ObGynLongBench: Revealing the Evidence-to-EHR Gap in Longitudinal EHR Decision-Making

Jun Xiang[1*] **Zhijie Bao**[1,2*] **Rong Hu**[4†]
Kaizhou Qin[3] **Wei Chen**[5†] **Zhongyu Wei**[1,2†]
[1]School of Data Science, Fudan University, China
[2]Shanghai Innovation Institute, China
[3]Obstetrics & Gynecology Hospital of Fudan University, China
[4]Department of Obstetrics, Obstetrics & Gynecology Hospital of Fudan University,
Shanghai Key Lab of Reproduction and Development,
Shanghai Key Lab of Female Reproductive Endocrine Related Diseases, Shanghai, China
[5]School of Software Engineering, Huazhong University of Science and Technology, China
jxiang25@m.fudan.edu.cn, zjbao24@m.fudan.edu.cn, hurong@fudan.edu.cn
lemuria_chen@hust.edu.cn, zywei@fudan.edu.cn

Abstract

The application of large language models (LLMs) to personalized medical assistants has garnered growing interest. However, existing medical benchmarks largely rely on static question answering with pre-selected evidence, leaving unclear whether LLMs can make reliable clinical decisions from real longitudinal electronic health records (EHRs). To bridge this gap, we introduce **ObGynLongBench**, a rule-grounded long-context EHR benchmark for obstetric and gynecologic decision-making, comprising 1,500 clinical decision-point cases from 976 real pregnancy EHR histories and traceable rules. Each case is anchored to a patient, a pregnancy-timeline point, and a pre-decision information boundary, enabling Evidence-only, Visit-level EHR, and History-level EHR evaluation. Evaluating 17 LLMs reveals a substantial **Evidence-to-EHR Gap**: models perform well when evidence is directly provided, but accuracy drops when evidence must be extracted from same-day records or full pre-decision EHR histories. Further analyses identify evidence utilization as a key bottleneck: performance decreases with longer EHR contexts and more complex evidence requirements, and earlier failures often predict later failures within the same patient history. Finally, active-search agents perform best among EHR access strategies, highlighting patient-specific evidence utilization as a central challenge for reliable personalized medical assistants. Resources are available at https://github.com/xiangjun2003/ObgynLongbench.

1 Introduction

Large language models (LLMs) have achieved strong performance on medical examinations and static clinical question answering benchmarks such as MedQA, MedMCQA, PubMedQA, and MultiMedQA (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019; Singhal et al., 2023). However, real-world personalized medical assistants must go beyond static QA and reason directly over raw patient EHRs. Obstetrics and gynecology provides a representative setting: clinical decisions are often anchored to gestational age and require integrating longitudinal, multi-source evidence accumulated throughout prenatal care (American College of Obstetricians and Gynecologists, 2017; Peahl and Howell, 2021).

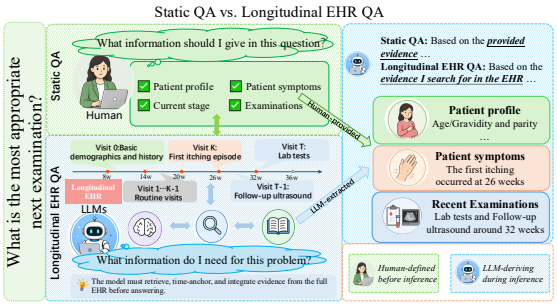

Figure 1: Static QA vs. Longitudinal EHR QA. Static QA uses human-provided decision-relevant evidence, while longitudinal EHR QA requires LLMs to extract patient-specific evidence from the pre-decision EHR history during inference.

Existing LLM evaluations in obstetrics and gynecology mainly rely on specialty examinations, board-style questions, or patient-facing question answering (Bachmann et al., 2024; Ługowski et al., 2025; Srikanth et al., 2024). As summarized in Table 1, these settings largely follow a static QA paradigm analogous to our **Evidence-only** scope, where relevant clinical facts are provided directly in the question, primarily assessing models' medi-

*These authors contributed equally to this work.
†Corresponding authors.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 15764–15788
October 24-29, 2026 ©2026 Association for Computational Linguistics

Work	Size	Data Source	Task Format	Input	Guideline
Obstetrics, gynecology, and maternal-infant QA benchmarks					
Exploring ChatGPT in MRCOG (Bachmann et al., 2024)	1,824	Exam	MCQ	Evidence	✗
Comparative Ob/Gyn Board-style QA (Ługowski et al., 2025)	352	Exam	MCQ	Evidence	✗
Pregnant Questions(Srikanth et al., 2024)	500	Patient QA	Open	Evidence	✗
EHR benchmarks					
MedAlign(Fleming et al., 2024)	983	Hospital	Open	History	✗
MIMIC-Instr(Wu et al., 2024)	400K+	Hospital	Mixed	History	✗
TIMER(Cui et al., 2025)	402 / 248	Hospital	Mixed	History	✗
ObGynLongBench (ours)	1,500	Hospital	MCQ	History	✓

Table 1: Comparison between existing obstetrics and gynecology QA benchmarks, general EHR benchmarks, and ObGynLongBench. In the "Input" column, **Evidence** denotes the **Evidence-only** setting, where evidence is provided directly in the question; **History** denotes the **History-level EHR** setting, where LLMs reason over longitudinal EHR records. "Guideline" indicates grounding in traceable clinical rules from guidelines or textbooks.

cal knowledge and clinical reasoning ability. More broadly, general EHR benchmarks have shown that strong performance on static QA does not readily translate to robust performance on EHR-based clinical tasks (Fleming et al., 2024; Cui et al., 2025).

As illustrated in Figure 1, static QA typically presents evidence in an **Evidence-only** manner directly in the question, thereby primarily assessing models' medical knowledge. In real-world obstetric care, however, the relevant evidence is no longer pre-provided in the question, but scattered across the **History-level EHR**. Before making a decision, the model must first infer what information is needed for the current clinical problem, retrieve the corresponding evidence from heterogeneous records, time-anchor each piece of evidence along the gestational timeline, and integrate findings across visits, examinations, and clinical notes.

To study this problem, we propose **ObGynLongBench**, a rule-grounded long-context EHR benchmark for obstetric decision-making along the pregnancy timeline. The benchmark is built from real longitudinal EHRs of 976 pregnant patients, covering the process from early pregnancy registration to pre-delivery assessment. Based on textbooks and clinical guidelines, we extract 1,500 traceable clinical decision points across 16 obstetric clinical domains and instantiate them on patient-specific gestational histories. Each question is anchored to a concrete decision point and associated with a clinical rule and a gold-standard answer. ObGynLongBench therefore evaluates whether LLMs can locate, time-anchor, and integrate evidence from real longitudinal pregnancy EHRs before making clinical decisions.

ObGynLongBench evaluates LLMs along two controlled axes: evidence complexity and input scope. For evidence complexity, clinical rules are assigned to three levels: **L1** single-evidence recognition, **L2** multi-source integration, and **L3** long-horizon reasoning over temporally distributed evidence. The level of a case is determined by the clinical rule used to construct it. For input scope, each case is tested under three within-case settings: **Evidence-only** (Evidence), **Visit-level EHR** (Visit), and **History-level EHR** (History). Here, Evidence provides pre-localized key evidence, Visit provides same-day records, and History provides the full pre-decision EHR history, enabling us to measure the **Evidence-to-EHR Gap** and the sufficiency of visit-level information.

Experimental results reveal a clear **Evidence-to-EHR Gap**: LLMs perform well under Evidence-only, but accuracy drops substantially when they answer from Visit-level EHR or History-level EHR.

We highlight three main findings. First, the gap is largely driven by **evidence utilization**: LLM accuracy decreases as EHR contexts become longer and evidence requirements become more complex. Second, errors show **patient-level associations**: LLMs that fail at an earlier decision point are more likely to fail again at a later decision point from the same patient history. Third, among different EHR access strategies, a medically motivated **active-search agent** performs best by first identifying the current clinical decision target and then actively gathering relevant patient-specific evidence before answering.

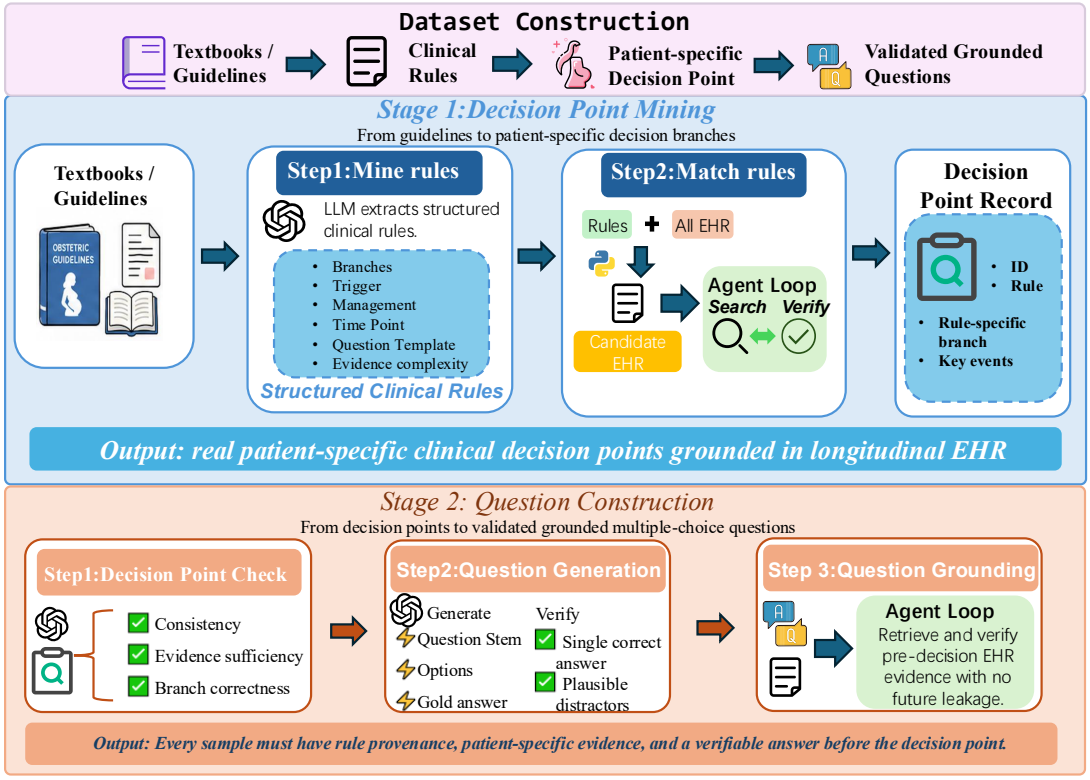

Figure 2: Construction pipeline of ObGynLongBench. Stage 1 derives traceable clinical decision rules from obstetrics and gynecology textbooks and guidelines, structures them into executable rules, and matches them to longitudinal EHR histories through programmatic screening and agent-based review. Stage 2 constructs rule-grounded clinical questions, verifies patient-specific evidence, enforces the pre-decision cutoff, and generates the final benchmark cases.

2 ObGynLongBench

Overall, we organize real pregnancy EHRs into longitudinal patient histories, derive traceable clinical decision rules from textbooks and guidelines, and use these rules to locate corresponding decision points within each patient's EHR history. Each benchmark case is grounded in a real patient, a concrete gestational-timeline point, and a rule-defined clinical decision.

2.1 Task Definition

Each evaluation sample corresponds to a **clinical decision point** in a real patient's longitudinal EHR history. At each decision point, we construct one sample consisting of a patient EHR input and a decision-oriented clinical question:

$$c = (x, q),$$

where x denotes the patient EHR available at the decision point, and q denotes the clinical question to be answered.

For a patient p, the longitudinal EHR history is represented as an ordered sequence of timestamped records:

$$\mathcal{H}_p = \{(t_1, e_1), (t_2, e_2), \ldots, (t_n, e_n)\},$$
$$t_1 < t_2 < \cdots < t_n.$$

Each e_i contains heterogeneous clinical events recorded at time t_i. For a decision point at time t_k, the input EHR is the prefix of the longitudinal EHR history:

$$x = \mathcal{H}_{p,\leq k} = \{(t_1, e_1), (t_2, e_2), \ldots, (t_k, e_k)\}.$$

All records after t_k are removed to avoid future information leakage.

The question q specifies the clinical decision to be made at t_k:

$$q = (m, \mathcal{E}, r),$$

where m denotes the multiple-choice question instance, including the question stem, candidate answers, and the gold-standard answer; $\mathcal{E} \subseteq x$ denotes the EHR evidence supporting the answer; and r denotes the clinical rule from which the decision point is derived. Given only the pre-decision EHR prefix x, the LLM selects the correct option for m based on the evidence contained in x.

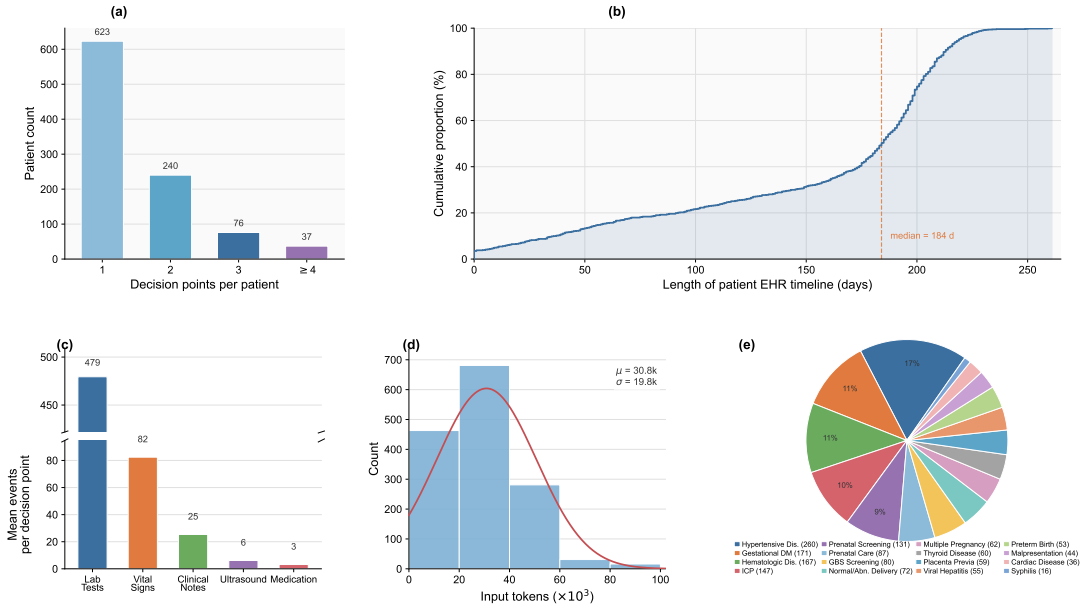

Figure 3: Dataset overview of ObGynLongBench. **(a)** Distribution of decision points per patient. **(b)** Cumulative distribution of patient EHR history duration. **(c)** Mean number of pre-decision clinical events per decision point across five major event types. **(d)** Distribution of pre-decision input length under the **History-level EHR** setting, measured by the Qwen3.5 tokenizer. **(e)** Case distribution across 16 obstetric and gynecologic subspecialties.

2.2 Data Construction

ObGynLongBench is built from real pregnancy EHR histories and traceable obstetric and gynecologic clinical rules. The EHRs come from a single-center tertiary hospital cohort, covering the pregnancy process from early registration to pre-delivery assessment. Clinical rules are derived from textbooks and guidelines, define decision-point types, and guide question construction and evidence grounding.

Figure 2 summarizes the construction pipeline. Inspired by MIMIC-Instr (Wu et al., 2024), we align heterogeneous EHR events onto a gestational-age timeline, derive structured clinical rules, match them to longitudinal EHR histories through programmatic screening and agent-based review, and construct grounded decision-point questions. The complete construction pipeline and independent clinician audit are presented in Appendix A.

2.3 Dataset

Figure 3 summarizes the scale and composition of the underlying EHR cohort and the resulting benchmark. The dataset is built from 976 real pregnancy EHR histories and instantiated into 1,500 rule-grounded decision-point cases across 16 obstetric and gynecologic subspecialties.

Most patients contribute one decision point, while 240 patients contribute two and smaller groups contribute three or more. The longitudinal EHR histories span a median of 184 days from early registration to pre-delivery assessment. At each decision point, the available pre-decision EHR contains rich heterogeneous clinical events, with an average of 479 laboratory tests, 82 vital-sign measurements, 25 clinical notes, 6 ultrasound reports, and 3 medication records. Under the **History-level EHR** setting, the mean pre-decision input length is 30.8K tokens ($\sigma = 19.8$K). The cases cover 16 obstetric and gynecologic subspecialties, including hypertensive disorders, gestational diabetes, hematologic disorders, ICP, prenatal screening, and other pregnancy-related conditions.

3 Experiments

This section builds on the benchmark and evaluation protocol defined in Section 2. We first describe the experimental setup and then report the main results under the three input-scope settings.

3.1 Experimental Setup

Information Scope. For each case $c = (x, q)$ with $q = (m, \mathcal{E}, r)$, we evaluate LLMs under three input-scope settings at the same decision point t_k: **Evidence-only** (**Evidence**) provides only the supporting evidence $\mathcal{E} \subseteq x$; **Visit-level EHR** (**Visit**) provides all EHR records recorded on the same day as the decision point t_k; and **History-level EHR**

Model		Evidence-only	History-level EHR				Visit-level EHR
Name	Context window	Overall	Overall	L1	L2	L3	Overall
Commercial LLMs (Flash Level)							
Claude-Haiku-4.5	256k	75.5	59.8(-15.7)	63.0	61.0	52.3	58.5(+1.3)
GPT-5.4-mini	400k	77.4	64.9(-12.5)	61.8	68.0	60.7	62.3(+2.6)
Gemini-3-Flash	1000k	79.1	68.7(-10.4)	66.0	71.5	65.0	67.1(+1.6)
DeepSeek-V4-Flash	1000k	78.8	64.4(-14.4)	66.5	66.5	56.0	61.2(+3.2)
Long-context open-source LLMs							
Qwen3-VL-4B	262k	72.1	59.8(-12.3)	61.8	58.2	61.3	54.7(+5.1)
Qwen3-VL-8B	262k	74.3	61.8(-12.5)	64.8	62.4	56.3	58.4(+3.4)
Qwen3.5-4B	262k	75.4	59.4(-16.0)	63.2	57.5	59.3	52.7(+6.7)
Qwen3.5-9B	262k	73.1	59.3(-13.8)	62.7	60.1	52.3	55.4(+3.9)
Qwen3.5-35B-A3B	262k	76.9	60.1(-16.8)	58.0	62.1	57.7	55.7(+4.4)
Gemma-4-E4B	131k	68.1	52.0(-16.1)	56.0	53.2	43.3	47.4(+4.6)
Gemma-4-26B-A4B	262k	78.2	60.7(-17.5)	65.0	61.4	53.0	60.6(+0.1)
Phi-4-mini	131k	58.7	51.1(-7.6)	56.2	47.8	53.3	48.6(+2.5)
Long-context Medical LLMs							
HealthGPT-Pro-4B	262k	69.6	57.1(-12.5)	59.0	56.1	57.3	54.1(+3.0)
HealthGPT-Pro-8B	262k	73.5	61.3(-12.2)	67.2	59.8	57.3	58.1(+3.2)
Hulu-Med-30A3	75k	74.9	53.9(-21.0)	59.2	53.0	49.0	55.8(-1.9)
Lingshu-7B	128k	72.1	59.6(-12.5)	61.0	59.5	58.0	60.4(-0.8)
MedGemma-1.5-4B	131k	64.0	41.3(-22.7)	45.8	43.5	29.7	48.1(-6.8)

Table 2: Main results on ObGynLongBench. We report accuracy (%) for each LLM under three input-scope settings: **Evidence-only**, **History-level EHR**, and **Visit-level EHR**. Parenthesized values in the **Overall** column of **History-level EHR** denote the accuracy gap relative to **Evidence-only** ($\Delta = History - Evidence$), and those in the **Overall** column of **Visit-level EHR** denote the gap relative to **History-level EHR** ($\Delta = History - Visit$); positive gaps are shown in green as $(+\Delta)$ and negative gaps in red as $(-\Delta)$. For open-source and medical LLMs, thinking mode is disabled for fair comparison because it often causes context-window overflow.

(**History**) provides the full pre-decision EHR history $x = \mathcal{H}_{p,\leq k}$. The LLM is required to answer m from the provided input under each setting.

Models. We evaluate 17 LLMs from three groups that are practically relevant to personalized medical assistant scenarios: **Commercial LLMs**, which provide fast and accessible inference, including Claude Haiku 4.5 (Anthropic, 2025), GPT-5.4-mini (OpenAI, 2026b), Gemini-3-Flash (Doshi, 2025), and DeepSeek-V4-Flash (DeepSeek-AI et al., 2026); **Long-context open-source LLMs**, including Qwen3-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026), Gemma 4 (Google DeepMind, 2026), and the Phi-4 series (Abouelenin et al., 2025); and **Long-context medical LLMs**, including HealthGPT-Pro (Lin et al., 2025), Hulu-Med (Jiang et al., 2025), Lingshu (LASA Team et al., 2025), and MedGemma-1.5-4B (Sellergren et al., 2026).

Evaluation metric. All experiments use a multiple-choice question (MCQ) format, where each LLM outputs a single answer option. Given a test set $\mathcal{D}$, accuracy is defined as:

$$\mathrm{Acc} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbb{I}(\hat{y}_i = y_i), \tag{1}$$

where $\hat{y}_i$ is the LLM-predicted answer and y_i is the gold-standard answer.

3.2 Main Results

Table 2 reports the main results of the three LLM groups under three input-scope settings: **Evidence**, **Visit**, and **History**, corresponding to **Evidence-only**, **Visit-level EHR**, and **History-level EHR**, respectively. Patient-clustered bootstrap confidence intervals are reported in Appendix C.1.

The gap between Evidence-only and History-level EHR is large. Under **Evidence**, the best-performing LLMs in the three groups are Gemini-3-Flash (79.1%), Gemma-4-26B-A4B (78.2%), and Hulu-Med-30A3 (74.9%). Under **History**, the best scores drop to Gemini-3-Flash (68.7%), Qwen3-VL-8B (61.8%), and HealthGPT-Pro-8B (61.3%). This shows that **Evidence-only** performance does not directly transfer to **History-level EHR**.

Medical LLMs show a larger gap under History-level EHR. Medical LLMs do not show a clear advantage in the **History** setting. The gap between the best general-purpose LLM and the best medical LLM increases from 4.2pp under **Evidence** to 7.4pp under **History**. This suggests that existing medical LLM training does not sufficiently prepare LLMs to extract decision-relevant information from realistic EHR histories.

Visit is incomplete; History can introduce additional noise. Under **Visit**, LLM performance

drops substantially compared with **Evidence**, indicating that same-day records often lack the longitudinal context needed for obstetric and gynecologic decision-making. For most LLMs, **History** improves over **Visit**, showing that longitudinal EHR history provides useful clinical context. However, for LLMs with context windows shorter than 256K, **History** performs worse than **Visit**. This indicates that full history-level EHR input can also introduce additional noise, especially when LLMs cannot reliably process long EHRs.

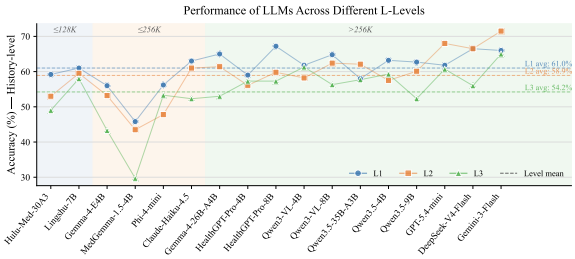

Figure 4: LLM performance across evidence-complexity levels under the **History-level EHR** setting. Accuracy decreases from L1 to L3 as evidence utilization and integration become more difficult.

4 Further Analysis

Section 3.2 establishes the **Evidence-to-EHR Gap** under different input scopes. This section provides additional analyses under the **History-level EHR** setting, examining how evidence complexity, EHR input length, and patient-level history affect LLM performance.

4.1 Evidence Complexity Analysis

Under the **History** setting, we further analyze how LLM performance changes with evidence complexity. Clinical rules are assigned to three evidence-complexity levels, and cases are grouped by the rule used to construct them: **L1 (single-evidence recognition)** depends on one explicit and locally visible key evidence item; **L2 (multi-source integration)** requires combining multiple evidence items or modalities within the same visit or local context; and **L3 (long-horizon reasoning)** requires retrieving, time-anchoring, and integrating evidence distributed across the longitudinal EHR history. Full stratification criteria are provided in Appendix B.

Figure 4 shows a consistent ordering under the **History** setting: accuracy is highest on L1, lower on L2, and lowest on L3. This indicates that the difficulty of locating and integrating the supporting evidence $\mathcal{E} \subseteq x$ from the pre-decision EHR history directly affects LLM performance under the **History** setting.

4.2 Context Length Analysis

We further examine whether longer EHR histories make the **History** setting more difficult. Specifically, we group cases by the length of the pre-decision EHR input and analyze both final accuracy and evidence utilization.

Longer EHR histories lead to lower accuracy. Figure 5(a) reports LLM accuracy under the **History** setting across four input-length bins. Accuracy consistently decreases as the pre-decision EHR input becomes longer.

The decline is accompanied by weaker evidence utilization. To understand why longer EHR inputs reduce accuracy, we use DeepSeek-V4-Flash as the primary LLM judge to assess evidence utilization by comparing each LLM's rationale with the gold supporting evidence $\mathcal{E}$ (Appendix D). Figure 5(b) shows that evidence utilization also decreases as input length increases across six representative LLMs. This suggests that the accuracy drop is partly driven by weaker evidence utilization: as the EHR history becomes longer, LLMs are less able to identify and use decision-relevant evidence, leading to less reliable clinical decisions.

4.3 Patient-level Analysis

We further analyze 240 patients who each have two consecutive decision points in the same pregnancy EHR history. For each patient, we denote the two decision points as DP1 and DP2. To control for gestational timing, we compare DP2 with a temporally matched control group: DP1 cases after 30 gestational weeks from patients with only one decision point. Table 3 reports four quantities under **History-level EHR**: overall DP2 accuracy, accuracy on the temporally matched DP1 control group, DP2 accuracy when DP1 is answered incorrectly, and DP2 accuracy when DP1 is answered correctly. Values in parentheses denote the change relative to overall DP2 accuracy.

At the aggregate level, LLM performance on DP2 is close to that on the temporally matched DP1 control group, suggesting that DP2 has comparable difficulty to other decision points from the same pregnancy period. However, a clear patient-level pattern emerges when we condition on whether the

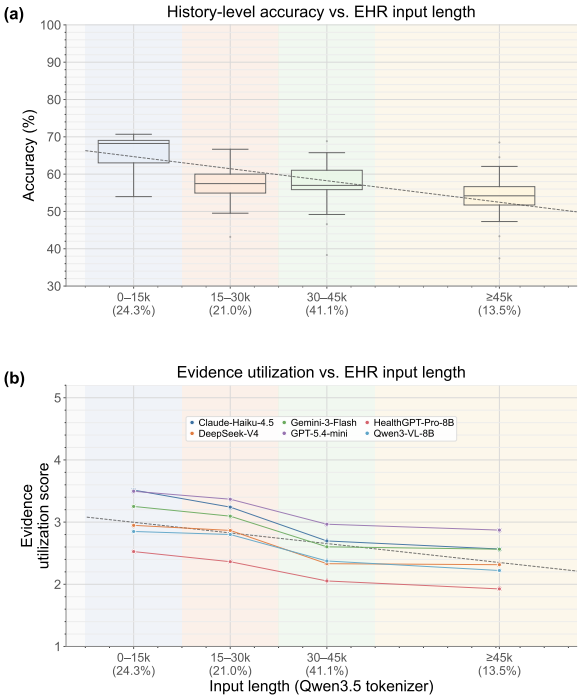

Figure 5: Context-length analysis under **History-level EHR**. **(a)** Accuracy across pre-decision EHR input-length bins, with boxes computed over 17 LLMs. **(b)** LLM-judged evidence utilization across the same bins for six representative LLMs.

earlier decision point from the same patient is answered correctly. For most LLMs, DP2 accuracy decreases when DP1 is wrong and increases when DP1 is correct. This indicates that LLM errors are patient-dependent: if an LLM fails to interpret one decision point in a patient's EHR history, it is more likely to fail again on a later decision point from the same history. In realistic personalized medical assistant settings, LLMs would often be used repeatedly across multiple decision points throughout the same patient's longitudinal care. Appendix C.2 further examines patient-level performance using all 1,500 cases and all 17 models.

Model	Acc@DP2 (All)	Acc@DP1 ($GA \geq 30$w)	Acc@DP2 (DP1 wrong)	Acc@DP2 (DP1 correct)
Gemini-3-Flash	67.1	69.3(+2.2)	66.7(-0.4)	67.3(+0.2)
GPT-5.4-mini	65.4	63.3(-2.1)	64.4(-1.0)	65.9(+0.5)
DeepSeek-V4-Flash	61.7	64.0(+2.3)	59.0(-2.7)	63.0(+1.3)
HealthGPT-Pro-8B	57.5	59.5(+2.0)	53.3(-4.2)	60.1(+2.6)
Claude-Haiku-4.5	57.1	57.1(0.0)	48.8(-8.3)	61.4(+4.3)
Qwen3-VL-8B	57.1	57.6(+0.5)	51.2(-5.9)	60.3(+3.2)

Table 3: Patient-level analysis under **History-level EHR**. We report accuracy (%). Values in parentheses denote changes relative to overall DP2 accuracy.

5 EHR Access Strategies under History-level EHR

Section 4.2 shows that direct reading of the full pre-decision EHR history cannot recover the performance achieved under the **Evidence** setting. We therefore compare alternative EHR access strategies under the **History** input scope. For each case, we keep the same full pre-decision EHR history fixed and vary only how LLMs access, organize, and compress the records. Our goal is to identify strategies that improve accuracy while controlling token cost in long-context medical record scenarios.

5.1 Strategy Settings

We use **Direct** access to the **History-level EHR** input as the baseline. We compare five alternative EHR access strategies against this baseline under the same full pre-decision EHR history, varying only how LLMs represent, retrieve, and interact with the records.

Image. We render the full pre-decision EHR history into multiple page-level images and provide them to LLMs.

Static RAG. We split the EHR history into date-based chunks, retrieve relevant snippets according to the current question in a single step, and provide them to LLMs.

Static RAG + Recency (Static RAG + Rec.). We add recency weighting to Static RAG, assigning higher retrieval weights to records closer to the decision point.

Rolling Summary. We input the EHR history every 7 days in chronological order and ask LLMs to generate stage-wise summaries. Each subsequent round contains the previous summary and the new 7-day records.

Agent. The agent first reads the clinical question to identify the decision target, then obtains basic patient information to anchor the case in the EHR history. It then actively searches for decision-relevant evidence through an iterative retrieval loop, rather than relying on a single-pass reading of the full **History-level EHR** input.

Implementation details for the RAG and Agent settings are provided in Appendix F.

5.2 Strategy Comparison

Table 4 compares the accuracy and average token cost of each EHR access strategy, averaged over three Qwen3.5-series LLMs under the **History-level EHR** setting. Values in parentheses indicate changes relative to the **Direct** baseline. Paired-bootstrap uncertainty estimates for the overall accuracy differences relative to **Direct** are provided in Appendix C.3. For **Rolling Summary**, token cost is computed cumulatively over the full rolling summarization process, rather than using only the input length of the final round.

Strategy	Overall	L1	L2	L3	Tok.
Direct	59.6	61.3	59.9	56.4	30,951
Image	52.7(-6.9)	55.1(-6.2)	52.8(-7.1)	49.0(-7.4)	11,433
Static RAG	58.9(-0.7)	58.8(-2.5)	59.7(-0.2)	56.8(+0.4)	4,308
Static RAG + Rec.	59.2(-0.4)	60.6(-0.7)	60.1(+0.2)	54.5(-1.9)	4,089
Rolling Summary	57.9(-1.7)	62.4(+1.1)	58.1(-1.8)	51.7(-4.8)	61,968
Agent	64.2(+4.6)	63.8(+2.5)	65.1(+5.2)	62.1(+5.7)	35,786

Table 4: Accuracy (%) and average token cost of different EHR access strategies under the **History-level EHR** setting, averaged over three Qwen3.5-series LLMs. Values in parentheses denote changes relative to the **Direct** baseline.

Agent. **Agent** achieves the best overall accuracy (64.2%, +4.6pp) and improves over **Direct** access across all evidence complexity levels, with the largest gain on L3 (+5.7pp). This suggests that history-level EHR reasoning benefits from question-guided active search: LLMs should first identify the current clinical decision target and then actively locate decision-relevant evidence, rather than simply shorten or read the full input in a single pass.

Static RAG and Static RAG + Rec. Both retrieval-based strategies substantially reduce token cost while maintaining accuracy close to **Direct** access. **Static RAG** performs slightly better on L2/L3, whereas **Static RAG + Rec.** improves L1 and L2 but drops on L3. This indicates a recency bias: emphasizing recent records helps locally anchored decisions, but may miss earlier evidence needed for long-horizon reasoning.

Image. **Image** reduces token count but degrades performance at all complexity levels, especially on L3 (−7.4pp). This suggests that page-level rendering introduces visual parsing noise and hinders cross-page evidence integration.

Rolling Summary. **Rolling Summary** incurs the highest token cost and does not surpass **Direct** overall (−1.7pp). Its drop on L3 (−4.8pp) suggests that sequential summarization may lose cross-time evidence needed for harder questions.

6 Related Work

Medical and obstetric QA benchmarks. Medical LLM evaluation commonly relies on examinations and benchmarks, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MultiMedQA (Singhal et al., 2023), and MedRCube (Bao et al., 2026). In static QA benchmarks, relevant clinical evidence is usually provided in the question. Obstetric and gynecologic evaluations similarly focus on specialty examinations, board-style questions, or patient-facing maternal health QA (Bachmann et al., 2024; Ługowski et al., 2025; Srikanth et al., 2024), evaluating domain knowledge and response quality rather than reasoning over real longitudinal pregnancy EHR histories.

EHR benchmarks. Recent EHR benchmarks move closer to real clinical records. MedAlign evaluates clinician-generated instructions grounded in longitudinal EHRs (Fleming et al., 2024); MIMIC-Instr constructs large-scale EHR-grounded instruction-following data from MIMIC-IV (Wu et al., 2024); and TIMER focuses on temporal reasoning over longitudinal patient histories (Cui et al., 2025). These benchmarks reveal that EHR-based reasoning is more difficult than static medical QA. However, they mainly target general EHR tasks such as retrieval, summarization, instruction following, or temporal reasoning. In contrast, ObGynLongBench focuses on rule-grounded obstetric decision points along the pregnancy EHR history, requiring LLMs to retrieve, time-anchor, and integrate patient-specific evidence before making clinical decisions.

7 Conclusion

We introduced **ObGynLongBench**, a rule-grounded long-context EHR benchmark for obstetric and gynecologic decision-making. Experiments reveal a clear **Evidence-to-EHR Gap**: LLMs perform well with directly provided evidence, but accuracy drops under **Visit-level EHR** and **History-level EHR**. Further analyses show that this gap is driven by evidence utilization and worsens with longer and more complex EHR evidence; model errors are also associated across longitudinal decision points within patients. Active-search agents

perform best among EHR access strategies, highlighting patient-specific evidence utilization as a key requirement for reliable personalized medical assistants.

Limitations

This paper has the following limitations:

- **Single-center EHR cohort.** ObGynLongBench is built from real pregnancy EHRs from a single tertiary hospital, and may reflect institution-specific documentation habits, clinical workflows, and regional population characteristics. Future work should validate the benchmark across hospitals and regions.
- **Chinese-language clinical records.** The EHRs are written in Chinese, so performance may be affected by Chinese clinical language understanding, abbreviation handling, and multilingual robustness. The conclusions may not directly generalize to EHRs in other languages.
- **MCQ-based evaluation.** We use MCQs for controlled automatic evaluation, but real clinical assistance often requires open-ended reasoning, structured recommendations, uncertainty expression, and interactive clarification. Future benchmarks should extend to more realistic clinical output formats.
- **Inference-setting differences.** As detailed in Appendix F, open-source and medical LLMs are evaluated with thinking mode disabled, whereas commercial LLMs use their APIs' default inference configurations. These differences may affect absolute performance comparisons across model families. However, the **Evidence-to-EHR Gap** is measured within each model under a fixed configuration and therefore does not require identical inference settings across families.

Ethical Considerations

This work uses real-world pregnancy EHRs, which contain sensitive health information. All data processing and benchmark construction was conducted under institutional ethics approval and appropriate data-use agreements. Patient records were used only in de-identified form, and the benchmark is intended solely for research evaluation. Researchers should not attempt patient re-identification, linkage with external data sources, or use the benchmark to infer private attributes. Consent requirements were handled according to the approved institutional protocol.

ObGynLongBench is not a clinical decision-support system. Model outputs may be incomplete, poorly grounded, or clinically unsafe, even when benchmark accuracy is high. Any real clinical use of LLMs would require careful validation, clinician oversight, patient consent where appropriate, and compliance with local medical, legal, and regulatory requirements. The goal of this benchmark is to support safer research on long-context EHR reasoning, not to encourage autonomous medical deployment.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2023YFF1204800) and the AI for Science Program of the Shanghai Municipal Commission of Economy and Informatization (Grant No. 2025-GZL-RGZN-BTBX-02028). This work was also supported by the National Natural Science Foundation of China (Grant No. 62406121) and the Shanghai Basic Research Program (Natural Science Foundation) (Grant No. 25ZR1401034). Computational resources for this project were partially provided by the CFFF platform at Fudan University.

AI writing assistants were used only for language polishing and editing. All scientific claims, experimental results, analyses, and final manuscript content were reviewed and verified by the authors.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, H. Awadalla, Nguyen Bach, Jianmin Bao, A. Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dongdong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 54 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

American College of Obstetricians and Gynecologists. 2017. Committee opinion no 700: Methods for estimating the due date. *Obstetrics & Gynecology*, 129(5):e150–e154.

American College of Obstetricians and Gynecologists. 2020a. Gestational hypertension and preeclampsia: Acog practice bulletin, number 222. *Obstetrics & Gynecology*, 135(6):e237–e260.

American College of Obstetricians and Gynecologists. 2020b. Prevention of group b streptococcal early-onset disease in newborns: Acog committee opinion, number 797. *Obstetrics & Gynecology*, 135(2):e51–e72.

American College of Obstetricians and Gynecologists' Committee on Practice Bulletins–Obstetrics, Committee on Genetics, and Society for Maternal-Fetal Medicine. 2020. Screening for fetal chromosomal abnormalities: Acog practice bulletin, number 226. *Obstetrics & Gynecology*, 136(4):e48–e69.

Anthropic. 2025. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5. Accessed: 2026-05-17.

Anthropic. 2026a. Introducing claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6. Accessed: 2026-05-25.

Anthropic. 2026b. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-08-27.

Magdalena Bachmann, Ioana Duta, Emily Mazey, William Cooke, Manu Vatish, and Gabriel Davis Jones. 2024. Exploring the capabilities of chatgpt in women's health: obstetrics and gynaecology. *npj Women's Health*, 2:26.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Zhijie Bao, Fangke Chen, Licheng Bao, Chenhui Zhang, Wei Chen, Jiajie Peng, and Zhongyu Wei. 2026. Medrcube: A multidimensional framework for fine-grained and in-depth evaluation of mllms in medical imaging. *Preprint*, arXiv:2604.13756.

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

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *Preprint*, arXiv:2606.19348.

Tulsee Doshi. 2025. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. Google Blog. Accessed: 2026-05-17.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, and 11 others. 2024. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030.

Google DeepMind. 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-17.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, and 6 others. 2025. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *Preprint*, arXiv:2510.08668.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *Preprint*, arXiv:2506.07044.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song Xiaohui, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. 2025. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37975–37995. PMLR.

OpenAI. 2026a. Introducing gpt-5.3-codex. https://openai.com/index/introducing-gpt-5-3-codex/. Accessed: 2026-05-25.

OpenAI. 2026b. Introducing gpt-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Accessed: 2026-05-17.

OpenAI. 2026c. Introducing gpt-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-08-27.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Alex F. Peahl and Joel D. Howell. 2021. The evolution of prenatal care delivery guidelines in the united states. *American Journal of Obstetrics and Gynecology*, 224(4):339–347.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. Medgemma 1.5 technical report. *Preprint*, arXiv:2604.05081.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Society for Maternal-Fetal Medicine (SMFM), Richard H. Lee, Mara Greenberg, Torri D. Metz, and Christian M. Pettker. 2021. Society for maternal-fetal medicine consult series #53: Intrahepatic cholestasis of pregnancy: Replaces consult #13, april 2011. *American Journal of Obstetrics and Gynecology*, 224(2):B2–B9.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.

Zhenbang Wu, Anant Dadu, Mike Nalls, Faraz Faghri, and Jimeng Sun. 2024. Instruction tuning large language models to understand electronic health records. In *Advances in Neural Information Processing Systems*, volume 37, pages 54772–54786. Curran Associates, Inc.

xAI. 2026. Introducing grok 4.5. https://x.ai/news/grok-4-5. Accessed: 2026-08-27.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Franciszek Ługowski, Julia Babińska, Artur Ludwin, and Paweł Jan Stanirowski. 2025. Comparative analysis of chatgpt 3.5 and chatgpt 4 obstetric and gynecological knowledge. *Scientific Reports*, 15(1):21133.

A Data Construction

This appendix presents the complete construction pipeline summarized in Section 2.2 and Figure 2, together with an independent clinician audit of the benchmark questions and gold answers. Specifically, Appendix A.1 describes the EHR and clinical rule sources, Appendix A.2 details the case construction pipeline, and Appendix A.3 reports the independent clinician audit.

A.1 Data and Clinical Rule Sources

Our benchmark is built from 976 real pregnancy EHR histories. For each patient p, the longitudinal EHR history $\mathcal{H}_p$ comes from a real obstetric and gynecologic clinical cohort at a single-center tertiary hospital, covering the pregnancy process from early pregnancy registration to pre-delivery assessment. Each timestamped record e_i contains heterogeneous clinical events, including diagnoses, laboratory tests, vital signs, ultrasound reports, clinical notes, medication records, and screening items.

Based on these EHR histories, we construct clinical decision-point samples $c = (x, q)$. For a decision point at time t_k, the EHR input $x = \mathcal{H}_{p,\leq k}$ consists only of the patient records available up to that point, while all later records are excluded to prevent future information leakage. The average length of the full pre-decision EHR input x is approximately 30K tokens when tokenized by Qwen3.5, making the benchmark challenging for evidence utilization over long contexts and longitudinal information integration.

Clinical rules r are derived from textbooks and clinical guidelines. Each rule r defines a type of clinical decision point and provides the basis for constructing the question $q = (m, \mathcal{E}, r)$, where m is the multiple-choice question instance and $\mathcal{E} \subseteq x$ is the supporting evidence extracted from the available EHR input. In this way, the rule source determines the clinical decision to be evaluated, while the patient EHR history determines what evidence is available at the corresponding decision point.

A.2 Case Construction Pipeline

The construction pipeline consists of two stages: decision-point mining and question construction with grounding. Starting from textbooks, clinical guidelines, and longitudinal EHR histories, the pipeline first mines specialty-specific clinical rules, matches them to patient EHR histories, and then performs multi-stage review, question drafting, grounding verification, and artifact generation.

Inspired by MIMIC-Instr (Wu et al., 2024), which constructs EHR-grounded instructions from structured EHR events using programmatic scripts and LLM rewriting, we adapt this paradigm to rule-grounded obstetric and gynecologic clinical decision-making. Different from general EHR instruction generation, our construction pipeline is centered on clinically meaningful decision points. It introduces LLM-based rule mining from specialty knowledge sources, programmatic high-recall candidate screening, and multi-stage agent-based gating to ensure that each case is medically valid, temporally grounded, and answerable from pre-decision EHR evidence.

We construct evaluation cases through four steps.

1. **Longitudinal EHR history construction.** We first align raw obstetric and gynecologic EHRs from different modalities and event sources onto a unified gestational-age timeline. Each patient's records are converted into a patient-centered longitudinal EHR history ordered by time. This process integrates structured events, such as laboratory tests, vital signs, diagnoses, prescriptions, and screening items, with semi-structured or unstructured records, such as ultrasound reports and clinical notes.

2. **Clinical rule mining.** We use GPT-5.3-Codex (OpenAI, 2026a) to extract traceable clinical decision rules from obstetrics and gynecology textbooks and clinical guidelines. Each rule specifies the triggering context, applicable gestational-age or temporal window, required evidence fields, evidence-complexity level, management action, candidate branches, question template, and supporting medical background. We further use Claude Opus 4.6 (Anthropic, 2026a) to assign each rule to one of three evidence-complexity levels: L1, L2, or L3.

3. **Programmatic candidate screening with agent-based gating.** For each clinical rule, we implement a Python screening program to identify patient-rule candidates from the longitudinal EHR history. This programmatic step serves as a high-recall initial filter. We then use a GPT-5.3-Codex-based agent (OpenAI, 2026a) to review each candidate, check whether the predicted branch is consistent

with the patient's pre-decision EHR evidence, and determine whether the case corresponds to a valid clinical decision point.

4. **Question construction and multi-stage verification.** For each accepted decision point, an LLM drafts an examination-style clinical question and candidate options based on the matched clinical-rule branch, which determines the gold-answer direction. We then use GPT-5.4-mini-based verification agents (OpenAI, 2026b) to check three dimensions: whether the case satisfies the intended clinical rule, whether it forms a valid current-time reasoning question, and whether it is answerable from information available before the decision point. The grounding agent further verifies the supporting evidence, confirms the pre-decision cutoff, collects evidence targets, and records grounding information for constructing the **Evidence-only**, **Visit-level EHR**, and **History-level EHR** input scopes.

The construction pipeline is designed to produce **verifiable clinical decision-point cases**, each grounded in a real patient EHR history, a traceable clinical rule, a pre-decision information boundary, and explicit supporting evidence. This design enables ObGynLongBench to evaluate both rule application under the **Evidence-only** setting and evidence utilization from full pre-decision EHR histories. For each instance, the gold-answer direction is fixed by the matched clinical-rule branch and verified against the patient's pre-decision EHR. The benchmark answers and evaluation scores are therefore objectively defined. Evaluation uses fixed-choice MCQ accuracy, avoiding the potential circularity and model-family bias associated with open-ended generation or LLM-based subjective evaluation. The independent clinician audit in Appendix A.3 provides additional validation of the benchmark questions and gold answers.

A.3 Clinician Validation

To assess the reliability of the benchmark, we conducted an independent clinician audit of 100 instances randomly sampled from the 1,500-instance benchmark. Three obstetrics and gynecology resident physicians with clinical experience participated in the audit, none of whom was involved in benchmark construction. Each clinician independently reviewed the same 100 instances against the complete EHR history available up to the corresponding pre-decision cutoff.

Clinician / statistic	Gold-answer correctness	Question quality
Clinician 1	95/2/3	97/2/1
Clinician 2	93/2/5	97/3/0
Clinician 3	98/1/1	100/0/0
Aggregate result	**96/1/3**	**98/2/0**
At least two clinicians agreed	99%	100%
All three clinicians agreed	90%	96%
Gwet's AC1	0.927	0.973

Table 5: Results of the independent clinician audit on 100 randomly sampled benchmark instances. For gold-answer correctness, values are reported in the order Correct/Incorrect/Uncertain. For question clarity and clinical meaningfulness, values are reported in the order Clear and meaningful/Ambiguous/Unreasonable. Clinician-level and aggregate results are counts out of 100.

For each instance, the clinicians were provided with the question and answer options, the designated gold answer, and the complete pre-decision EHR history. They independently assessed two dimensions using predefined three-category rubrics: (1) gold-answer correctness, categorized as correct, incorrect, or uncertain; and (2) question clarity and clinical meaningfulness, categorized as clear and meaningful, ambiguous and requiring revision, or unreasonable.

For gold-answer correctness, majority voting was applied to the 99 instances for which at least two clinicians agreed. The single instance with three-way disagreement was conservatively categorized as uncertain. Overall, 96 gold answers were classified as correct, one as incorrect, and three as uncertain, corresponding to an observed clearly incorrect rate of 1% in the audit sample. Based on this clinician audit, we estimate the benchmark-wide rate of clearly incorrect gold answers to be below 5%. For question quality, 98 instances were classified as clear and clinically meaningful, two as ambiguous and requiring revision, and none as unreasonable.

Inter-rater agreement was high for both dimensions. For gold-answer correctness, at least two clinicians agreed on 99% of the instances, all three agreed on 90%, and Gwet's AC1 was 0.927. For question clarity and clinical meaningfulness, the corresponding results were 100%, 96%, and 0.973. We report Gwet's AC1 because the annotations were highly concentrated in one category, under

Model	Context window	Evidence-only Overall	History-level EHR Overall	Visit-level EHR Overall
Commercial LLMs (Flash Level)				
Claude-Haiku-4.5	256k	75.5±2.2	59.8±2.6	58.5±2.5
GPT-5.4-mini	400k	77.4±2.2	64.9±2.4	62.3±2.4
Gemini-3-Flash	1000k	79.1±2.1	68.7±2.4	67.1±2.4
DeepSeek-V4-Flash	1000k	78.8±2.1	64.4±2.5	61.2±2.5
Long-context open-source LLMs				
Qwen3-VL-4B	262k	72.1±2.3	59.8±2.5	54.7±2.6
Qwen3-VL-8B	262k	74.3±2.3	61.8±2.5	58.4±2.5
Qwen3.5-4B	262k	75.4±2.2	59.4±2.6	52.7±2.6
Qwen3.5-9B	262k	73.1±2.2	59.3±2.6	55.4±2.5
Qwen3.5-35B-A3B	262k	76.9±2.3	60.1±2.5	55.7±2.5
Gemma-4-E4B	131k	68.1±2.4	52.0±2.6	47.4±2.5
Gemma-4-26B-A4B	262k	78.2±2.2	60.7±2.5	60.6±2.5
Phi-4-mini	131k	58.7±2.6	51.1±2.5	48.6±2.5
Long-context Medical LLMs				
HealthGPT-Pro-4B	262k	69.6±2.3	57.1±2.5	54.1±2.6
HealthGPT-Pro-8B	262k	73.5±2.3	61.3±2.6	58.1±2.5
Hulu-Med-30A3	75k	74.9±2.2	53.9±2.6	55.8±2.6
Lingshu-7B	128k	72.1±2.4	59.6±2.7	60.4±2.5
MedGemma-1.5-4B	131k	64.0±2.4	41.3±2.6	48.1±2.5

Table 6: Patient-clustered bootstrap uncertainty for overall accuracy under the three input-scope settings. Each entry is reported as $\hat{A} \pm m$ in percentage points, where m is the half-width of a 95% patient-clustered percentile bootstrap confidence interval (Eqs. (3)–(5)).

which conventional kappa statistics can be affected by the prevalence paradox. Overall, the audit provides independent clinician-based evidence supporting the reliability of the benchmark questions and gold answers.

B Difficulty Stratification and Metrics

To analyze LLM performance under different levels of evidence complexity, we assign clinical rules to three mutually exclusive levels. Cases are then grouped by the rule used to construct them.

L1: Single-evidence recognition. L1 rules depend on one explicit and locally visible key evidence item, such as a single abnormal indicator, one positive screening result, or a clear diagnosis.

L2: Multi-source integration. L2 rules require integrating multiple evidence items or modalities within the same visit or local context, such as jointly interpreting laboratory results and ultrasound findings, or using multiple indicators together to support a management choice.

L3: Long-horizon reasoning. L3 rules require previous visit records, cross-time trends, or historical evidence far from the decision point, such as events more than 7 days before the decision point. These rules therefore require LLMs to perform temporal anchoring and longitudinal integration over long-range EHR histories.

The stratification is based on **evidence complexity**, rather than strict medical difficulty, in order to better characterize LLMs' ability to identify and integrate historical evidence in long-context EHRs.

For metrics, we mainly report overall accuracy, and further report stratified accuracy by evidence complexity, input scope, and EHR access strategy.

C Statistical Uncertainty and Additional Analyses

C.1 Patient-Clustered Bootstrap Uncertainty for Main Results

To quantify the statistical uncertainty of model evaluation, we compute patient-clustered percentile bootstrap confidence intervals for each model under each input-scope setting.

For a fixed model and input scope, let $\mathcal{D}_p = \{(y_{pj}, \hat{y}_{pj})\}_{j=1}^{n_p}$ denote the decision-point cases contributed by patient p, where y_{pj} is the gold answer and $\hat{y}_{pj}$ is the model prediction. Across P patients, the total number of cases is $N = \sum_{p=1}^{P} n_p$. The point accuracy is

$$\hat{A} = \frac{1}{N} \sum_{p=1}^{P} \sum_{j=1}^{n_p} \mathbb{I}(\hat{y}_{pj} = y_{pj}), \tag{2}$$

which is consistent with Eq. (1).

For each bootstrap replicate $b = 1, \ldots, B$, we sample P patients with replacement. Let $i_1^{(b)}, \ldots, i_P^{(b)}$ denote the sampled patient indices and let $N^{(b)} = \sum_{k=1}^{P} n_{i_k^{(b)}}$. All decision-point cases belonging to each sampled patient are retained. The bootstrap accuracy is

$$\hat{A}^{(b)} = \frac{1}{N^{(b)}} \sum_{k=1}^{P} \sum_{j=1}^{n_{i_k^{(b)}}} \mathbb{I}\left(\hat{y}_{i_k^{(b)} j} = y_{i_k^{(b)} j}\right). \tag{3}$$

We report the two-sided $100(1-\alpha)\%$ percentile confidence interval

$$\left[Q_{\alpha/2}(\hat{A}^*),\ Q_{1-\alpha/2}(\hat{A}^*)\right], \tag{4}$$

where $\hat{A}^*$ denotes the empirical distribution of the bootstrap replicates and $Q_q(\hat{A}^*)$ denotes its q-th quantile. In Table 6, we report the half-width

$$m = \frac{Q_{1-\alpha/2}(\hat{A}^*) - Q_{\alpha/2}(\hat{A}^*)}{2}, \tag{5}$$

so that each result is presented as $\hat{A} \pm m$ in percentage points. We use $B = 10{,}000$ and $\alpha = 0.05$, with a fixed random seed for each model–scope pair.

C.2 Expanded Patient-Level Analysis

We further evaluate model performance across longitudinal decision points at the patient level using all 1,500 cases and all 17 models under the **History-level EHR** setting. We define each patient's first decision point as DP1, yielding 976 cases, and group all subsequent decision points as DP2+, yielding 524 cases. For each model, we test the one-sided hypothesis that DP2+ accuracy is lower than DP1 accuracy.

Model	DP1 Acc.	DP2+ Acc.	Diff. (pp)	p
Claude-Haiku-4.5	61.99	55.73	−6.26	0.0095^{**}
GPT-5.4-mini	65.57	63.55	−2.02	0.2177
Gemini-3-Flash	69.67	66.98	−2.69	0.1438
DeepSeek-V4-Flash	66.29	60.88	−5.41	0.0192^{**}
Qwen3-VL-4B	60.45	58.59	−1.86	0.2419
Qwen3-VL-8B	63.83	58.02	−5.82	0.0140^{**}
Qwen3.5-4B	60.66	57.06	−3.59	0.0890^{*}
Qwen3.5-9B	61.17	55.73	−5.44	0.0208^{**}
Qwen3.5-35B-A3B	63.93	53.05	−10.88	$< 0.0001^{***}$
Gemma-4-E4B	54.30	47.71	−6.59	0.0073^{**}
Gemma-4-26B-A4B	63.93	54.58	−9.35	0.0002^{***}
Phi-4-mini	54.00	45.80	−8.19	0.0012^{**}
HealthGPT-Pro-4B	57.89	55.73	−2.16	0.2101
HealthGPT-Pro-8B	63.11	57.82	−5.29	0.0231^{**}
Hulu-Med-30A3	56.05	49.81	−6.24	0.0105^{**}
Lingshu-7B	60.66	57.63	−3.02	0.1284
MedGemma-1.5-4B	43.85	36.64	−7.21	0.0031^{**}

Table 7: Expanded patient-level analysis under the **History-level EHR** setting. Accuracy is reported in percent, and the difference is computed as DP2+ accuracy minus DP1 accuracy. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.001$.

DP2+ accuracy is lower than DP1 accuracy for all 17 models, with decreases ranging from 1.86 to 10.88 percentage points and an average decrease of 5.41 points. The decrease is statistically significant at $p < 0.1$ for 12 of the 17 models. These results show that later decision points are generally more challenging across models. Together with the conditional analysis in Section 4.3, they further show that model errors across longitudinal decision points are associated within patients.

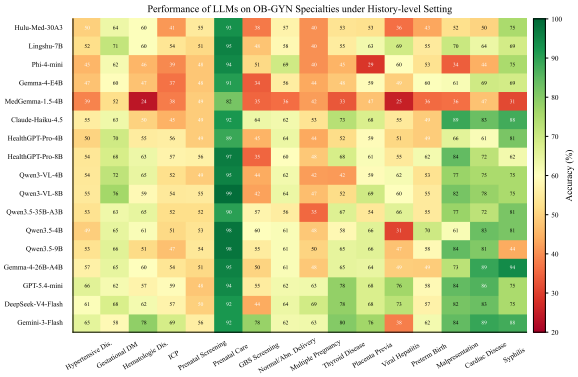

Figure 6: Obstetric and gynecologic chapter-level accuracy heatmap under the **History-level EHR** setting.

C.3 Paired-Bootstrap Uncertainty for EHR Access Strategies

To quantify the statistical uncertainty in the comparison of EHR access strategies, we perform a paired-bootstrap analysis by pairing each strategy with **Direct** access on the same benchmark cases. Table 8 reports the overall accuracy difference relative to **Direct** access and the half-width of its 95% paired-bootstrap confidence interval.

Access strategy	Accuracy gap vs. Direct (pp)
Image	-6.91 ± 1.59
Static RAG	-0.71 ± 1.44
Static RAG + Rec.	-0.44 ± 1.47
Rolling Summary	-1.67 ± 1.51
Agent	$+4.53 \pm 1.60$

Table 8: Paired-bootstrap uncertainty for the EHR access-strategy comparison. Each entry reports the overall accuracy difference relative to **Direct** access and the half-width of its 95% paired-bootstrap confidence interval.

C.4 Chapter-level Analysis

This subsection provides a chapter-level analysis under the **History-level EHR** setting. We group benchmark cases by obstetric and gynecologic chapters to examine how LLM performance varies across clinical topics when LLMs must answer from the full pre-decision EHR history.

Figure 6 shows that LLM performance varies substantially across chapters under **History-level EHR**. In the **Prenatal Care** chapter, LLMs main-

tain relatively high accuracy. Clinically, prenatal care usually starts with early pregnancy registration and gestational-age confirmation; many prenatal care tasks have clear gestational-age anchors, repeated visit structures, and relatively standardized examination pathways (Peahl and Howell, 2021). These features make it easier for LLMs to locate decision-relevant evidence from the pre-decision EHR history.

In contrast, chapters such as **ICP**, **Prenatal Screening**, **GBS Screening**, **Hypertensive Disorders**, and **Multiple Pregnancy** show lower accuracy under **History-level EHR**. Taking **ICP** as an example, clinical judgment requires jointly considering pruritus symptoms, total bile acids, liver enzymes, gestational age, and timing of delivery. SMFM recommends measuring total bile acids and transaminases when ICP is suspected, and determining monitoring and delivery strategies based on bile acid levels and gestational age (Society for Maternal-Fetal Medicine (SMFM) et al., 2021). Thus, LLMs must not only identify an abnormal laboratory result, but also determine whether it occurs at the relevant gestational age, whether it matches the symptoms, whether it reaches the management threshold, and how it relates to delivery timing.

Similarly, **GBS Screening** depends on the screening window of 36+0–37+6 weeks of gestation, culture results, and history of GBS bacteriuria or previous neonatal infection (American College of Obstetricians and Gynecologists, 2020b); **Hypertensive Disorders** requires distinguishing blood pressure before and after 20 weeks, proteinuria, platelet count, liver and kidney function, and severe features (American College of Obstetricians and Gynecologists, 2020a); and **Prenatal Screening** requires integrating gestational age, screening method, fetal number, previous screening results, and subsequent diagnostic pathways (American College of Obstetricians and Gynecologists' Committee on Practice Bulletins–Obstetrics et al., 2020). The correct answers in these chapters are not determined by a single record, but depend jointly on gestational age, repeated examinations, and historical results. This explains why these chapters place greater demands on evidence utilization, temporal anchoring, and longitudinal integration under the **History-level EHR** setting.

Judge	DS-V4	CS-4.6	GPT-5.5	Grok-4.5
DS-V4	1.00	0.77	0.74	0.73
CS-4.6	0.77	1.00	0.85	0.83
GPT-5.5	0.74	0.85	1.00	0.85
Grok-4.5	0.73	0.83	0.85	1.00

Table 9: Pairwise quadratic weighted kappa among four LLM judges on the same 100 randomly sampled instances. The judges are DeepSeek-V4-Flash (DS-V4) (DeepSeek-AI et al., 2026), Claude-Sonnet-4.6 (CS-4.6) (Anthropic, 2026b), GPT-5.5 (OpenAI, 2026c), and Grok-4.5 (xAI, 2026). All judges used the identical prompt, inputs, and 0–5 evidence-utilization rubric; DS-V4 was the primary judge used in Figure 5(b).

D LLM-as-Judge Robustness

DeepSeek-V4-Flash was used as the primary judge for the evidence-utilization scores reported in Figure 5(b), following the 0–5 rubric described in Appendix G.8. To assess robustness to judge-model choice, we used three additional judges, Claude-Sonnet-4.6, GPT-5.5, and Grok-4.5, to re-score the same 100 randomly sampled instances using the identical prompt, inputs, and rubric. Table 9 reports the pairwise quadratic weighted kappa results.

Pairwise QWK ranged from 0.73 to 0.85, with a mean of 0.80 across the four judge models. These results indicate that the four models produced consistently aligned evidence-utilization scores when applying the same 0–5 rubric.

E Computational Budget

Commercial LLMs are evaluated through their official APIs, so their parameter counts and serving infrastructure are not publicly available. For open-source and medical LLMs, model sizes are indicated by their released model names or official model cards when available. Local inference experiments are run with vLLM on NVIDIA H100 80GB GPUs. Because our experiments combine API-based commercial LLMs and locally served open-source or medical LLMs with different inference infrastructures, a single unified GPU-hour estimate is not directly comparable across model groups. For EHR access strategies, we report average token cost in Table 4.

F Implementation Details

We use the Qwen3.5 tokenizer to measure EHR input length. Local inference for open-source and medical LLMs is performed with vLLM 0.19.1 on

NVIDIA H100 80GB GPUs. For these locally deployed models, we disable thinking mode when applicable and use the generation parameters recommended in each model's official generation configuration. Commercial LLMs are evaluated through their official APIs, using the default automatic reasoning or thinking configuration when provided by the API.

For **Rolling Summary**, EHR histories are processed in consecutive 7-day windows. All EHR access strategies are evaluated on the same benchmark instances using the same three Qwen3.5-series base models. Because **Agent** and **Rolling Summary** involve multiple inference rounds, we do not impose an identical total token budget across strategies. Instead, Table 4 reports the actual average token cost of each strategy. All prompt templates are provided in Appendix G.

F.1 RAG Implementation Details

Retrieval uses Qwen3-Embedding-4B (Zhang et al., 2025) embeddings and cosine similarity. Each pre-decision EHR is split at natural calendar-date boundaries, and chunks do not span multiple dates. Within each date, records are organized by event type. The maximum chunk length is 10,000 characters; longer records are split at event-type or event-batch boundaries.

Static RAG uses the question text as the retrieval query and returns the top-10 chunks. **Static RAG + Recency** uses the same retrieval setting and applies linear temporal weighting:

$$s_i^{\text{final}} = s_i^{\text{retrieval}} \left(1 + 0.3 \frac{i}{n-1}\right), \qquad (6)$$

where the retrieved chunks are ordered chronologically, with $i = 0, \ldots, n-1$. The oldest chunk therefore receives a weight of 1.0, whereas the most recent chunk receives a weight of 1.3.

F.2 Agent Implementation Details

Each call to `search_ehr` returns five chunks by default, while the model may request up to ten. The **Agent** may issue multiple searches using different queries and runs for at most 20 interaction turns. It stops when `submit_answer` is invoked. If the maximum number of turns is reached, the model is instructed to submit its final answer immediately. The Agent prompt is provided in Appendix G.7.

Licenses

We use publicly available models, benchmarks, and related resources in accordance with their respective licenses or terms of use. Our evaluation code is publicly available under the MIT License.

G Prompts

This appendix lists the data-construction, MCQ inference, Agent, and evaluation prompt templates used in ObGynLongBench. Unless otherwise noted, all single-pass settings share one two-message chat template: one `system` message and one `user` message. The **Agent** access strategy uses a multi-turn tool-calling loop instead (Appendix G.7). The **Image** access strategy sends the pre-decision EHR as rendered page images to a vision-language model (Appendix G.5). Each case is first converted to a multiple-choice question string (`format_question`): the question stem followed by options A–D on separate lines. The original deployed construction and MCQ prompts are written in Chinese; we provide English translations below. Model rationales are also requested in Chinese in the MCQ output JSON. All prompt figures are collected at the end of this appendix.

G.1 Data Construction Prompts

Figure 2 summarizes the overall construction pipeline (Appendix A.2). This subsection lists the LLM prompts used in rule mining, evidence-complexity assignment, candidate branch classification, and question construction with verification. Candidate branch classification and grounding verification use multi-turn agents; we show only the core `system` instructions in the figures below.

Clinical rule mining. Rule mining proceeds in two LLM steps: (1) discover clinical decision chains from textbook/guideline text (Figure 7), and (2) convert each chain into a structured executable rule (Figure 8).

Evidence-complexity level assignment. Each rule branch is assigned to L1, L2, or L3 based on the evidence a test-taker must locate in the EHR to answer correctly (Figure 9).

Candidate branch classification. After programmatic high-recall screening, an agent reviews each patient–rule candidate and assigns a branch ID consistent with pre-decision EHR evidence (Figure 10).

Question construction and verification. Accepted candidates pass four LLM steps: chain audit (Figure 11), draft question generation (Figure 12), grounding verification (Figure 13), and final rewrite (Figure 14). These steps are implemented as separate prompts rather than three independent verification agents.

G.2 Inference and Evaluation Prompts

The prompts below are used at benchmark inference and analysis time.

G.3 Shared MCQ Inference Template

Figure 15 shows the shared template used by **Evidence-only**, **Visit-level EHR**, and **History-level EHR**. Table 10 summarizes the setting-specific *{SCOPE_INSTRUCTION}* and user input layout; **Static RAG**, **Static RAG + Recency**, and **Image** reuse the same system-message template with different scope instructions. **Image** renders the full pre-decision EHR as page-level images and uses a vision-language input format (Figure 17). **Rolling Summary** uses a separate rolling-summarization prompt for intermediate windows and a final MCQ prompt on the last window (Figure 18 and Figure 19). **Agent** uses a multi-turn tool-calling prompt under the **History-level EHR** setting (Figure 20).

G.4 Static RAG User Message

Static RAG and **Static RAG + Recency** reuse the system-message template with *{SCOPE_INSTRUCTION}* from Table 10, but use the user-message layout in Figure 16.

G.5 Image Prompt

Under **Image**, the full pre-decision EHR text is paginated and rendered as grayscale PNG page images. The vision-language model reuses the same system-message template with *{SCOPE_INSTRUCTION}* from Table 10, but receives rendered page images together with the formatted question in a multimodal user message (Figure 17).

G.6 Rolling Summary Prompts

Under **Rolling Summary**, the pre-decision EHR is split into consecutive 7-day windows. For each non-final window, the model receives the previous cumulative summary and the current window's EHR events, and outputs an updated cumulative summary. On the final window, the model receives the cumulative summary, the last window's EHR events, and the question, and outputs the MCQ answer in JSON format.

G.7 Agent Prompt

Under the **History-level EHR** setting, **Agent** answers each MCQ through a multi-turn tool-calling loop. The agent has three tools: `get_overview` for patient overview, `search_ehr` for semantic retrieval over the full pre-decision EHR, and `submit_answer` for submitting the final option and rationale. Only records on or before the decision cutoff are accessible.

G.8 Evidence Utilization Judge Prompt

For the context-length analysis in Section 4.2, we use an LLM judge to score how substantively each model's rationale draws on the gold supporting evidence $\mathcal{E}$. The judge receives the formatted question, the filtered numbered evidence list (using the same cutoff and scrub rules as **Evidence-only**), and the candidate model's `chosen_option` and `reasoning`. Unlike the MCQ prompts above, this judge prompt was deployed in English (Figure 21).

Instantiations of the MCQ Inference Prompt		
Setting	**Scope instruction**	**User input header**
Evidence-only	The following text contains **key clinical evidence** extracted from the patient's EHR. These items were curated by another physician reviewing the record and cover the information relevant to the current decision. Answer using only these evidence items and the question.	Key Clinical Evidence
Visit-level EHR	The following text contains the patient's basic profile and **EHR events from the most recent 1 day**. Earlier records are omitted. Answer using only the provided information and the question.	EHR Excerpt
History-level EHR	Answer using only the provided **EHR excerpt** and the question.	EHR Excerpt
Static RAG / Static RAG + Rec.	The following text contains the **record snippets retrieved** from the full EHR that are most relevant to the question. The full history may contain additional information, but only the retrieved snippets are shown. If the retrieved information is insufficient, make the best decision based on what is available.	See Fig. 16
Image	The following **images** contain the patient's EHR records (text rendered as page-level images). Read all images carefully and answer the question based on the clinical content.	See Fig. 17

Table 10: Instantiations of the MCQ inference prompt across input-scope and access settings. Static RAG + Recency uses the same prompt as Static RAG; only the retrieval weighting differs. Image uses a vision-language multi-modal user message.

Decision Chain Discovery Prompt (English Translation)

System message

You are an expert in obstetric clinical guideline analysis. Your task is to systematically discover all **clinical decision chains** from textbook text—patterns of the form "if a pregnant patient presents condition X → management action Y should be performed within time T."
Core principles:

1. Extract only from the given text; do not add external knowledge.
2. Each chain must contain an explicit trigger-condition → management-action causal relation.
3. There must be a temporal span (trigger and action are separated by days to months, not instantaneous operations).
4. Preserve numeric thresholds from the source text (e.g., "fasting glucose $\geq$5.1 mmol/L") and medication doses.
5. Identify all branch decision points—under the same trigger, different situations follow different management paths.
6. Do not miss chains: one chapter may contain multiple distinct decision chains.

Granularity control (very important):

- Each chain corresponds to one independently answerable **core clinical question**, e.g., "After GDM diagnosis, is full-pregnancy management needed?"
- Do not split sub-steps of the same management action into separate chains; merge them as branches under one chain.
- Independent decisions across different time windows should remain separate chains.
- Prenatal management, delivery-mode decisions, and postpartum follow-up are usually three independent chains.

User message

Discover all clinical decision chains from the following textbook/guideline text.
Chapter/guideline name: *{CHAPTER_NAME}*
Page range: *{PAGE_RANGE}*
{HINT_SECTION}
Text content
{OCR_TEXT}
Output JSON listing all discovered chains with fields such as `chain_id`, `trigger_condition`, `trigger_criteria`, `management_action`, `expected_timeline`, `branch_points`, `confidence`, and `textbook_quotes`. A chapter typically contains 3–8 chains; each chain should have at least two branch points, including a negative "not found/not performed" branch.

Figure 7: Prompt for discovering clinical decision chains from textbook/guideline OCR text.

Structured Rule Generation Prompt (English Translation)

System message

You are a clinical data-mining engineer. Your task is to convert natural-language clinical decision chains into structured rules executable over EHR event streams.
You will receive:

1. One or more clinical decision chains (from textbook analysis).
2. A description of EHR data fields (Event and PatientContext structure).

You must output structured rules for keyword matching over longitudinal EHR events.
Rule-merging principles (very important):

- If multiple input chains share the same trigger but have complementary management actions, merge them into one rule with branches.
- If the same disease has independent "pregnancy management" and "postpartum follow-up" windows, keep separate rules.
- Each final rule should correspond to one independently answerable clinical question.
- Branches must be mutually exclusive, exhaustive, and include a `_NOT_FOUND` negative branch.

User message

Convert the following clinical decision chains into structured data-mining rules.
Clinical decision chains
{CHAINS_JSON}
Data schema description
{DATA_SCHEMA}
Output JSON with one structured rule per chain, including `rule_id`, `trigger` (keywords, event types, value/context conditions, gestational-age anchor), `action`, `branches`, `question_template`, and `medical_background`. Keyword design should cover Chinese synonyms, English abbreviations, drug generic/brand names, and multiple lab-name variants; branches must be exhaustive.

Figure 8: Prompt for converting discovered decision chains into structured executable rules.

Evidence-Complexity Level Assignment Prompt (English Translation)

System message

You are an obstetric clinical evaluation expert. Given branch information, assign a data-dependency level (L1/L2/L3).
Level definitions:

- **L1 (single-point recognition)**: the test-taker needs only one key finding in the record (one screening result, one lab value, or one diagnosis) at one time point.
- **L2 (multi-source integration)**: the test-taker must consult $\geq$2 different data types within the same time window (e.g., blood pressure + proteinuria, or diagnosis + medication).
- **L3 (longitudinal lookback)**: the test-taker must review clinical data spanning $>$7 days (repeated BP measurements, serial visits, repeated labs, before/after treatment, cross-gestational-age checks).

Assignment principles:

1. Base the level on what the test-taker must look up in the EHR to answer correctly, not on how long the condition text is.
2. Explicit cross-time comparisons ("improved after treatment", "multiple follow-ups", "previously present", "before 20 weeks", "postpartum recovery") $\rightarrow$ L3.
3. Multiple data sources within one time window $\rightarrow$ L2.
4. One lab/screening result suffices (e.g., "TBA$\geq$40", "screening negative/positive") $\rightarrow$ L1.
5. "And/simultaneously" across different data sources $\rightarrow$ L2; multiple thresholds for the same test $\rightarrow$ still L1.
6. A long condition text whose core still needs only one test type $\rightarrow$ L1.
7. Seemingly simple conditions with implicit cross-time needs (e.g., "hypertension before 20 weeks gestation") $\rightarrow$ L3.

Output a single-line JSON: `{"level": "L1/L2/L3", "reason": "one-sentence rationale"}`.

User message

{BRANCH_METADATA} (rule ID, branch ID, condition, label, etc.)

Figure 9: Prompt for assigning evidence-complexity levels (L1/L2/L3) to rule branches.

Candidate Branch Classification Prompt (English Translation)

System message

You are a senior obstetric clinical expert agent. Based on the patient's EMR data, determine which branch of the specified clinical rule applies.
Workflow:

1. Review basic patient information.
2. Read key data targeted by the rule trigger and medical background.
3. Request additional record slices if needed.
4. Submit a branch classification when sufficient evidence is available.

Clinical rule: *{CATEGORY_NAME}*
Medical background
{MEDICAL_BACKGROUND}
Candidate branches (choose exactly one branch ID; do not invent new IDs)
{BRANCHES_DESC}
Valid branch ID list
{BRANCH_IDS_JSON}
Important notes:

- The `branch` field must be one of the valid branch IDs above; do not use Chinese labels or custom IDs.
- For hypertension-related rules, even a single abnormal BP reading together with a relevant diagnosis may support classification.
- Admission/discharge diagnoses are valid clinical evidence.

User message

Patient summary and candidate metadata: *{PATIENT_SUMMARY}*, *{RULE_ID}*, *{PATIENT_ID}*.

Figure 10: Core system prompt for agent-based candidate branch classification after programmatic screening.

Chain Audit Prompt (English Translation)

System message

You are an obstetric clinical item-review expert. Determine whether the case's reasoning chain is sufficient to support construction of a medical reasoning question.
Criteria:

1. Do `key_events` form an interpretable medical reasoning chain?
2. Even if retrospective information is mixed in, can the case still be rewritten as a valid current-time reasoning question?
3. Is there an obvious scene mismatch between the rule/category and `key_events` (e.g., preterm rule on a term case, syphilis rule without syphilis evidence, tocolysis rule used for antihypertensive management)?

Output JSON: `{"audit_pass": bool, "audit_reason": str, "focus_terms": list}`.
Set `audit_pass=false` only when:

- the chain is medically invalid;
- the branch strongly conflicts with `key_events`;
- the rule/case scene is clearly mismatched;
- no valid current-time question can be inferred.

Do not generate a question; do not output extra fields.

User message

{RULE_ID, CATEGORY_NAME, BRANCH, QUESTION_TEMPLATE, MEDICAL_BACKGROUND, EVIDENCE_SUMMARY, KEY_EVENTS, TOOL_CALLS, TRIGGER_EVENTS_DATES}

Figure 11: Prompt for auditing whether a candidate case supports question construction.

Draft Question Generation Prompt (English Translation)

System message

You are an obstetric clinical item-writing expert. Draft a medical reasoning question based on the case information.
Core principles:

1. Test clinical reasoning and decision-making, not static label classification.
2. The stem gives a clinical scenario and a decision question; the test-taker must look up evidence in the EHR.
3. The semantic direction of `gold_option` is locked by `branch_label` and must not change.

Stem constraints:

- Must bind the exact `cutoff` date as the decision time point.
- Use neutral background wording (e.g., "follow-up in a high-risk obstetric clinic"); do not translate disease names from the rule/category.
- Do not include specific lab values, specific diagnoses, exclusion statements ("no XX seen"), or full abnormal chains.
- Do not use meta-conclusion phrasing such as "already recorded/already decided".
- Do not translate EHR key findings line-by-line into the stem.

Option constraints:

- Correct and distractor options must sit at the same abstraction level.
- Write options as clinical decisions, not retrieval-style wording such as "evidence not found".
- Distractors should resemble realistic alternative management choices.

Output JSON: `{"draft_question_stem", "options" [4], "gold_option", "grounding_focus"}`; if a valid item cannot be written, return `{"reject_reason"}` with other fields empty.

User message

{RULE_ID, CATEGORY_NAME, BRANCH, BRANCH_LABEL, QUESTION_TEMPLATE, BRANCHES, EVIDENCE_SUMMARY, KEY_EVENTS, TOOL_CALLS, AUDIT_REASON, FOCUS_TERMS}

Figure 12: Prompt for drafting an examination-style clinical question and four options.

Grounding Verification Prompt (English Translation)

System message

You are an obstetric clinical evidence-verification agent. Take the draft question into the EHR for targeted verification.
Verification steps

1. **First step (required):** verify the stem premise—confirm that the EHR contains original evidence for the core disease implied by the rule. If not found (e.g., all syphilis tests negative, bile acids never abnormal) → `grounded_pass=false`. Upstream `key_events` may be wrong; do not trust them blindly.
2. Verify that the gold answer is supported by original record evidence.
3. Select a `cutoff` date:
   - Using only data on or before `cutoff`, the test-taker must be able to derive the gold answer.
   - Every `support_evidence` item must have date $\leq$ `cutoff`.
   - All post-`cutoff` data are hidden from the test-taker; set `cutoff` at the end of the evidence chain, not the beginning.
   - If repeated measurements are needed, `cutoff` $\geq$ the date of the last key repeat measurement.
   - For inpatient cases, set `cutoff` to the admission date (the program automatically removes all diagnoses on that day).
   - If no `cutoff` can satisfy both answerability and non-leakage, `grounded_pass=false`.
4. Generate `precise_scrub_targets` (exact deletion list).
5. If verification cannot be completed within the tool-call budget, `grounded_pass=false`.

Scrub rules

- After choosing `cutoff`, the program automatically removes all diagnoses on that day; notes/procedures on that day that leak the answer must be listed for scrubbing.
- Post-`cutoff` data are globally masked; do not list them in scrub targets.
- Do not delete pre-`cutoff` labs or vitals.
- Pre-`cutoff` notes/medications may be deleted only when they state management conclusions or discharge summaries.
- Labs on `cutoff` day are kept by default; same-day follow-up tests (e.g., OGTT 1h/2h) may be scrubbed with reason prefixed by "[same-day follow-up]" and no conditional wording.
- `match_hint` contains only `value_text`, without a name prefix.
- Scrub only answer leakage, retrospective outcomes, or management conclusions; do not output spurious scrub items.

Gold direction

- `branch_label` locks the semantic direction of the gold answer; it must not change.
- If the EHR does not support the branch, set `grounded_pass=false` with `drop_reason`; do not invent a new direction.
- You do not assign L1/L2/L3; the level is determined by the rule system.

User message

Draft question, options, gold option, grounding focus, audit reason, focus terms, key events, and trigger-event dates: *{CASE_AND_DRAFT_FIELDS}*.

Figure 13: Core system prompt for grounding verification: cutoff selection, supporting evidence collection, and leakage scrubbing.

Final Question Rewrite Prompt (English Translation)

System message

You are an obstetric clinical item editor. Without changing the core testing point or answer, revise the draft into the final item.
Core principles:

1. Preserve medical-reasoning item style; the stem may use only information supported by `support_evidence`.
2. The semantic direction of `gold_option` is locked by `branch_label` and must not change.
3. Do not introduce new facts or change option abstraction level.

Stem constraints:

- Explicitly state `confirmed_cutoff_date` as the decision time point.
- Use neutral background wording; do not translate disease names from rule labels.
- Do not include specific values, specific diagnoses, exclusion statements, full abnormal chains, or meta-conclusion phrasing.
- Keep the stem short and examination-like; the test-taker must consult the EHR to answer.

Option constraints:

- Write options as clinical decisions, not retrieval-style wording.
- Do not change gold semantics; only improve natural phrasing.

Output JSON: `{"final_question_stem", "final_options" [4], "final_gold_option"}`.

User message

{CONFIRMED_CUTOFF_DATE, BRANCH_LABEL, DRAFT_STEM, DRAFT_OPTIONS, DRAFT_GOLD, SUPPORT_EVIDENCE, REWRITE_NOTES}

Figure 14: Prompt for rewriting a grounded draft into the final benchmark question.

Shared MCQ Inference Prompt (English Translation)

System message

You are a senior obstetric attending physician. *{SCOPE_INSTRUCTION}*
Your output must be a single-line JSON object without markdown code fences. Keys:

- `chosen_option`: string "A" | "B" | "C" | "D"
- `reasoning`: brief rationale in Chinese

User message

{USER_INPUT_HEADER}
{CLINICAL_INPUT}
Question
{FORMATTED_QUESTION}
Please output JSON.

Figure 15: Shared MCQ inference prompt. The system message is reused by all settings in Table 10; this figure shows the text-based user-message layout used by Evidence-only, Visit-level EHR, and History-level EHR.

Static RAG MCQ Prompt (English Translation)

System message

You are a senior obstetric attending physician. *{SCOPE_INSTRUCTION}*
Your output must be a single-line JSON object without markdown code fences. Keys: `chosen_option` ("A"|"B"|"C"|"D"), `reasoning` (brief rationale in Chinese).

User message

Patient Overview
{PATIENT_OVERVIEW}
Retrieved Record Snippets (*{K}* snippets, ranked by relevance)
1. *{SNIPPET_1}*
2. *{SNIPPET_2}*
...
Question
{FORMATTED_QUESTION}
Please output JSON.

Figure 16: MCQ prompt for Static RAG and Static RAG + Recency. The scope instruction is instantiated in Table 10.

Image MCQ Prompt (English Translation)

System message

You are a senior obstetric attending physician. *{SCOPE_INSTRUCTION}*
Your output must be a single-line JSON object without markdown code fences. Keys: `chosen_option` ("A"|"B"|"C"|"D"), `reasoning` (brief rationale in Chinese).

User message (multi-modal)

{RENDERED_EHR_PAGE_1}
{RENDERED_EHR_PAGE_2}
...
Question
{FORMATTED_QUESTION}
Please output JSON.

Figure 17: MCQ prompt for the Image access strategy. The scope instruction is instantiated in Table 10. Each *{RENDERED_EHR_PAGE_i}* denotes one paginated PNG image of the pre-decision EHR text.

Rolling Summarization Prompt (English Translation)

System message

You are a senior obstetric attending physician, reading a patient's EHR in segments. Based on the **previous summary** and the **new records in the current segment**, output an updated **cumulative summary** that includes:

1. Key personal information (age, gravidity/parity, BMI, important history, etc.)
2. Important clinical indicators and trends (e.g., blood pressure, glucose, weight, laboratory results)
3. Abnormal indicators and confirmed or suspected diagnoses
4. Current medications and important procedures

The summary should be concise yet complete, and should not omit important information recorded in earlier segments.

User message

Patient Overview
{PATIENT_OVERVIEW}
Previous Cumulative Summary
{PREVIOUS_SUMMARY} (omitted in the first window)
Segment *{i}* Records (*{DATE_RANGE}*, *{N}* events)
{CURRENT_WINDOW_EHR}
Please output the updated cumulative summary.

Figure 18: Rolling summarization prompt used for all non-final 7-day windows.

Rolling Summary Final MCQ Prompt (English Translation)

System message

You are a senior obstetric attending physician. You have read the patient's full EHR in segments. The following text contains the cumulative summary from earlier segments and the records from the final segment. Answer the question based on all provided information.
Your output must be a single-line JSON object without markdown code fences. Keys: `chosen_option` ("A"|"B"|"C"|"D"), `reasoning` (brief rationale in Chinese).

User message

Previous Cumulative Summary
{CUMULATIVE_SUMMARY}
Final Segment Records (*{DATE_RANGE}*, *{N}* events)
{FINAL_WINDOW_EHR}
Question
{FORMATTED_QUESTION}
Please output JSON.

Figure 19: Final-window MCQ prompt used in the Rolling Summary access strategy.

Agent Prompt (English Translation)

System message

You are a clinical physician answering a multiple-choice question based on the patient's EHR. You may use the following tools to consult the record:

- `get_overview`: retrieve the patient overview
- `search_ehr`: search EHR snippets by keyword (may be called multiple times with different queries)
- `submit_answer`: submit the final answer

Note: only records on or before *{CUTOFF_DATETIME}* (inclusive) may be used. First review the patient overview, then search for information needed by the question, and finally call `submit_answer`.

Initial user message

{FORMATTED_QUESTION}
Please first call `get_overview` to review the patient's basic information.

Tool: `submit_answer`

Submit the final multiple-choice answer and rationale. Required fields: `chosen_option` ("A"|"B"|"C"|"D"), `reasoning` (brief rationale in Chinese, citing key evidence).

Figure 20: Initial Agent prompt used in the History-level EHR access strategy. Subsequent turns append tool responses and assistant tool calls until `submit_answer` is invoked.

Evidence Utilization Judge Prompt

System message

You are an independent clinical case-item LLM judge.
Your task: using ONLY the question stem/options, the numbered key clinical evidence list, and the candidate's reasoning text below, rate how substantively the candidate's **reasoning draws on items in that evidence list** (explicitly or implicitly—labs, course, orders, dates, etc.).
Rules:

- Do NOT require the candidate to restate every evidence item; focus on decision-relevant evidence in the reasoning chain.
- If the filtered evidence count is 0, assign score **0** per the user message.
- Output a single JSON object (no markdown fences) parseable by `json.loads`.
- `support_evidence_utilization_score` must be an integer **0–5**.
- All string fields in the JSON must be in **English**.

User message

Question stem and options
{FORMATTED_QUESTION}
Key clinical evidence for this item (post-filter; same as at exam time; ***{N_EVIDENCE}*** items)
Indices match the numbered list below. In `matched_evidence_indices`, list **1-based** indices of items substantively used in the reasoning.
{FILTERED_SUPPORT_EVIDENCE}
—
Candidate response (model under evaluation)
chosen_option: *{CHOSEN_OPTION}*
reasoning:
{CANDIDATE_REASONING}
—
Rubric: support_evidence_utilization_score (integer 0–5)

- **5**: Reasoning clearly weaves **multiple decision-relevant** evidence items into the main argument; specific clinical facts tie to the conclusion (not vague platitudes).
- **4**: **Most** relevant items are used; strong alignment with the list; minor items may be omitted.
- **3**: **Some** items used, but important ones missing or only generic references ("labs", "chart") without substantive alignment.
- **2**: Touches only one minor point, or relies on inference poorly aligned with the list.
- **1**: Almost no substantive use of the list (or contradicts evidence while forcing a choice).
- **0**: No use of the list (generic/off-chart reasoning), or *{N_EVIDENCE}* = 0 and reasoning does not reasonably ground on the stem / is pure speculation.

If *{N_EVIDENCE}* = 0: `support_evidence_utilization_score` must be **0**; `matched_evidence_indices = []`; note in `rubric_note` that no filtered evidence items were available.
Return JSON with **only** these keys:

- `support_evidence_utilization_score`: integer 0–5
- `matched_evidence_indices`: array of integers (1-based indices used; `[]` if none)
- `rubric_note`: string, one-sentence rationale ($\leq$200 chars, English)
- `summary`: string, 2–5 sentences in English on alignment between reasoning and evidence

Figure 21: LLM-as-judge prompt for scoring evidence utilization in model rationales. *{FILTERED_SUPPORT_EVIDENCE}* is built with the same cutoff and scrub filtering as the **Evidence-only** input.