LongMedBench: Benchmarking Medical Agents for Long-Horizon Clinical Decision-Making

Yanzhen Chen[1*], Zihan Xu[1*], Xiaocheng Zhang[1*], Zhiting Fan[1], Weiqi Zhai[2], Hongxia Xu[1,3], and Zuozhu Liu[1,3†]

[1] Zhejiang University, Hangzhou, China
{yanzhen.22,zihan1.22,xiaocheng.22,zhiting.23}@intl.zju.edu.cn
einstein@zju.edu.cn, zuozhuliu@intl.zju.edu.cn
[2] Alibaba Group, Hangzhou, China
zhaiweiqi.zwq@alibaba-inc.com
[3] Transvascular Implantation Devices Research Institute, Hangzhou, China

Abstract. We introduce LongMedBench, a real-world EHR benchmark for long-horizon clinical reasoning with medical agents. Prior evaluations of LLM-based medical agents have largely emphasized short-context medical QA, tool use, or single-encounter interactions, whereas real-world care requires integrating evidence across repeated visits, tests, treatments, and discharge outcomes. LongMedBench integrates MIMIC-IV admission records and clinical notes into time-series event streams and controlled long-context memory datasets. It comprises 355 patients, averaging 19.72 inpatient visits per patient and 44.91 medical events per visit. We propose an evaluation taxonomy with three suites: factual QA, temporal reasoning, and long-horizon decision-making. These suites measure how agents retrieve, order, and use historical patient information over extended horizons. Experiments show that recent LLMs use explicit timestamps effectively but still struggle with implicit temporal inference. Longer context, RAG, and agent memory improve factual retrieval more than long-horizon decision-making, revealing a gap between accessing historical evidence and using it for clinical decision reasoning.

Keywords: Computer-Aided Diagnosis · Medical Agents · EHR.

1 Introduction

In clinical practice, diagnosis is inherently longitudinal and time-dependent. Clinicians do not merely react to isolated symptoms; instead, they synthesize evidence from multiple visits, diagnostic tests, and evolving treatment responses over years [5]. Such long-horizon clinical reasoning is fundamental to high-quality care. As large language models (LLMs) transition to autonomous medical agents, their ability to navigate complex trajectories in real-world electronic health records (EHRs) [9] has become a central benchmark for clinical readiness.

[*] These authors contributed equally to the work.
[†] Corresponding author: Zuozhu Liu.

Table 1: Comparison of Medical Agent Benchmarks.

Benchmark	**Long.**[a]	**Ctx.**[b]	**EHR.**[c]	**Dec.**[d]	**Temp.**[e]	**Focus**
MedAgentBench [8]	×	×	✓	✓	×	EHR Tool Integration
AgentClinic [20]	×	×	×	✓	×	Simulated Interaction
DiagBench [18]	×	×	✓	✓	×	Diagnostic Trajectory
MedBench v4 [4]	×	✓	✓	×	×	Medical Knowledge QA
ReflecTool [12]	×	×	×	✓	×	Reflective Tool-Use
EHRSQL [11]	×	×	✓	×	×	Relational Querying
TIMER [2]	✓	×	✓	×	✓	Temporal QA
ER-Reason [15]	✓	×	✓	✓	×	EHR-Based Reasoning
LongMedBench	✓	✓	✓	✓	✓	Long-horizon Reasoning

[a] *Longitudinal*: Reasoning across multiple discrete clinical visits. [b] *Context*: Benchmark environments with multi-turn context. [c] *EHR*: Dataset built from real-world EHRs. [d] *Decision-making*: Evaluating long-horizon clinical decision-making beyond static retrieval. [e] *Temporal Sensitivity*: Evaluating understanding of clinical timeline and urgency.

Although general-purpose benchmarks evaluate long-horizon or multi-turn interactions [22, 21, 10, 23, 7, 4], they primarily emphasize information retrieval, namely locating specific facts within a long context [13]. They test needle-in-a-haystack retrieval but overlook temporal dynamics essential to clinical reasoning. In real-world medical scenarios, the challenge extends to understanding how a patient's state evolves over dozens of visits and how this progression informs treatment strategies.

Current medical evaluation paradigms also do not fully meet this requirement, as shown in Table 1. Recent frameworks [18, 8, 20, 4, 12, 11] have advanced medical-agent evaluation for tool use, simulated encounters, and structured EHR querying. However, they largely evaluate short-horizon or single-encounter behavior and do not systematically test how models integrate clinical evidence across repeated visits. More recent benchmarks explore complementary dimensions: TIMER [2] evaluates temporal reasoning over longitudinal EHRs but in a static QA setting without agent interaction; ER-Reason [15] models the multi-stage ER workflow but is limited to a single clinical episode without multi-visit agent trajectories.

To address these limitations, we introduce LongMedBench, a MIMIC-IV [9]-based benchmark for evaluating how medical agents use cross-visit EHR histories under controlled memory settings. By converting 355 patient records into longitudinal event streams (*Avg.* 19.72 visits per patient), LongMedBench provides a temporally dense environment for measuring how LLM-based medical agents retrieve, organize, and use patient history across visits.

Our contributions are: (1) a tri-level memory dataset architecture representing patient history at different granularities; (2) a progressive evaluation taxonomy spanning *factual QA*, *temporal reasoning*, and *long-horizon decision-making*; and (3) empirical findings showing that state-of-the-art LLMs exploit explicit timestamps but still struggle with implicit temporal reasoning. Although RAG and agent memory systems improve factual retrieval, long-horizon decision-

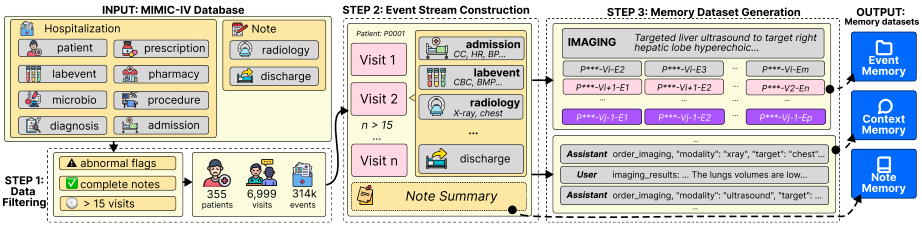


Fig. 1: Data processing pipeline for LongMedBench.

making remains highly dependent on immediate context, revealing a gap in reasoning over long-term clinical trajectories.

2 Methodology

2.1 EHR Data Processing Pipeline

LongMedBench is built on MIMIC-IV v3.1 [9], a public database with medical records for more than 100,000 patients. To convert static EHRs into an interactive agent environment, we design a three-stage pipeline (Figure 1). The pipeline integrates structured MIMIC-IV tables and clinical notes through patient and admission identifiers. Patient and admission tables provide demographic information and visit boundaries; laboratory, medication, procedure, microbiology, and diagnosis tables provide structured clinical events; and radiology and discharge notes from MIMIC-IV-Note provide note-derived imaging reports and visit summaries. This design preserves both timestamped structured evidence and narrative clinical summaries, allowing the same patient trajectory to support retrieval, temporal reasoning, and decision-oriented tasks.

Data Filtering To reduce signal redundancy while preserving clinically salient information, we retain abnormal lab results following prior work [6]. We further filter for patients with ≥ 15 hospitalizations and complete admission/discharge records, obtaining 355 patients and 6,999 visits, with an average of 19.72 visits per patient (median=18.00, SD=5.72). This threshold provides sufficient temporal depth for long-horizon tasks while retaining an adequate cohort size. The resulting cohort is a dense, high-acuity stress-test setting rather than a representative sample of the full MIMIC-IV population. This filtering may remove normal measurements that indicate recovery or help rule out disease; thus, abnormal-event filtering should be viewed as a benchmark construction choice rather than a universal EHR preprocessing rule.

Event Stream Construction Following the three-level trajectory design, each patient $\mathcal{P}$ is represented by an ordered visit sequence $\mathcal{S}_{\mathcal{P}} := \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n\}$, where $\mathcal{V}_i$ is the i-th visit ordered by admission time. This patient level preserves

the cross-visit trajectory needed for long-horizon memory and temporal reasoning. On average, each visit $\mathcal{V}_i$ contains 44.91 medical events (median=25.00, SD=70.32).

At the visit level, each visit $\mathcal{V}_i$ is represented as an ordered event stream. The event stream begins with an admission event E_i^{adm}, is followed by a series of specific medical events, and ends with a discharge event E_i^{dis}. It can therefore be denoted as $\mathcal{V}_i := \{E_i^{\text{adm}}, E_i^1, E_i^2, \ldots, E_i^{\text{dis}}\}$. This level aligns structured hospital events with note-derived admission and discharge summaries, providing visit boundaries for temporal tasks.

At the event level, a medical event is any discrete clinical occurrence extracted from EHR tables or note-derived sources. Formally, the k-th event in $\mathcal{V}_i$ is represented as $E_i^k := (a_i^k, t_i^k, p_i^k, o_i^k)$, where a_i^k is the action type from the clinical action space $\mathcal{A}$, such as lab test, medication, procedure, microbiology test, or imaging; t_i^k is the timestamp; p_i^k is the action parameter; and o_i^k is the corresponding observation result.

Memory Dataset Generation We create three memory modules to compare compressed narrative history, detailed event history, and immediate interaction context under the same patient trajectory.

1. Note Memory: $\mathcal{M}_N^{(i,j)} := \{N_i, N_{i+1}, \ldots, N_{j-1}\}$, which contains continuous visit-level summaries from $\mathcal{V}_i$ to $\mathcal{V}_{j-1}$. These summaries emphasize the global trajectory rather than detailed event recall.

2. Event Memory: $\mathcal{M}_E^{(i,j)} := \{\mathcal{V}_i, \mathcal{V}_{i+1}, \ldots, \mathcal{V}_{j-1}\}$, which contains the ordered visit-level event streams from $\mathcal{V}_i$ to $\mathcal{V}_{j-1}$. Since each $\mathcal{V}_k$ is defined as $\{E_k^{\text{adm}}, E_k^1, E_k^2, \ldots, E_k^{\text{dis}}\}$, event memory can be flattened into a single chronological patient trajectory.

3. Context Memory: $\mathcal{M}_C^{(j,T)} := \{(r_0, m_0), (r_1, m_1), \ldots, (r_P, m_P)\}$, where r_k and m_k denote the dialog role and message content. Any E_j^k with $t_j^k < T$ is transformed into an LLM-style trajectory using an LLM: {(assistant,{a_j^k,t_j^k,p_j^k}), (user,{f_j^k,$t_j^{k\prime}$})}, where f_j^k and $t_j^{k\prime}$ denote simulated feedback from the clinical environment and the corresponding timestamp inferred from o_j^k. This simulates the agent's immediate interaction state.

2.2 Benchmark Question Generation

Based on the memory modules $\{\mathcal{M}_N^{(i,j)}, \mathcal{M}_E^{(i,j)}, \mathcal{M}_C^{(j,T)}\}$, we construct three progressively challenging task families spanning *factual QA*, *temporal reasoning*, and *long-horizon decision-making*, reflecting increasingly global memory dependency.

Factual QA This task evaluates precise retrieval and temporal alignment within the agent's event memory. For each question, we sample a ground-truth event E_i^k from $\mathcal{V}_i$ and reformat it into one of two formats: *explicit questions* include t_i^k in the query text, evaluating direct retrieval capacity (e.g. *what is the Chief Complaint at 2108-07-12 19:12:00 at Visit n?*); *relative questions* remove

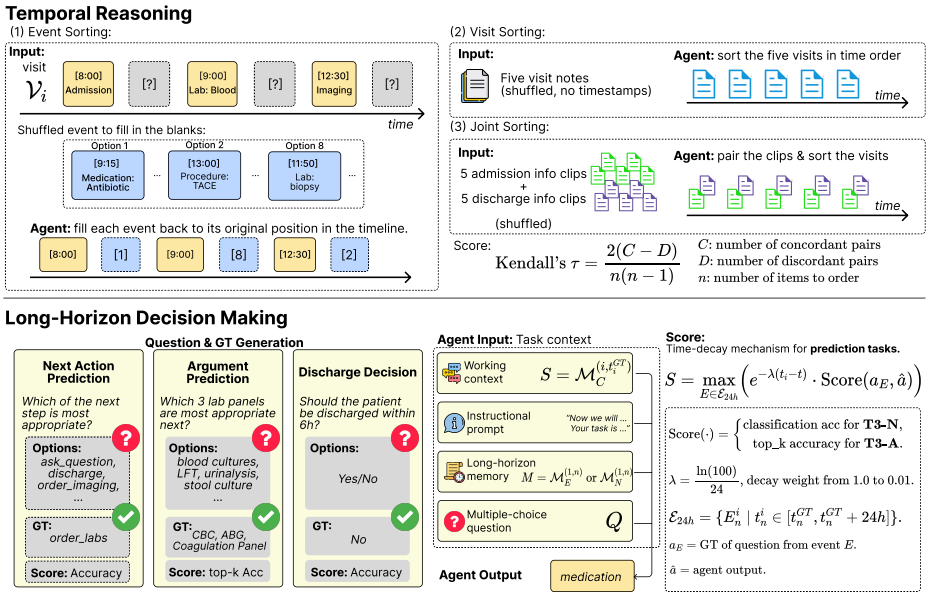


Fig. 2: Workflow for Temporal Reasoning and Long-Horizon Decision Making.

timestamps and specify relative temporal relationships (e.g. *what is INR(PT) in the 7th lab of Visit n?*). The questions require the agent to recover source-derived factual details p_i^k or o_i^k from a redundant memory window $\mathcal{M}_E^{(i-m,i+m+1)}$, where m controls the window size. We place $\mathcal{V}_i$ in the middle of the window to evaluate retrieval from non-recency-favored regions of long-context memory [13]. For automatic scoring, models are prompted with few-shot examples to return exact source values without paraphrasing. Outputs are normalized and evaluated by exact matching against source-derived targets, prioritizing retrieval precision over semantic paraphrase flexibility.

Temporal Reasoning This task evaluates the agent's ability to reconstruct chronological order at different memory granularities. The evaluation score is measured by Kendall's τ. Given a target event stream $\mathcal{S}$ or one of its visits $\mathcal{V}_i$, we design three subtasks using the original order as ground truth, each targeting a distinct level of temporal reasoning. *Visit cloze* tests event-level reasoning: several intervention events $E_i^k \in \mathcal{V}_i$ with timestamps preserved are masked and provided to the agent as a shuffled list. The agent must insert each event back to its original position. *Visit sorting* evaluates visit-level ordering: five shuffled visit summaries $N \in \mathcal{M}_N^{(i,i+5)}$ must be ordered solely by clinical progression cues with timestamps removed. *Joint sorting* assesses integrated capacity: five visits are split into ten admission/discharge clips. The agent must correctly pair the clips for each visit and order them chronologically.

Long-Horizon Decision Making This task measures history-consistent action decision-making, as recorded EHR actions reflect real practice rather than prospectively validated gold standards; questions are multiple-choice with LLM-generated clinically plausible distractors. From $\mathcal{V}_n$, we sample a ground-truth event E_i^{GT}, and define the agent's working context $S = \mathcal{M}_C^{(i,t_i^{GT})}$. For question Q, we randomly select one of the following three formats: *next action prediction* requires the agent to predict a_i^{GT}; *argument prediction* evaluates inference for p_i^{GT}; *discharge decision* queries if the patient can be discharged within 6h. To handle the gold-answer limitation and lab result cut, lab events are excluded from building two *prediction* tasks, and a time-decay scoring mechanism is applied to recognize events that occur within $24h$ after t_i^{GT}, as shown in Figure 2. The agent can refer to injected memory $M = \mathcal{M}_E^{(1,n)}$ or $\mathcal{M}_N^{(1,n)}$ during inference. Unlike prior tasks, this task requires integrating immediate state S with long-term memory M to produce clinically coherent decisions.

3 Experiments and Results

3.1 Experimental Setup

To evaluate medical agents in long-horizon clinical decision-making, we deployed representative long-context LLMs with their default configurations as agent backbones, including *GPT-5-mini* [16], *Qwen-Turbo* [19], and *DeepSeek-V3.2* [3]. Following common retrieval protocols in long-horizon tasks [22, 14], we set up three memory architectures: (1) *Naive Long-Context.* Memories from preceding visits are directly packed into the LLM's context window without external retrieval tools. (2) *RAG. text-embedding-3-small* [17] is used to vectorize memory entries; the agent can retrieve top-K similar memory fragments when answering the question. (3) *Agent Memory System.* Mem0 [1], an open-source agent memory framework, is deployed to maintain a persistent vector store with embedding-based semantic search. The agent can add, update, retrieve, and analyze memories across extended interaction histories [1].

3.2 Results

Factual QA We evaluate naive long-context LLMs using *Qwen-Turbo*, with window sizes from $\{3, 5, 7\}$ (*i.e.*, $m \in \{1, 2, 3\}$) and full history ($m = \infty$), and compare them with the Agent Memory system (Mem0). Results in Table 2 report Lab-T (recall) and Lab-F (rejection) to assess hallucination impact. Table 2 shows that the naive LLM's performance decays severely as history grows, dropping from 0.882 to 0.423 in explicit retrieval, with low Lab-F scores ($\leq$0.570) highlighting a pervasive hallucination bias. The benefit of the Agent Memory system varies across task types. Mem0 achieves near-optimal results in explicit Lab-T (0.993) and Medication (0.983), but its relative performance (*Avg.* 0.331) remains inferior. This suggests that memory augmentation can improve access to certain factual records but does not by itself provide temporal alignment or relational indexing across events.

Table 2: Performance on Factual Retrieval under Explicit and Relative Settings.

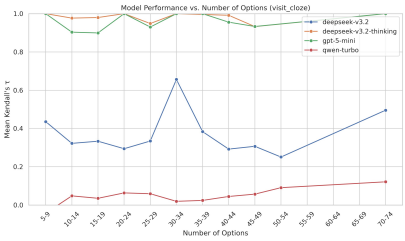
	Explicit							Relative					
Model	L-T^f	L-F^g	L-O^h	Med.	Img.	Adm/Dis	Avg.	L-T^f	L-F^g	L-O^h	Med.	Img.	Avg.
$m = 1$	0.96	**0.57**	**0.83**	0.84	**0.89**	0.97	**0.88**	**0.91**	**0.58**	**0.80**	**0.78**	**0.84**	**0.80**
$m = 2$	0.92	0.51	0.78	0.71	0.81	0.92	0.81	0.86	0.52	0.74	0.64	0.74	0.71
$m = 3$	0.86	0.50	0.74	0.61	0.75	0.87	0.74	0.80	0.50	0.70	0.55	0.68	0.64
$m = \infty$	0.63	0.29	0.51	0.28	0.41	0.46	0.42	0.52	0.28	0.44	0.23	0.36	0.34
Mem0	**0.99**	0.43	0.80	**0.98**	0.07	**0.98**	0.74	0.71	0.57	0.66	0.17	0.13	0.33

^{f}L-T: Lab-T, recall of existing lab indicators. ^{g}L-F: Lab-F, rejection of absent lab indicators.
^{h}L-O: Overall accuracy of lab events.

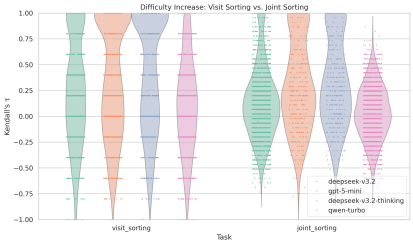
Table 3: Temporal reasoning performance (Kendall's τ) on LongMedBench.

	visit_cloze			*visit_sorting*			*joint_sorting*		
Model	N	Mean	Std	N	Mean	Std	N	Mean	Std
DeepSeek-V3.2	6776	0.316	0.352	765	0.258	0.523	2878	0.132	0.346
GPT-5-mini	6776	0.925	0.220	765	0.376	0.552	2878	0.295	0.439
DeepSeek-V3.2-Thinking	6776	**0.969**	0.124	765	**0.424**	0.549	2878	**0.330**	0.425
Qwen-Turbo	6776	0.047	0.257	765	0.114	0.456	2878	0.033	0.256

Temporal Reasoning Table 3 presents Kendall's τ across three progressive temporal tasks. In event-level Visit Cloze, top models achieve strong performance (*GPT-5-mini*: 0.925; *DeepSeek-V3.2-Thinking*: 0.969), confirming that strong LLMs can use explicit timestamp cues. Performance drops sharply in Visit Sorting, where timestamps are removed and models must infer order from clinical progression. Joint Sorting is harder still, with the best mean τ decreasing to 0.330, showing the added difficulty of fragmented admission-discharge evidence. This explicit-to-implicit performance gap indicates that current models can leverage explicit time anchors but still struggle to infer clinical progression from fragmented multi-visit evidence. Figure 3 summarizes these trends.



(a) Mean Kendall's τ vs. number of options in *visit_cloze*.



(b) Comparison between *joint_sorting* and *visit_sorting*.

Fig. 3: Temporal reasoning analysis.

Table 4: Long-context LLM performance on Long-Horizon Decision Making.

	Note Memory				Event Memory			
Model	**T3-A**	**T3-D**	**T3-N**	**Avg**	**T3-A**	**T3-D**	**T3-N**	**Avg**
Qwen-Turbo	0.475	0.604	0.309	0.434	0.494	0.604	0.309	0.420
DeepSeek-V3.2	**0.478**	0.607	0.278	0.422	0.493	**0.607**	0.304	0.439
GPT-5-mini	0.478	**0.625**	**0.324**	**0.445**	**0.494**	**0.607**	**0.335**	**0.452**
DeepSeek-V3.2-Thinking	0.471	0.578	0.294	0.420	0.466	0.593	0.290	0.419

Table 5: Ablation study on Long-Horizon Decision Making.

Visit Injection						Memory Architecture					
n	Memory	T3-A	T3-D	T3-N	Avg	Arch	Memory	T3-A	T3-D	T3-N	Avg
0^i	None	**0.49**	**0.61**	**0.32**	**0.45**	RAG	event	0.45	0.51	0.29	0.40
2	event	0.49	0.60	0.32	0.44	Mem0	event	0.47	**0.62**	0.31	**0.44**
5	event	0.48	0.60	0.31	0.43	RAG	note	**0.48**	0.60	**0.31**	0.44
2	note	0.48	0.60	0.32	0.44	Mem0	note	0.46	0.56	0.29	0.41
5	note	0.48	0.60	0.32	0.44						

^{i}Baseline with no memory injected.

Long-Horizon Decision Making We first evaluate Naive Long-Context under a common 128K-token input cap for this experiment. As shown in Table 4, structured event memory outperforms note memory for most models, and *GPT-5-mini* achieves the best average performance. Thinking mode does not consistently improve long-horizon decision-making, suggesting that explicit reasoning traces alone do not guarantee better use of longitudinal history. This further separates decision-making performance from the timestamp-based ordering success observed in temporal reasoning.

To analyze the Lost-in-Middle [13] effect with *Qwen-Turbo* as the baseline model, we (1) adjusted the number of injected historical visits and (2) implemented RAG and Mem0 architectures as retrieval sources in separate experiments. As shown in Table 5, performance changes are small and remain close to the full-memory condition, while increasing the injected history length yields slight declines. The performance differences between Mem0 and RAG are also minimal, both close to the baseline performance, indicating that existing memory augmentation architectures offer limited gains for long-horizon decision-making in this setting.

Notably, while RAG and Mem0 help factual retrieval, they provide limited gains for long-horizon decision-making. This negative result is an intended diagnostic finding of LongMedBench: agents may retrieve patient history without reliably integrating it into clinical decision reasoning. Therefore, the benchmark exposes a gap between memory access and history-aware decision reasoning rather than merely measuring whether more history can be inserted into context.

4 Conclusion

We introduce LongMedBench, a real-world EHR benchmark for long-horizon clinical reasoning with medical agents. By converting MIMIC-IV records into multi-session event streams and controlled memory datasets, LongMedBench evaluates factual QA, temporal reasoning, and long-horizon decision-making. Experiments show that models use explicit timestamps effectively but still struggle with implicit temporal reasoning. Memory augmentation improves factual retrieval more than long-horizon decision-making, revealing a gap between accessing historical evidence and using it for clinical decision reasoning. Since recorded EHR actions are observational references rather than unique optimal decisions, future work should add expert-validated annotations and account for multiple valid next steps. Overall, LongMedBench supports the study of cross-visit memory, temporal inference, and the limitations of current medical agents in longitudinal EHR settings.

Acknowledgments. This work is supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2026C01021) and the Transvascular Implantation Devices Research Institute (TIDRI) (Grant No. KY052025008). We also acknowledge the contributors of MIMIC-IV [9] and the PhysioNet team for providing open access to this large-scale, de-identified clinical dataset.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Chhikara, P., Khant, D., Aryan, S., Singh, T., Yadav, D.: Mem0: Building production-ready ai agents with scalable long-term memory. arXiv preprint arXiv:2504.19413 (2025)
2. Cui, H., Unell, A., Chen, B., Fries, J.A., Alsentzer, E., Koyejo, S., Shah, N.H.: TIMER: Temporal instruction modeling and evaluation for longitudinal clinical records. npj Digital Medicine **8**(1), 577 (2025)
3. DeepSeek-AI: DeepSeek-V3.2: Pushing the frontier of open large language models (2025)
4. Ding, J., Lu, L., Ding, C., Bian, M., Chen, J., Pang, W., Chen, R., Peng, X., Lu, R., Ren, S., Zhu, G., Wu, X., Liu, Z., Zhang, R., Jiang, L., Han, B., Wang, Y., Xu, J.: MedBench v4: A robust and scalable benchmark for evaluating chinese medical language models, multimodal models, and intelligent agents (2025), https://arxiv.org/abs/2511.14439
5. Gong, E.J., Bang, C.S., Lee, J.J., Baik, G.H.: Knowledge-practice performance gap in clinical large language models: Systematic review of 39 benchmarks. Journal of medical Internet research **27**, e84120 (2025). https://doi.org/10.2196/84120
6. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., et al.: Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature medicine **30**(9), 2613–2622 (2024)

7. Hsieh, C.P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., Ginsburg, B.: RULER: What's the real context size of your long-context language models? (2024), https://arxiv.org/abs/2404.06654
8. Jiang, Y., Black, K.C., Geng, G., Park, D., Zou, J., Ng, A.Y., Chen, J.H.: MedAgentBench: A virtual ehr environment to benchmark medical llm agents. NEJM AI p. AIdbp2500144 (2025)
9. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV. PhysioNet (Oct 2024). https://doi.org/10.13026/kpb9-mt58, https://doi.org/10.13026/kpb9-mt58, version 3.1
10. Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The NarrativeQA reading comprehension challenge (2017), https://arxiv.org/abs/1712.07040
11. Lee, G., Hwang, H., Bae, S., Kwon, Y., Shin, W., Yang, S., Seo, M., Kim, J.Y., Choi, E.: EHRSQL: A practical text-to-sql benchmark for electronic health records. Advances in Neural Information Processing Systems **35**, 15589–15601 (2022)
12. Liao, Y., Jiang, S., Wang, Y., Wang, Y.: ReflecTool: Towards reflection-aware tool-augmented clinical agents (2024), https://arxiv.org/abs/2410.17657
13. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts (2023), https://arxiv.org/abs/2307.03172
14. Maharana, A., Lee, D.H., Tulyakov, S., Bansal, M., Barbieri, F., Fang, Y.: Evaluating very long-term conversational memory of llm agents. arXiv preprint arXiv:2402.17753 (2024)
15. Mehandru, N., Golchini, N., Bamman, D., Zack, T., Molina, M.F., Alaa, A.: ER-Reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. arXiv preprint arXiv:2505.22919 (2025)
16. OpenAI: GPT-5 mini model. https://developers.openai.com/api/docs/models/gpt-5-mini (2026), accessed 22 Feb 2026
17. OpenAI: text-embedding-3-small model. https://developers.openai.com/api/docs/models/text-embedding-3-small (2026), accessed 22 Feb 2026
18. Qiu, P., Wu, C., Liu, J., Zheng, Q., Liao, Y., Wang, H., Yue, Y., Fan, Q., Zhen, S., Wang, J., Gu, J., Wang, Y., Zhang, Y., Xie, W.: Evolving Interactive Diagnostic Agents in a Virtual Clinical Environment (2026), https://arxiv.org/abs/2510.24654
19. QwenTeam: Qwen3: Think deeper, act faster. https://qwen.ai/blog?id=qwen3 (2024), accessed: 22 Feb 2026
20. Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., Moor, M.: AgentClinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments (2025), https://arxiv.org/abs/2405.07960
21. Shen, Y., Huang, Z., Wang, Z., Tian, M., Guo, Z., Zhang, C., Zhou, S., Hu, Z., Li, D., Xu, J., Wang, K., Liu, W., Li, T., Yue, F., Hong, F., Liu, C., Zeng, K.: TRIP-Bench: A benchmark for long-horizon interactive agents in real-world scenarios (2026), https://arxiv.org/abs/2602.01675
22. Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.W., Yu, D.: LongMemEval: Benchmarking chat assistants on long-term interactive memory. https://arxiv.org/abs/2410.10813 (2024)
23. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 2369–2380 (2018)