的敏感性差，亦不 能显示滑膜炎、骨髓水肿等炎性病变，因此对RA的 早期诊断和短期疗效判定有非常大的局限性，目前 主要用于随访中，以判定患者骨结构改变的进展情 况。X 线上 RA 的改变可分为 4 期：Ⅰ期见骨质疏 松，无骨破坏；Ⅱ期见骨质疏松，可有轻度软骨下骨 质破坏，无关节畸形；Ⅲ期见骨质疏松、骨质破坏和 关节畸形，无纤维性或骨性强直；Ⅳ期出现纤维或 骨性强直。 2. 超声：可以发现 RA 关节滑膜增生/滑膜炎 症、关节腔积液、肌腱炎/腱鞘炎、滑囊炎、骨侵蚀等 多种RA常见的病变。超声对滑膜炎症有较高的检 出率，可以非常早期发现骨侵蚀，因此借助超声可明显提高 RA 的诊断率。3. 磁共振成像（MRI）：MRI 在识别滑膜炎、骨 侵蚀等典型病变的基础上，还可以发现代表早期炎 症的骨髓水肿，因此可以为RA的早期诊断、判定疾 病活动性和评估预后提供重要依据。MRI 在辅助 诊断和监测疗效方面的价值与超声相当，但受到费 用高、耗时长、设备等限制，在 RA 患者随访中较少 应用，在全国，尤其是基层医院的应用也受到很大 制约。("page": 3,"side": "left")

1987年美国风湿病学会（ARA）制定的RA分 类标准 ：见表1。 2. 2010 年 ACR/EULAR 制定的 RA 分类标 准：（1）至少一个关节表现为临床滑膜炎；（2）滑膜炎不能用其他疾病解释；（3）X 线未见到典型的 骨侵蚀改变。 如满足上述 3个条件，则进行 4项评分（表 2）， 最高分为10分，当总分≥6分时可诊断RA。 3. 1987 年 ARA 制定的 RA 分类标准可很好地 区分炎性和非炎性关节炎，但不能敏感地识别早期 RA。("page": 3,"side": “right")


```
{
  "诊断名称": "类风湿关节炎",
  "诊断依据": { "症状": {"逻辑": "OR","项目": [
{"内容": "关节炎","来源": {"page": 2,"side": "left","path": ""}},
{"内容": "不同程度的疼痛肿胀","来源": {"page": 2,"side": "left","path": ""}},
{"内容": "不同程度的疼痛肿胀,晨僵持续时间长达1小时以上","来源": {"page": 2,"side":
                "left","path": ""}},
{"内容": "关节畸形","来源": {"page": 2,"side": "left","path": ""}},
{"内容": "关节周围肌肉逐渐萎缩","来源": {"page": 2,"side": "left","path": ""}}]},
  "体征": {},"病史": {},
  "实验室检查": {"逻辑": "COREAND+OPTIONAL","核心项": [
{"内容": "血常规","来源": {"page": 3,"side": "left","path": ""}},
{"内容": "自身抗体，抗瓜氨酸蛋白自身抗体（Anti-citrulline protein autoantibody, ACPA）
                和类风湿因子（rheumatoid factor, RF）是RA的重要血清学标志物",
                "来源": {"page": 3,"side": "left","path": ""}}],
  "可选项": [
{"内容": "抗角蛋白抗体(AKA)对诊断RA有一定价值","来源": {"page": 3,
                "side": "left","path": ""}},
{"内容": "抗核周因子(APF)对诊断RA有一定价值","来源": {"page": 3,"side":
                "left","path": ""}},
{"内容": "抗突变型瓜氨酸波形蛋白(MCV)抗体对诊断RA有一定价值",
                "来源 ": {"page": 3,"side": "left","path": ""}},
{"内容": "抗氨甲酰化蛋白(CarP)抗体对诊断RA有一定价值","来源": {"page":
                3,"side": "left","path": ""}},
{"内容": "抗葡萄糖-6-磷酸异构酶(GPI)抗体对诊断RA有一定价值",
                "来源": {"page": 3,"side": "left","path": ""}}]},
  "体格检查": {},
  "影像学检查": {"逻辑": "OR","核心项": [
{"内容": "X线","来源": {"page": 3,"side": "left","path": ""}},
{"内容": "超声","来源": {"page": 3,"side": "left","path": ""}},
{"内容": "磁共振成像（MRI）","来源": {"page": 3,"side": "left","path": ""}}]},
  "辅助检查": {},"病理学检查": {},"分型与分期标准": {},"鉴别诊断": {},
  "诊断标准": {"逻辑": "OR","项目": [
{"内容": "1987 年美国风湿病学会类风湿关节炎分类标准以持续不少于 6 周的客观滑膜炎为前提，要求在晨僵持续至少 1 小时、多关节炎、手关节炎、对称性关节炎、类风湿结节、类风湿因子阳性以及手腕关节 X 线片显示骨侵蚀改变这 7 项指标中满足 4 项及以上，用于区分炎性与非炎性关节炎","来源": {"page": 3,"side": "right","path": ""}},
{"内容": "至少一个关节存在临床滑膜炎表现，且该滑膜炎不能由其他疾病解释，并且关节 X 线检查尚未出现典型骨侵蚀改变的患者。在满足上述前提条件后，分别从受累关节数、自身抗体状态、急性期反应物水平以及滑膜炎持续时间四个方面进行量化评分，其中受累关节数依据大关节和小关节受累情况计 0 至 5 分，自身抗体根据类风湿因子和抗环瓜氨酸肽抗体的阴性、低滴度阳性或高滴度阳性计 0 至 3 分，急性期反应物依据 C 反应蛋白或红细胞沉降率是否升高计 0 或 1 分，滑膜炎持续时间依据是否达到 6 周计 0 或 1 分，四项评分总分最高为 10 分，当总评分达到或超过 6 分时，可将患者分类为类风湿关节炎。","来源": {"page": 3,"side": "right","path":
""}}]}},
  "知识来源": "类风湿关节炎诊疗规范"
}
```


Figure A.1: Examples of structured guideline representations in CEDAR. The upper figure shows the Chinese version, and the lower figure shows the corresponding English version.

Relation Type Definitions			
Relation Type	Core Definition	Clinical Evidence Example	Guideline Item
Support	The evidence semantically satisfies the corresponding guideline criterion.	Blood glucose 11.2 mmol/L	Elevated blood glucose
Performed	The corresponding examination or assessment has been completed, while the result is normal or negative.	Routine blood test shows no obvious abnormalities	Complete routine blood test
Contradiction	The evidence semantically conflicts with the corresponding guideline criterion.	Patient has no fever	Fever

Figure A.3: Examples of evidence–guideline relation definitions in CEDAR.

Prompt for Symptoms

```
Please extract the patient's relevant symptoms from the medical
records, make sure not to omit any relevant information, and try
to output as accurately as possible; do not fabricate information
Please extract the above information from the provided medical
records in the following format (no fabrication allowed):
{{
    "Symptoms": [[
      "Symptom 1",
      "Symptom 2",
      "……"
    ]]
}}
```

```
请从病历中提取病人的相关症状，务必不要遗漏任何相关信息，尽量按照原文输
出，不得编造
请按照以下格式，从提供的病历中提取以上信息（所有信息不得编造）：
{{
    "症状": [[
      "症状1",
      "症状2",
      "……"
    ]]
}}
```

Figure A.4: Prompt template for symptom extraction.

Prompt for Patient Information

```
You are a professional medical record reviewer skilled at extracting basic patient information
from the original medical records.
Please extract [all information related to the patient's medical history] from the original medical
record.
[Draw Scope Explanation]
Medical history information includes but is not limited to:
- The time of symptom onset, how it develops, and how it progresses
- Description of current medical history
- Past history
- Personal history (such as smoking, drinking, occupation, drug or toxic exposure, etc.)
- Family history
- Menstrual history / marital and childbearing history (as mentioned in medical records)
- Medication history (including long-term medication, recent medication, special medication)
- History of substance abuse or addiction
- Characteristics of the occurrence and progression of neurological symptoms (such as onset time,
progression/remission process, etc.)
- Any other information that clearly belongs to the "medical history."
[Important Rules]
1. Be sure not to omit any medical history-related information that appears in the medical records.
2. Strictly extract from the original text, strive to preserve the original expression, and do not
summarize, speculate, or fabricate content.
3. If a specific type of medical history information is not mentioned in the medical record, this
field will not be output.
4. Field names can be flexibly added or removed according to the medical record content, not
limited to the fields shown in the example.
[Output Format Requirements]
Please output in JSON format. The structure example is as follows. Specific fields can be
adjusted according to the actual medical record content:
{
 "Medical history": {
   "Symptom onset time": ".....",
   "Past history": ".....",
   "Personal History": ".....",
   "Family History": ".....",
   "Menstrual History": ".....",
   "Medication history": ".....",
   "Other Medical History": "....."
 }
}
Please note:
- The above fields are for example only;
- If there are other types of medical history information in the medical record, please add the
corresponding field under "Medical History";
- Only output content that truly exists in the medical records.
```

```
你是一位专业病历审查专家，擅长从病历原文中找出病人基本信息。
请从病历原文中提取【所有与病人病史有关的信息】。
【抽取范围说明】
病史信息包括但不限于：
- 症状出现时间、起病方式及演变过程
- 现病史相关描述
- 既往史
- 个人史（如吸烟、饮酒、职业、药物或毒物接触等）
- 家族史
- 月经史 / 婚育史（如病历中涉及）
- 用药史（含长期用药、近期用药、特殊用药）
- 药物滥用史或成瘾史
- 神经症状发生及进展特征（如出现时间、加重/缓解过程等）
- 其他任何明确属于“病史”的信息
【重要规则】
1. 务必不要遗漏病历中出现的任何病史相关信息。
2. 严格依据原文提取，尽量保持原文表述，不得总结、推测或编造。
3. 若病历中未提及某一类病史信息，则不输出该字段。
4. 字段名称可根据病历内容灵活增减，不限于示例中的字段。
【输出格式要求】
请以 JSON 格式输出，结构示例如下，具体字段可根据实际病历内容调整：
{
  "病史": {
    "症状出现时间": "……",
    "既往史": "……",
    "个人史": "……",
    "家族史": "……",
    "月经史": "……",
    "用药史": "……",
    "其他病史": "……"
  }
}
请注意：
- 上述字段仅作为示例；
- 若病历中存在其他类型的病史信息，请在“病史”下新增相应字段；
- 仅输出病历中真实存在的内容。
```

Figure A.5: Prompt template for medical history extraction.

Prompt for Physical Examination

You are a professional medical record reviewer, skilled at identifying relevant physical examinations from the original medical records.
Please extract [all physical examination/special examination-related information] from the original medical record.
[Draw Scope Explanation]
Physical examination / specialized examinations include but are not limited to:
- General information: body temperature, pulse, respiration, blood pressure, height and weight, consciousness, nutritional status, mental state, etc
- Head and neck examination: skull shape, facial features, sclera, conjunctiva, pupils, eye movement, ears, nose, oropharynx, thyroid, cervical lymph nodes, etc
- Chest examination: visual inspection, palpation, percussion, auscultation, apical pulse, cardiac border, heart sounds, murmurs, lung breath sounds, dry and wet rales, etc
- Abdominal examination: visual inspection, palpation, percussion, auscultation, liver and spleen palpation, tenderness, rebound pain, mass, ascites signs, etc
- Spine and limbs: spinal morphology, joint mobility, muscle strength, muscle tone, edema, deformities, etc
- Neurological examination: consciousness, speech, cranial nerves, muscle strength, muscle tone, sensation, tendon reflexes, pathological reflexes, mutual aids, meningeal stimulation signs, etc
- Respiratory function/airway assessment: breathing patterns, airway patency, signs of dyspnea, etc
- Circulatory function assessment: heart rate, rhythm, peripheral pulse, peripheral circulation, skin temperature, etc
- Any other physical examination or specific sign descriptions that appear in the medical record
[Important Rules]
1. Be sure not to omit any information related to physical examinations or special examinations that appear in the medical records.
2. Strictly extract from the original text, strive to preserve the original expression, and do not summarize, speculate, or fabricate content.
3. If a particular type of physical examination is not mentioned in the medical record, that field will not be output.
4. Field names can be flexibly added or removed according to the medical record content, not limited to the fields shown in the example; If the same site/system is described in a medical record, please merge or retain multiple entries as they are.
[Output Format Requirements]
Please output in JSON format. The structure example is as follows, and the fields can be adjusted according to the actual medical record content:
{
"Physical Examination": {
"General Situation": ".....",
"Head and Neck Examination": ".....",
"Chest Examination": ".....",
"Abdominal Examination": ".....",
"Airway function": ".....",
"Respiratory function": ".....",
"Loop Function": ".....",
"Neurological Examination": "....."
}
}
Please note:
- The above fields are for example only;
- If there are other physical examinations or special examination items in the medical record, please add the corresponding field under "Physical Examination";
- Only output content that truly exists in the medical records.

你是一位专业病历审查专家，擅长从病历原文中找出相关身体检查。
请从病历原文中提取【所有体格检查 / 专项检查相关信息】。
【抽取范围说明】
体格检查 / 专项检查包括但不限于：
- 一般情况：体温、脉搏、呼吸、血压、身高体重、意识状态、营养状况、精神状态等
- 头颈部检查：头颅外形、面容、巩膜、结膜、瞳孔、眼球活动、耳、鼻、口咽、甲状腺、颈部淋巴结等
- 胸部检查：视诊、触诊、叩诊、听诊、心尖搏动、心界、心音、杂音、肺部呼吸音、干湿啰音等
- 腹部检查：视诊、触诊、叩诊、听诊、肝脾触诊、压痛、反跳痛、包块、腹水征等
- 脊柱与四肢：脊柱形态、关节活动度、肌力、肌张力、水肿、畸形等
- 神经系统检查：意识、语言、颅神经、肌力、肌张力、感觉、腱反射、病理反射、共济运动、脑膜刺激征等
- 呼吸功能 / 气道评估：呼吸模式、气道通畅性、呼吸困难体征等
- 循环功能评估：心率、节律、周围脉搏、末梢循环、皮肤温度等
- 其他任何在病历中出现的体格检查或专项体征描述
【重要规则】
1. 务必不要遗漏病历中出现的任何体格检查 / 专项检查相关信息。
2. 严格依据原文提取，尽量保持原文表述，不得总结、推测或编造。
3. 若病历中未提及某一类体格检查内容，则不输出该字段。
4. 字段名称可根据病历内容灵活增减，不限于示例中的字段；如同一部位/系统在病历中分散描述，请合并或按原文保留多条。
【输出格式要求】
请以 JSON 格式输出，结构示例如下，字段可根据实际病历内容调整：
{
"体格检查": {
"一般情况": "……",
"头颈部检查": "……",
"胸部检查": "……",
"腹部检查": "……",
"气道功能": "……",
"呼吸功能": "……",
"循环功能": "……",
"神经系统检查": "……"
}
}
请注意：
- 上述字段仅作为示例；
- 如病历中存在其他体格检查或专项检查内容，请在“体格检查”下新增对应字段；
- 仅输出病历中真实存在的内容。

Figure A.6: Prompt template for physical examination extraction.

Prompt for Lab Examination

You are a professional medical record reviewer skilled at identifying relevant laboratory tests from the original medical records.
Please extract information about [all laboratory tests and their results] from the original medical records.
[Draw Scope Explanation]
Laboratory tests include but are not limited to:
- Hematology tests: complete blood count/complete blood count, classification count, hemoglobin, platelets, reticulocytes, coagulation function (PT, APTT, INR, TT, Fibrinogen, D-dimer, etc.)
- Clinical biochemical tests: liver function (ALT, AST, TBil, DBil, ALB, etc.), kidney function (BUN, Cr, UA, etc.), blood glucose, glycated hemoglobin, blood lipids, electrolytes, myocardial enzyme profile/myocardial injury markers, etc
- Immunological tests: autoimmune (ANA, ENA, etc.), infection-related (hepatitis B, C, SYPHILIS, HIV antibodies, etc.), inflammatory markers (CRP, PCT, ESR, etc.)
- Fluid examination: urinalysis, urine sediment, quantitative urine protein, occult blood, parasite eggs, cerebrospinal fluid, pleural and ascites fluid, etc
- Blood gas analysis and related indicators (pH, PaO_2, $PaCO_2$, oxygen saturation, etc.)
- Any other "lab/laboratory/laboratory tests" appearing in the medical record and their corresponding results, values, units, reference scopes, and tips (including original statements such as "abnormality/no abnormality/pending report"), etc.
[Important Rules]
1. Be sure not to omit any laboratory test information (test name + result/value/conclusion/tip) that appears in the medical record.
2. Strictly extract from the original text, strive to preserve the original expression, and do not summarize, speculate, or fabricate content.
3. If a particular examination is not mentioned in the medical record, the field will not be output.
4. Field names can be flexibly added or removed according to the medical record content, not limited to the fields shown in the example; If the same check appears multiple times, please list them separately as original or keep multiple original entries in the same field.
[Output Format Requirements]
Please output in JSON format. The structure example is as follows, and the fields can be adjusted according to the actual medical record content:
{
"Laboratory Test": {
"Blood sugar": ".....",
"HbA1c Hemoglobin": ".....",
"Creatinine": ".....",
"Blood lipids": ".....",
"Myocardial Ischemia Biomarkers": ".....",
"Complete blood count": ".....",
"Prothrombin Time PT": ".....",
"Activated partial thromboplastin time APTT": ".....",
"Electrolyte": ".....",
"Oxygen Saturation": "....."
}
}
Please note:
- The above fields are for example only;
- If there are other types of laboratory tests in the medical record, please add the corresponding field under "Laboratory Examinations";
- Only output content that truly exists in the medical records.

你是一位专业病历审查专家，擅长从病历原文中找出相关实验室检查。
请从病历原文中提取【所有实验室检查及其结果】的信息。
【抽取范围说明】
实验室检查包括但不限于：
- 血液学检查：血常规/全血细胞计数、分类计数、血红蛋白、血小板、网织红细胞、凝血功能（PT、APTT、INR、TT、纤维蛋白原、D-二聚体等）
- 临床生化检查：肝功能（ALT、AST、TBil、DBil、ALB等）、肾功能（BUN、Cr、UA等）、血糖、糖化血红蛋白、血脂、电解质、心肌酶谱/心肌损伤标志物等
- 免疫学检查：自身免疫（ANA、ENA等）、感染相关（乙肝、丙肝、梅毒、HIV 抗体等）、炎症指标（CRP、PCT、ESR等）
- 体液学检查：尿常规、尿沉渣、尿蛋白定量、隐血、寄生虫卵、脑脊液、胸腹水等
- 血气分析及相关指标（pH、PaO_2、$PaCO_2$、氧饱和度等）
- 其他任何在病历中出现的“化验/检验/实验室检测”及其对应结果、数值、单位、参考范围、提示（含“异常/未见异常/待回报”等原文表述）
【重要规则】
1. 务必不要遗漏病历中出现的任何实验室检查相关信息（检查名称 + 结果/数值/结论/提示）。
2. 严格依据原文提取，尽量保持原文表述，不得总结、推测或编造。
3. 若病历中未提及某项检查，则不输出该字段。
4. 字段名称可根据病历内容灵活增减，不限于示例中的字段；如同一检查多次出现，请按原文分别列出或在同一字段内保留多条原文。
【输出格式要求】
请以 JSON 格式输出，结构示例如下，字段可根据实际病历内容调整：
{
"实验室检查": {
"血糖": "……",
"糖化血红蛋白": "……",
"肌酐": "……",
"血脂": "……",
"心肌缺血标志物": "……",
"全血细胞计数": "……",
"凝血酶原时间PT": "……",
"活化部分凝血活酶时间APTT": "……",
"电解质": "……",
"氧饱和度": "……"
}
}
请注意：
- 上述字段仅作为示例；
- 如病历中存在其他类型的实验室检查，请在“实验室检查”下新增对应字段；
- 仅输出病历中真实存在的内容。

Figure A.7: Prompt template for laboratory examination extraction.

Prompt for Imaging Examination

You are a professional medical record reviewer skilled at identifying relevant imaging tests from original medical records.
Please extract information on [all imaging examinations and their results] from the original medical records.
[Draw Scope Explanation]
Imaging examinations include but are not limited to:
- X-ray examinations: standard X-rays, digital X-rays (DR), barium meal/barium enema and other contrasts, mammography, dental panoramas, etc
- CT examinations: plain CT (plain scan), enhanced CT, multi-stage/multimodal CT, perfusion CT, CT angiography (CTA), coronary CTA, etc
- MRI examinations: standard MRI, enhanced MRI, high-resolution magnetic resonance imaging (HRMRI), magnetic resonance angiography (MRA), magnetic resonance venography (MRV), etc
- Ultrasound examinations: abdominal ultrasound, echocardiography, gynecological/obstetric ultrasound, vascular ultrasound (such as carotid artery and lower limb veins), transcranial Doppler (TCD), etc
- Nuclear medicine tests: PET-CT, SPECT, etc
- Interventional/vascular imaging: digital subtraction angiography (DSA), etc
- Other special imaging examinations: endoscopic imaging recording, functional imaging, dynamic imaging, etc. (as shown in medical records)
[Important Rules]
1. Be sure not to omit any imaging examination information (exam name + examination method + location + result/conclusion/tip) that appear in the medical record.
2. Strictly extract from the original text, strive to preserve the original expression, and do not summarize, speculate, or fabricate content.
3. If a particular imaging examination is not mentioned in the medical record, that field will not be output.
4. Field names can be flexibly added or removed according to the medical record content, not limited to the fields shown in the example; If the same imaging examination occurs multiple times at different times or in different ways, please list them separately according to the original text or keep multiple original entries in chronological order within the same field.
[Output Format Requirements]
Please output in JSON format. The structure example is as follows, and the fields can be adjusted according to the actual medical record content:
{
"Imaging Examination": {
"Flat CT": ".....",
"Multi-mode CT": ".....",
"Perfusion CT": ".....",
"Carotid Ultrasound": ".....",
"Transcranial Doppler (TCD)": ".....",
"MRA": ".....",
"CTA": ".....",
"High-Resolution Magnetic Resonance Imaging (HRMRI)": ".....",
"Digital Subtraction Angiography (DSA)": "....."
}
}
Please note:
- The above fields are for example only;
- If there are other types of imaging examinations in the medical record, please add the corresponding field under "Imaging Examination";
- Only output content that truly exists in the medical records.

你是一位专业病历审查专家，擅长从病历原文中找出相关影像学检查。
请从病历原文中提取【所有影像学检查及其结果】的信息。
【抽取范围说明】
影像学检查包括但不限于：
- X线检查：普通X线平片、数字化X线摄影（DR）、钡餐/钡灌肠等造影、乳腺钼靶、牙科全景片等
- CT检查：普通CT（平扫）、增强CT、多期/多模式CT、灌注CT、CT血管成像（CTA）、冠脉CTA等
- MRI检查：普通MRI、增强MRI、高分辨磁共振成像（HRMRI）、磁共振血管成像（MRA）、磁共振静脉成像（MRV）等
- 超声检查：腹部彩超、心脏超声、妇科/产科超声、血管超声（如颈动脉、下肢静脉等）、经颅多普勒（TCD）等
- 核医学检查：PET-CT、SPECT 等
- 介入/血管影像：数字减影血管造影（DSA）等
- 其他特殊影像学检查：内镜下影像记录、功能成像、动态影像等（如病历中出现）
【重要规则】
1. 务必不要遗漏病历中出现的任何影像学检查相关信息（检查名称 + 检查方式 + 部位 + 结果/结论/提示）。
2. 严格依据原文提取，尽量保持原文表述，不得总结、推测或编造。
3. 若病历中未提及某项影像学检查，则不输出该字段。
4. 字段名称可根据病历内容灵活增减，不限于示例中的字段；如同一影像学检查在不同时间或不同方式下多次出现，请按原文分别列出或在同一字段内按时间顺序保留多条原文。
【输出格式要求】
请以 JSON 格式输出，结构示例如下，字段可根据实际病历内容调整：
{
"影像学检查": {
"平扫CT": "……",
"多模式CT": "……",
"灌注CT": "……",
"颈动脉超声": "……",
"经颅多普勒（TCD）": "……",
"MRA": "……",
"CTA": "……",
"高分辨磁共振成像（HRMRI）": "……",
"数字减影血管造影（DSA）": "……"
}
}
请注意：
- 上述字段仅作为示例；
- 如病历中存在其他影像学检查类型，请在“影像学检查”下新增对应字段；
- 仅输出病历中真实存在的内容。

Figure A.8: Prompt template for imaging examination extraction.

Prompt for Auxiliary Examination

You are a professional medical record review expert, skilled at identifying relevant auxiliary examinations from the original medical records.
Please extract information on [all auxiliary examinations and their results] from the original medical record.
[Draw Scope Explanation]
Auxiliary examinations include but are not limited to:
- Cardiovascular-related examinations: ECG, Holter (Holter), echocardiography/echocardiography, myocardial enzyme spectrum, BNP/NT-proBNP, cardiac CT/CTA, coronary angiography, etc
- Respiratory system tests: lung function, blood gas analysis, exhaled nitric oxide (FeNO), etc
- Neuroelectrophysiology: EEG, electromyography/nerve conduction, evoked potentials, etc
- Endoscopy / Endoscopic Ultrasound / Bronchoscope, etc
- Any other "examination/examination/assessment/measurement" appearing in the medical record, along with corresponding results, conclusions, values, or tips (including original statements such as "no abnormalities observed/no actions/pending report")
[Important Rules]
1. Be sure not to omit any auxiliary examination information that appears in the medical record (test name + result/conclusion/value/tip).
2. Strictly extract from the original text, strive to preserve the original expression, and do not summarize, speculate, or fabricate content.
3. If a particular examination is not mentioned in the medical record, the field will not be output.
4. Field names can be flexibly added or removed according to the medical record content, not limited to the fields shown in the example; If there are different or multiple results of similar checks, please list them separately in the original text or keep multiple entries in the original text using semicolons/line breaks.
[Output Format Requirements]
Please output in JSON format. The structure example is as follows. Specific fields can be adjusted according to the actual medical record content:
{
"Auxiliary Examination": {
"Electrocardiogram": ".....",
"EEG": ".....",
"Lung function": "....."
}
}
Please note:
- The above fields are for example only;
- If there are other types of auxiliary examinations in the medical record, please add the corresponding field under "Auxiliary Examinations";
- Only output content that truly exists in the medical records.

你是一位专业病历审查专家，擅长从病历原文中找出相关辅助检查。
请从病历原文中提取【所有辅助检查及其结果】的信息。
【抽取范围说明】
辅助检查包括但不限于：
- 心血管相关检查：心电图、动态心电图（Holter）、心脏彩超/超声心动图、心肌酶谱、BNP/NT-proBNP、心脏CT/CTA、冠脉造影等
- 呼吸系统相关检查：肺功能、血气分析、呼出气一氧化氮（FeNO）等
- 神经电生理：脑电图、肌电图/神经传导、诱发电位等
- 内镜/超声内镜/支气管镜等
- 其他任何在病历中出现的“检查/检验/评估/测定”及对应结果、结论、数值、提示（含“未见异常/未做/待回报”等原文表述）
【重要规则】
1. 务必不要遗漏病历中出现的任何辅助检查相关信息（检查名称 + 结果/结论/数值/提示）。
2. 严格依据原文提取，尽量保持原文表述，不得总结、推测或编造。
3. 若病历中未提及某项检查，则不输出该字段。
4. 字段名称可根据病历内容灵活增减，不限于示例中的字段；如存在同类检查的不同时间/多次结果，请按原文分别列出或用分号/换行保留多条原文。
【输出格式要求】
请以 JSON 格式输出，结构示例如下，具体字段可根据实际病历内容调整：
{
"辅助检查": {
"心电图": "……",
"脑电图": "……",
"肺功能": "……"
}
}
请注意：
- 上述字段仅作为示例；
- 如病历中存在其他类型的辅助检查，请在“辅助检查”下新增对应字段；
- 仅输出病历中真实存在的内容。

Figure A.9: Prompt template for auxiliary examination extraction.

Prompt for Pathological Examination

You are a professional pathology specialist, skilled at identifying relevant pathological tests from original medical records.
Please extract information on [all pathology-related examinations and results] from the original medical record.
[Draw Scope Explanation]
Pathological examinations include but are not limited to:
- Cytology examination
- Immunohistochemical tests
- Tissue biopsy / pathological biopsy
- Frozen slices
- Special dyeing
- Molecular pathology / genetic testing
- Any examination clearly belonging to the category of pathology and its results
[Important Rules]
1. Be sure not to omit any pathological examination information that appears in the medical record.
2. Strictly extract from the original text, strive to preserve the original expression, and do not speculate, summarize, or fabricate content.
3. If no specific examination appears in the medical record, this field will not be output.
4. Field names can be flexibly added or removed according to the medical record content, not limited to the fields shown in the example.
[Output Format Requirements]
Please output in JSON format. The structure example is as follows, and the fields can be adjusted according to the actual content:
{
"Pathological examination": {
"Cytology": ".....",
"Immunohistochemistry": ".....",
"Tissue biopsy": "....."
}
}
Please note:
- The above fields are just examples and do not need to appear in all of them;
- If there are other types of pathological examinations in the medical record, please add the corresponding field under "Pathological Examination";
- Only the information that truly exists in the medical record is output.

你是一位专业病理学检查专家，擅长从病历原文中找出相关病理学检查。
请从病历原文中提取【所有病理学相关检查及结果】的信息。
【抽取范围说明】
病理学检查包括但不限于：
- 细胞学检查
- 免疫组化检查
- 组织活检 / 病理活检
- 冰冻切片
- 特殊染色
- 分子病理 / 基因检测
- 任何明确属于病理学范畴的检查及其结果
【重要规则】
1. 务必不要遗漏病历中出现的任何病理学检查相关信息。
2. 严格依据原文提取，尽量保持原文表述，不得推测、总结或编造。
3. 若病历中未出现某类检查，则不输出该字段。
4. 字段名称可根据病历内容灵活增减，不限于示例中的字段。
【输出格式要求】
请以 JSON 格式输出，结构示例如下，字段可根据实际内容调整：
{
"病理学检查": {
"细胞学": "……",
"免疫组化": "……",
"组织活检": "……"
}
}
请注意：
- 上述字段只是示例，不要求全部出现；
- 如病历中存在其他类型的病理学检查，请在“病理学检查”下新增对应字段；
- 只输出病历中真实存在的信息。

Figure A.10: Prompt template for pathological examination extraction.

Prompt for Diagnostic Name

Please extract the patient's diagnosis name from the medical record, and extract the primary and secondary diagnoses from the discharge record
Please extract the above information from the provided medical records in the following format (no fabrication allowed):
{{
"Diagnosis Name": {{
"First diagnosis": "****",
"Secondary Diagnosis": ["Secondary Diagnosis 1", "Secondary Diagnosis 2", "....."]
}}
}}

请从病历中提取病人诊断名称，请提取出院记录中出院诊断的第一诊断及次要诊断
请按照以下格式，从提供的病历中提取以上信息（所有信息不得编造）：
{{
"诊断名称": {{
"第一诊断": "****",
"次要诊断":["次要诊断1", "次要诊断2","……"]
}}
}}

Figure A.11: Prompt template for diagnosis extraction.

Prompt for Relation Graph

You are a physician reviewing clinical diagnostic evidence. Please assess the relationship between the following medical record evidence and guideline items.
The types of relationships can only be selected: satisfactory, checked, refuted, insufficient information, or irrelevant.
Principles of judgment:
1. If the guideline item is a symptom, sign, or medical history, the medical record must have a corresponding positive result to be considered "satisfactory."
2. If the guideline item is an examination item, the medical record records that the examination has been completed and counts as "completed," without requiring abnormal results.
3. "No abnormalities observed" usually means the inspection item has been completed with normal results, and should be judged as "completed."
4. "None, denial, not seen" usually counts as "counter-evidence" for symptom-related guideline items.
5. Do not infer information not clearly recorded in the medical record.
6. Semantic consistency is sufficient; exact wording is not required. For example, "blood routine" can correspond to "complete blood cell analysis."
7. Each guideline item must output a relationship; If there is no matching evidence, output "insufficient information."
[Current Assessment Module]
[List of Medical Records and Evidence]
[List of Guidelines to Judge]
Please strictly output the following JSON, and the guideline item number must exactly match the input:

```
{{
  "Relationship Chart": [
    {{
      "Guide Item Number": "xxx",
      "Guideline Item Content": "xxx",
      "Relationship": "Satisfied/Completed Check/Counter-evidence/Insufficient Information/Not Relevant",
      "Matching Evidence Number": ["E1", "E2"],
      "Basis for relationship": "..."
    }}
  ]
}}
```

Output requirements:
- The final answer must be a JSON object.
The first character must be {{, and the last character must be }}.
- Do not output any JSON-external text such as "Thinking Process," "Thinking Process," "Analysis," or "JSON below."

你是一名临床诊断依据审核医生，请判断以下病历证据和指南项之间的关系。
关系类型只能选择：满足、已完成检查、反证、信息不足、无关。
判断原则：
1. 如果指南项是症状、体征或病史表现，病历需要有对应的阳性记录才算“满足”。
2. 如果指南项是检查项目，病历记录该检查已完成即可算“已完成检查”，不要求结果异常。
3. “未见异常”对于检查项目通常表示检查已完成且结果正常，应判为“已完成检查”。
4. “无、否认、未见”对于症状类指南项通常算“反证”。
5. 不得推断病历中没有明确记录的信息。
6. 语义一致即可，不要求字面完全一致，如“血常规”可对应“全血细胞分析”。
7. 每个指南项必须输出一条关系；如果没有匹配证据，输出“信息不足”。
【当前判断模块】

【病历证据列表】

【需要判断的指南项列表】

请严格输出如下JSON，指南项编号必须与输入完全一致：

```
{{
  "关系图": [
    {{
      "指南项编号": "xxx",
      "指南项内容": "xxx",
      "关系": "满足/已完成检查/反证/信息不足/无关",
      "匹配证据编号": ["E1", "E2"],
      "关系依据": "..."
    }}
  ]
}}
```

输出要求：
- 最终回答必须是一个JSON对象。
- 第一个字符必须是 {{，最后一个字符必须是 }}。
- 不要输出“思考过程”“Thinking Process”“分析”“下面是JSON”等任何JSON外文本。

Figure A.12: Prompt template for evidence–guideline relation graph construction.