MedLoCoMo: A Long-Context Multi-Session Medical Dialogue Benchmark for Large Language Models

Anonymous ACL submission

Abstract

`MedLoCoMo` is a **Med**ical **Lo**ng-**Co**ntext Me**mo**ry benchmark for patient-specific clinical reasoning over multi-admission medical dialogue. Existing medical QA benchmarks largely test short-context knowledge or single-document grounding, leaving open whether LLMs can use, connect, and abstain over longitudinal patient histories. We build `MedLoCoMo` from deidentified MIMIC-IV and MIMIC-IV-Note records by constructing admission-level clinical packets, synthesizing grounded doctor–patient conversations, and generating evidence-linked QA items over single-admission, cross-admission, and adversarial unanswerable settings. The benchmark contains 100 patient timelines averaging 1,669.8 turns, 29.7 sessions, and 74,512.2 tokens per conversation. Across the evaluated baselines, cross-admission reasoning is consistently harder than localized evidence use, even when models have long context windows or use external memory or retrieval methods. The anonymous code release is available at `https://anonymous.4open.science/status/MedLoCoMo-8C27` for review and reproducibility.

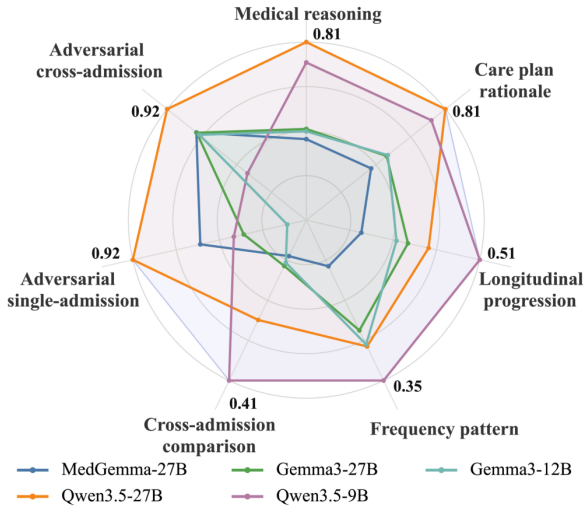


Figure 1: **Radar plot of selected model performance across question types in `MedLoCoMo`.** Axes report LLM-judge accuracy for answerable question types and abstention accuracy for adversarial types. Scores are shown on a 0–1 scale.

1 Introduction

Clinical reasoning is inherently longitudinal. Patient-specific questions often require evidence from prior admissions, discharge summaries, imaging reports, procedures, microbiology results, and evolving problem lists. These records form a heterogeneous sequence of clinical events rather than a single passage, making deidentified EHR resources such as MIMIC-IV (Johnson et al., 2023b) important for clinical NLP research.

The central evaluation gap is patient-level longitudinal reasoning. Exam-style and biomedical QA datasets such as MedQA, MedMCQA, and PubMedQA measure domain knowledge from compact questions, vignettes, or abstracts (Jin et al., 2020; Pal et al., 2022; Jin et al., 2019). EHR-grounded QA datasets such as emrQA and EHRNoteQA move closer to clinical documentation, but still mainly focus on single notes or bounded note contexts (Pampari et al., 2018; Kweon et al., 2024). They do not directly test whether LLMs can answer patient-specific questions whose evidence spans many admissions, or abstain when the longitudinal record lacks support.

This gap is challenging because longitudinal clinical QA stresses more than context length. Relevant evidence can be sparse, repeated, or updated across admissions, requiring models to decide which earlier events still matter and which later events change the patient's state. Prior long-context studies show degradation as contexts grow or relevant spans move away from salient prompt positions (Bai et al., 2024; Liu et al., 2024). In patient timelines, these failures appear as missed

Table 1: **Comparison of MedLoCoMo with long-context, dialogue, and clinical benchmark datasets.** Columns summarize medical domain coverage, dataset scale, average conversation length, session structure, token count, and temporal span where those fields are defined; MedLoCoMo combines medical grounding with long multi-session patient timelines for evaluation.

Dataset	Medical	# Conv./cases	Avg. turns/conv.	Avg. sessions/conv.	Avg. tokens/conv.	Time interval
LoCoMo (Maharana et al., 2024)	No	10	588.2	27.2	16,618.1	few months
MSC (Xu et al., 2022)	No	1,001	53.3	4	1,225.9	few days
MPChat (Ahn et al., 2023)	No	15,000	2.8	1	53.3	–
MedDialog-EN (Zeng et al., 2020)	Yes	257,332	2	1	173.0	–
ACI-Bench (Yim et al., 2023)	Yes	207	55	1	1,302	few minutes
MedLoCoMo (ours)	**Yes**	**100**	**1,669.8**	**29.7**	**74,512.2**	**few years**

cross-admission links, stale evidence use, or unsupported answers to plausible questions. A useful benchmark should therefore test both evidence use across encounters and abstention when no admission supports the requested fact (Rajpurkar et al., 2018). Figure 1 previews this variation across MedLoCoMo question types.

We introduce MedLoCoMo, a **Med**ical **Lo**ng-**Co**ntext Me**mo**ry benchmark for patient-specific clinical reasoning. Starting from deidentified MIMIC-IV structured data and MIMIC-IV-Note text (Johnson et al., 2023b,a), we construct admission-level clinical packets, synthesize grounded doctor–patient style conversations, and generate short-answer QA items with explicit evidence references. Each benchmark instance is a chronological patient timeline whose admissions are treated as sessions in one long conversation. MedLoCoMo is fixed before evaluation, so memory and retrieval systems are compared afterward on the same evidence-linked QA set. The resulting benchmark contains 100 patient conversations, averaging 1,669.8 turns, 29.7 sessions, and 74,512.2 tokens per conversation.

Our evaluation separates three sources of difficulty: *single-admission* questions answerable from one hospitalization, *cross-admission* questions requiring evidence from multiple admissions, and *adversarial unanswerable* questions that test abstention when plausible clinical questions lack support. Because long patient timelines also stress evidence preservation and updating, we include a memory-augmentation case study with retrieval methods as additional baselines. Figure 2 summarizes the pipeline; Appendices A–C provide additional dataset, evaluation, and prompt details.

Our contributions are:

- **A patient-level long-context clinical benchmark** built from deidentified EHR data and notes, with 100 longitudinal patient timelines averaging 1,669.8 turns and 74,512.2 tokens.
- **An evidence-linked QA design** that separates single-admission reasoning, cross-admission reasoning, and adversarial unanswerable questions.
- **A baseline study** of general-purpose and medical LLMs showing that, in the evaluated suite, model scale and domain specialization do not remove the cross-admission reasoning gap.
- **A memory-augmented case study** with retrieval baselines, showing that memory can improve abstention and aggregate QA Score but richer memory structure alone does not guarantee better answer extraction in this setting.

2 Related Work

Medical QA Benchmarks. Medical QA benchmarks largely measure domain knowledge in compact settings: MedQA (Jin et al., 2020) and MedMCQA (Pal et al., 2022) use exam-style questions, while PubMedQA (Jin et al., 2019) and BioASQ (Tsatsaronis et al., 2015) evaluate biomedical literature understanding. EHR-grounded QA moves closer to clinical documentation: emrQA constructs question-answer pairs from clinical notes (Pampari et al., 2018), and EHRNoteQA evaluates questions grounded in discharge summaries (Kweon et al., 2024). MedOdyssey targets medical long-context evaluation up to 200K tokens (Fan et al., 2024), but is organized around long medical inputs and retrieval-style stress tests rather than patient-level, multi-admission dialogue. In contrast, MedLoCoMo treats the patient as the unit of evaluation and tests evidence use across a longitudinal timeline.

Long-Context QA Evaluation. Long-context QA evaluation asks whether models can locate, retain, and combine evidence across thousands of tokens. LongBench (Bai et al., 2024), LongBench v2 (Bai et al., 2025), and L-Eval (An et al.,

2024) include long-document QA, summarization, dialogue understanding, and deeper reasoning tasks, while RULER probes controlled reasoning at different context lengths (Hsieh et al., 2024). Positional-robustness work shows that models can fail when evidence appears in less salient context positions (Liu et al., 2024), so longer prompts alone are not enough. Long patient histories also resemble long-term conversational memory problems: LoCoMo evaluates very long multi-session conversations (Maharana et al., 2024); MedMT-Bench tests memory, interference robustness, and safety defense in synthetic long multi-turn medical scenarios (Yang et al., 2026); MSC studies repeated interactions over time (Xu et al., 2022); and recent analysis suggests that multi-turn assumptions and errors can compound (Laban et al., 2026). `MedLoCoMo` follows this multi-session framing, but grounds each session in patient-specific EHR evidence.

Clinical Dialogue Generation. Medical dialogue corpora such as MedDialog (Zeng et al., 2020) and M^2-MedDialog (Yan et al., 2021) support healthcare conversation modeling, but online consultations are not typically linked to structured EHR timelines. ACI-Bench focuses on visit-note generation from encounter dialogue (Yim et al., 2023), while MedDialogRubrics evaluates multi-turn medical consultations with rubric-based assessment (Gong et al., 2026). MediLongChat synthesizes knowledge-guided, multi-encounter medical dialogues for in-dialogue, cross-dialogue, and synthesis reasoning (Hu et al., 2026); `MedLoCoMo` instead starts from deidentified EHR-derived admissions and emphasizes evidence-linked QA over long patient timelines. Prior work such as Self-Instruct shows that LLM-based generation can scale instruction-style data construction (Wang et al., 2023); in clinical benchmarking, this makes grounding, evidence references, and unsupported-answer cases essential.

How MedLoCoMo Differs. Table 1 positions `MedLoCoMo` against long-context, dialogue, and clinical benchmarks by medical relevance, scale, session structure, and context length. Across 100 conversations, `MedLoCoMo` averages 1,669.8 turns, 29.7 sessions, and 74,512.2 tokens; the longest conversation reaches 156,528 tokens and 64 sessions. Compared with short-context medical QA, single-note EHR QA, medical long-context document benchmarks, and consultation benchmarks, `MedLoCoMo` emphasizes evidence use across admissions, evolving clinical facts, and abstention when the record does not support an answer (Rajpurkar et al., 2018). Retrieval-augmented generation (Lewis et al., 2020), medical RAG benchmarks such as MIRAGE (Xiong et al., 2024), and clinical temporal information extraction benchmarks such as i2b2 (Sun et al., 2013) offer complementary baselines for evidence selection and event structuring, but they do not by themselves evaluate open-answer dialogue QA over complete patient timelines.

3 `MedLoCoMo` Generation Pipeline

`MedLoCoMo` is constructed through a staged pipeline that converts source EHR records into patient-level long-context QA instances. We first identify eligible longitudinal patient trajectories, then form admission-level clinical packets from notes and structured events. These packets support grounded conversation and summary synthesis, chronological patient-level aggregation, evidence-linked QA generation, and final quality-control checks. Figure 2 summarizes this process.

3.1 Source Data and Eligibility

`MedLoCoMo` is built from MIMIC-IV v3.1 and MIMIC-IV-Note v2.2, which provide deidentified structured EHR data and linked free-text clinical notes (Johnson et al., 2023b,a). The benchmark is organized at the patient level. Each patient is one benchmark instance, and each hospital admission is one session in that instance. Concatenating the sessions chronologically creates a multi-session, multi-turn long-context input.

Eligible admissions. An admission is eligible if it contains at least one discharge summary in MIMIC-IV-Note. We require discharge summaries because they provide the most complete narrative of the hospital course and serve as the backbone for grounded dialogue generation. Admissions without discharge notes are excluded.

Eligible patients. A patient is eligible if they have multiple eligible admissions. To support large-scale generation, we use a curated shortlist that ranks patients by the number of eligible admissions. This shortlist favors longitudinal trajectories with many encounters, which are especially useful for stress-testing long-context reasoning.

3.2 Clinical Packet Formation

For each eligible admission, we construct an *admission packet*: a structured JSON artifact that gathers

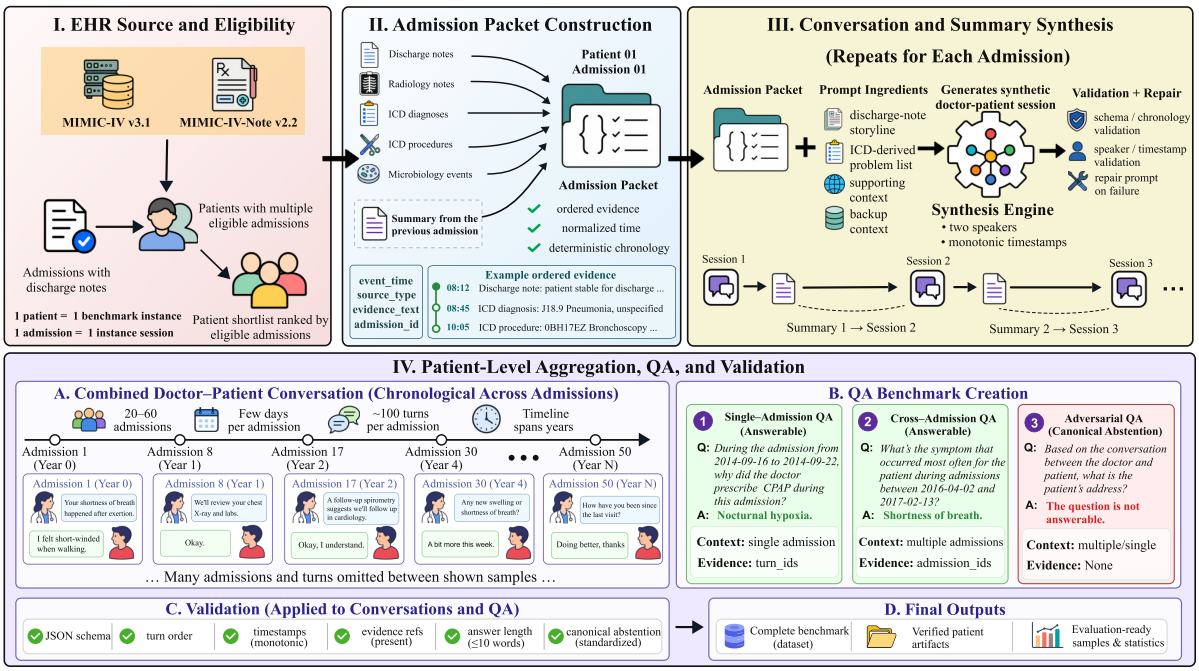


Figure 2: **End-to-end construction pipeline for `MedLoCoMo`.** Eligible MIMIC-IV and MIMIC-IV-Note records are converted into admission packets, used to synthesize grounded doctor–patient conversations and admission summaries, assembled into patient-level timelines, and converted into evidence-linked QA items for downstream evaluation and analysis.

the clinical evidence used for generation. Each packet contains:

- **Backbone narrative:** discharge notes for the admission, ordered by note sequence and time.
- **Supporting free text:** radiology notes linked to the same admission.
- **Structured clinical events:** admission-level ICD diagnoses, procedures, and microbiology entries.
- **Longitudinal context:** the generated summary from the previous admission, when available, to support cross-session continuity.

Timestamp handling for date-only fields and admission boundaries is described in Appendix A.5 for reproducible evaluation.

3.3 Conversation and Summary Synthesis

Given an admission packet, we prompt an LLM to generate a grounded doctor–patient style conversation for that admission. We enforce the following constraints through prompts and post-generation validation:

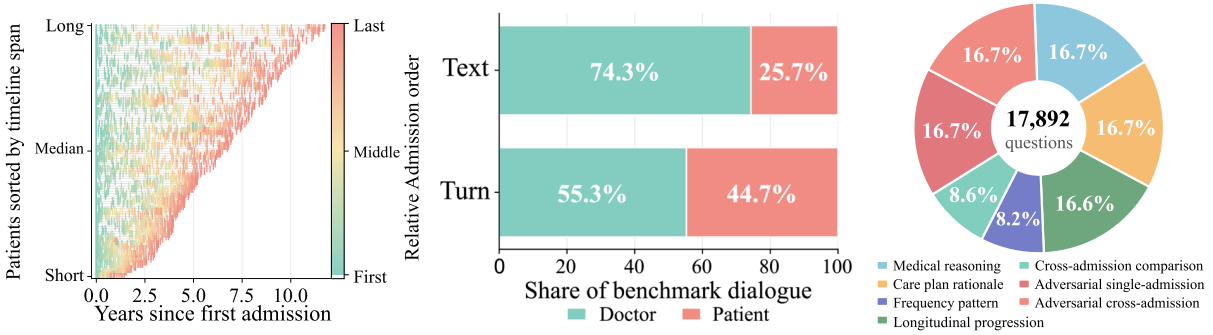
- Exactly two speakers: **Doctor** and **Patient**.
- The output is a list of timestamped turns with contiguous turn indices.
- Turn timestamps are monotonically increasing and fall within the admission boundaries.
- The discharge notes provide the backbone storyline; radiology, procedures, microbiology, and previous-admission summaries will provide the supporting context.

Per-admission summary. Alongside conversation lines, the model generates a structured admission summary containing admission boundaries, a hospital-course paragraph, and a problem list. The prompt requires the summary to be derived only from the generated conversation rather than directly from the packet, reducing direct reuse of note phrasing in downstream summary-based evaluation.

Chronological multi-session generation. Admissions are generated in chronological order. After each admission, its generated summary is passed to the next admission as lightweight longitudinal context. This mechanism encourages cross-session consistency, such as preserving chronic conditions or recurring problems, while avoiding repeated ingestion of the full prior dialogue.

3.4 Patient-Level Aggregation

After generating all eligible admissions for a patient, we assemble patient-level artifacts that define the long-context benchmark input. Appendix A describes the released artifact layout. The main



(a) Longitudinal Coverage **(b) Dialogue Composition** **(c) Question-Type Distribution**

Figure 3: **Benchmark overview for `MedLoCoMo`.** Panel (a) visualizes longitudinal coverage across patient admissions, panel (b) reports the doctor–patient composition of generated dialogue by text and turn share, and panel (c) shows the QA type mix spanning single-admission, cross-admission, and adversarial unanswerable questions.

patient-level artifacts are:

- **Combined conversation**: a chronological list of admission sessions, each containing its generated conversation lines.
- **Patient summary**: bookkeeping metadata and generation statistics, including processed versus eligible admissions, for reproducibility.

3.5 QA Generation

After patient-level timelines are assembled, the final generation stage converts them into evidence-linked short-answer QA items. We generate questions at two evidence scopes: *single-admission*, where the answer is supported by one session, and *cross-admission*, where the answer requires multiple sessions. Each QA item includes id, scope, question type, question, answer, and evidence. Answers are constrained to at most 10 words. Figure 3 summarizes the benchmark timeline, dialogue composition, and resulting distribution across clinical question types in the benchmark.

Single-admission QA. For each admission conversation, we generate three QA pairs: two answerable questions and one adversarial unanswerable question. Answerable questions are grounded in the admission conversation and cover medical reasoning and care plan rationale. The unanswerable question uses the adversarial type and stores the standard gold answer "the question is not answerable." Evidence cites the admission identifier and supporting turn identifiers.

Cross-admission QA. Cross-admission questions target patient-level reasoning across sessions. Answerable questions cover longitudinal progression, cross-admission comparison, and frequency pattern, and must cite evidence from at least two admissions. Adversarial cross-admission questions use the same standard abstention gold answer, but they draw on multiple admissions to create plausible unsupported longitudinal questions.

Human quality review. As the full benchmark is too large for complete manual inspection, we conducted a sampled human audit of about 50 generated admission conversations against their source EHR packets and about 300 QA items across single-admission, cross-admission, and adversarial settings. The review checked factual faithfulness of conversation content to the EHR-derived evidence, correctness and sufficiency of cited evidence, answerability for answerable QA, and true unanswerability for adversarial QA. All human-reviewed cases passed these checks. For the full benchmark, we apply automated quality-control checks before release; Appendix A provides details.

4 Experiment Results

4.1 Evaluation Protocol

`MedLoCoMo` evaluates open-answer QA over longitudinal patient conversations. We report results for two evidence scopes: *single-admission* questions, whose evidence is contained in one admission session, and *cross-admission* questions, whose evidence spans multiple sessions. For answerable questions, we report token-level F1 and LLM-judge accuracy (J). For adversarial unanswerable questions, we report abstention accuracy (Acc), which accepts normalized equivalent abstention phrasings rather than only the stored gold wording. We also report Score, a combined metric that aggregates J on answerable questions and Acc on adversarial

Table 2: **Baseline model results on MedLoCoMo.** F1 and LLM-judge accuracy (J) evaluate answerable questions, and adversarial accuracy (Acc) evaluates unanswerable questions. Score is the item-weighted combination of J on answerable items and Acc on adversarial items within the corresponding split; all values are percentages. In Score columns, **first**, second, and third mark summary ranks within each model group.

Model	Context Length	Overall				Single-admission				Cross-admission			
		F1	J	Acc	Score	F1	J	Acc	Score	F1	J	Acc	Score
General-Purpose LLMs													
GPT-5.1	400K	21.1	64.2	90.2	**72.8**	28.2	88.4	89.0	**88.6**	14.1	39.9	91.4	**57.0**
Qwen3.5-4B	256K	19.7	50.8	44.6	48.8	25.4	63.4	40.5	55.8	14.0	38.2	48.7	41.7
Qwen3.5-9B	256K	24.4	58.4	38.8	51.8	30.8	72.1	38.6	60.9	17.9	44.6	39.0	42.7
Qwen3.5-27B	256K	26.1	56.1	92.3	68.1	40.4	80.8	92.3	84.7	11.8	31.3	92.3	51.6
Gemma3-4B	128K	10.3	18.6	0.7	12.6	9.7	21.4	0.2	14.3	11.0	15.9	1.1	10.9
Gemma3-12B	128K	17.5	33.3	40.5	35.7	20.7	43.8	10.1	32.6	14.2	22.7	70.9	38.8
Gemma3-27B	128K	19.4	33.9	53.0	40.3	24.2	44.0	33.3	40.5	14.6	23.8	72.7	40.1
Medical-Specialized LLMs													
MedGemma-4B	128K	12.2	22.7	9.8	18.4	12.0	28.7	7.9	21.8	12.3	16.7	11.7	15.0
MedGemma-27B	128K	10.2	25.1	64.6	**38.2**	14.9	37.2	56.4	**43.6**	5.6	12.9	72.9	**32.9**
MedMO-8B	256K	11.5	18.0	2.2	12.7	11.4	19.9	0.0	13.3	11.7	16.1	4.4	12.2
MedMO-4B	256K	8.0	12.7	0.4	8.6	7.5	14.8	0.0	9.8	8.4	10.7	0.9	7.4
MediPhi	128K	8.0	13.9	0.0	9.2	8.0	15.6	0.0	10.4	7.9	12.1	0.0	8.1
Lingshu-32B	128K	9.8	24.8	0.0	16.5	9.4	27.5	0.0	18.4	10.3	22.0	0.0	14.7

questions, weighted by the number of items in the corresponding split. Score is intended only as a compact relative summary for a split; it should not be read as overall model competence because high abstention can raise Score even when answer extraction remains weak. We therefore interpret F1, J, Acc, and Score together. Tables 2 and 3 report the component metrics separately, and Appendix Tables 5–6 provide question-type breakdowns. All table values are percentages. Appendix B.1 gives the metric formulas, and Appendix B.2 gives the full judge prompt.

4.2 Models Evaluated

We evaluate two groups of non-memory baselines. The general-purpose group includes GPT-5.1, Qwen3.5 models at 4B, 9B, and 27B scales, and Gemma3 models at 4B, 12B, and 27B scales. The medical/specialized group includes MedGemma-4B, MedGemma-27B, Lingshu-32B, MedMO-8B, MedMO-4B, and MediPhi. We treat MedGemma as the medical-specialized counterpart to Gemma3, which allows a direct comparison between a general model family and its medical variant.

4.3 General Model Trends

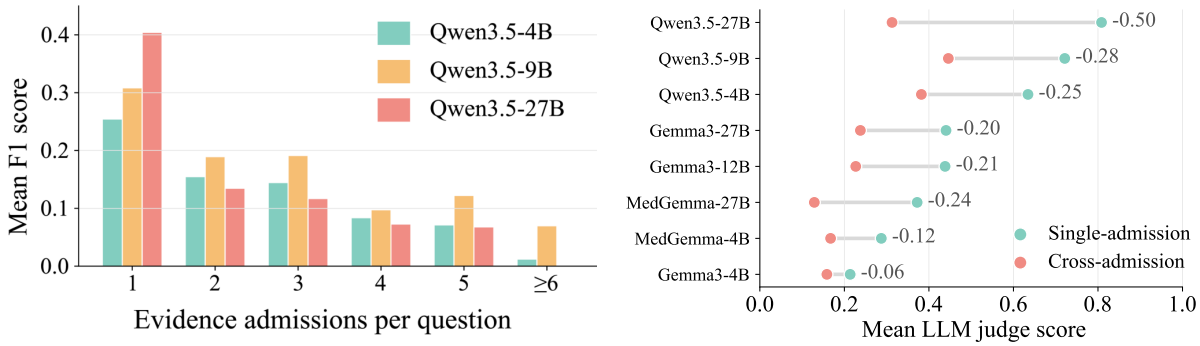
Among general-purpose models, Table 2 shows that GPT-5.1 has the strongest Score across all three summary columns, driven by high judged accuracy on single-admission QA and reliable abstention. The Qwen family still exposes the benchmark's scope effect: Qwen3.5-27B has the highest single-admission F1 and strongest abstention, but its judged answer quality drops sharply once evidence spans multiple admissions, while Qwen3.5-9B is stronger on cross-admission answerable metrics. This explains why Score should be read with its components rather than as a standalone ranking.

Gemma3 improves with scale mainly by abstaining more reliably. Its Score rises partly because Acc improves, but even the largest Gemma3 model remains below Qwen on judge-based answer quality, separating better abstention from stronger longitudinal evidence extraction. Appendix B.3 provides full question-type breakdowns.

4.4 Medical and Specialized Models

The medical/specialized models do not consistently outperform the strongest general-purpose baselines. MedGemma-27B has the highest group-level Score and Acc, but this mainly reflects abstention and local behavior rather than stronger cross-admission answer extraction. On cross-admission answerable questions, it remains weak despite strong adversarial abstention accuracy.

Other specialized models show complementary weaknesses. MedGemma-4B has the best specialized-model F1 overall and on cross-admission questions, while Lingshu-32B has the highest specialized-model cross-admission J but



(a) Evidence Scope Degradation **(b) Single–Cross Admission Gap**

Figure 4: **Answerable-question breakdowns on `MedLoCoMo`.** Panel (a) shows how Qwen F1 changes as supporting evidence spans more admissions, isolating the effect of evidence scope. Panel (b) compares single-admission and cross-admission LLM-judge scores for evaluated models, with labels reporting the cross-minus-single gap.

fails on adversarial abstention. MedMO and MediPhi remain low across both answer quality and abstention, suggesting that domain specialization alone does not resolve the benchmark's longitudinal evidence-use challenge in this model set.

4.5 Memory-Augmented Evaluation

Motivation. We focus memory augmentation on Gemma3 because its 128K token context is the main setting where `MedLoCoMo` can exceed the usable full-history prompt budget. Qwen3.5 models provide longer 256K token windows, while the longest `MedLoCoMo` timelines can approach or exceed the Gemma3 budget once questions and evaluation instructions are added. The Gemma3 setting therefore tests whether external memory can preserve useful longitudinal evidence when direct full-context prompting would otherwise require truncating the full patient history.

Memory methods and retrieval baselines. We evaluate three memory methods with Gemma3 models. *mem0* follows the long-term memory framework of Mem0 (Chhikara et al., 2025), which maintains an external memory store while sequentially processing conversation history. *mem0** uses the same overall memory interface but restricts the policy to write-only memory additions, and *memα* uses the multi-component memory structure introduced by Mem-α (Wang et al., 2025). To contextualize the memory results, we also include additional baseline results from BM25 (Robertson and Zaragoza, 2009) and Embedding retrieval based on retrieval-augmented generation (Lewis et al., 2020). Appendix B.4 summarizes the evaluated memory and retrieval methods.

Results. Table 3 shows that memory augmentation improves Gemma3 mainly through better abstention and, for larger models, stronger judged answer quality. The Score summaries also rise, but much of that movement comes from Acc, so the component metrics are essential for interpretation. The retrieval baselines provide a useful check on whether the gains come only from shortening the input: they improve over plain Gemma3, but remain below the strongest memory variants. In this Gemma3 augmentation study, the remaining gap suggests that `MedLoCoMo` stresses evidence selection and cross-admission linking, not only context-window length. Among memory methods, the highlighted summary cells favor mem0 and mem0*, while the component metrics show a mixed pattern: mem0* is stronger for single-admission behavior, and mem0 is more favorable on cross-admission questions. The more structured memα variant remains competitive but does not dominate, indicating that richer memory organization is not automatically beneficial without task alignment. Appendix B.3 provides the full question-type breakdown table results.

Implications for memory design. The main design implication is that memory should support evidence selection and cross-session linking, not just compress long histories into shorter prompts. Score is useful as a summary, but it should be read with F1/J/Acc and question-type results because stronger abstention can raise Score while answer extraction remains weak. Constrained variants such as mem0* and simple retrieval baselines are therefore important ablations: they test whether added memory actions and structure improve benchmark

Table 3: **Memory-augmented results for Gemma3 models on `MedLoCoMo`, with BM25 and Embedding included as additional baseline results.** The table uses the same percentage metrics as the main baseline evaluation: F1 and J for answerable questions, and Acc for adversarial unanswerable questions. Score is the item-weighted combination of J on answerable items and Acc on adversarial items within the corresponding split. In Score columns, **first**, second, and third mark summary ranks across rows.

Context Method	Model	Overall				Single-admission				Cross-admission			
		F1	J	Acc	Score	F1	J	Acc	Score	F1	J	Acc	Score
mem0	Gemma3-4B	16.3	32.3	20.2	28.2	17.3	37.8	10.6	28.7	15.3	26.8	29.7	27.8
mem0	Gemma3-12B	19.8	42.6	65.0	50.1	25.4	52.1	49.6	51.2	14.2	33.1	80.4	48.9
mem0	Gemma3-27B	18.3	42.6	87.5	**57.6**	25.4	52.5	87.7	64.2	11.3	32.6	87.5	**50.9**
mem0*	Gemma3-4B	13.1	22.2	23.8	22.8	12.7	23.1	11.7	19.3	13.5	21.4	35.9	26.2
mem0*	Gemma3-12B	16.9	35.2	66.6	45.7	21.1	39.3	50.2	43.0	12.8	31.2	83.0	48.5
mem0*	Gemma3-27B	18.3	41.1	90.6	**57.6**	24.9	52.0	91.4	**65.1**	11.7	30.3	89.9	50.1
memα	Gemma3-4B	11.0	12.6	24.8	16.6	10.1	12.6	20.9	15.3	11.9	12.6	28.6	17.9
memα	Gemma3-12B	14.3	27.8	72.0	42.5	16.5	29.7	62.5	40.7	12.0	25.8	81.5	44.4
memα	Gemma3-27B	14.7	33.5	89.4	52.1	18.0	38.0	89.0	55.0	11.4	29.0	89.9	49.3
BM25	Gemma3-4B	12.9	26.9	0.8	18.2	14.8	35.5	0.2	23.7	11.0	18.3	1.3	12.6
BM25	Gemma3-12B	20.4	40.1	37.4	39.2	25.9	53.2	8.6	38.3	14.8	27.0	66.3	40.1
BM25	Gemma3-27B	21.2	41.3	52.2	44.9	27.8	55.1	33.5	47.9	14.6	27.5	70.9	42.0
Embedding	Gemma3-4B	12.2	23.6	1.1	16.1	13.3	29.6	0.4	19.9	11.2	17.6	1.8	12.3
Embedding	Gemma3-12B	18.7	37.2	38.8	37.7	23.0	47.4	8.6	34.4	14.3	27.0	68.9	41.0
Embedding	Gemma3-27B	19.8	37.5	56.0	43.6	24.5	46.7	39.4	44.3	15.1	28.2	72.5	43.0

behavior beyond simpler context selection. Richer structures such as memα may still be useful, but the results suggest they need task-aligned training or tuning to translate structure into stronger cross-admission clinical reasoning.

4.6 Unanswerable Behavior and Cross-Admission Difficulty

Adversarial unanswerable performance is highly uneven. Qwen3.5-27B is the strongest abstainer across evidence scopes, and larger Gemma3 and MedGemma variants also improve on this dimension. However, abstention does not imply answer extraction: MedGemma-27B leads specialized models on Acc while trailing other specialized models on cross-admission answer extraction.

Figure 4 shows why cross-admission QA is the central stress test in `MedLoCoMo`. Qwen F1 drops as evidence spans more admissions, especially for Qwen3.5-27B: the strongest single-admission model loses its advantage as scope expands. Every displayed model also has lower cross-admission J than single-admission J, with the largest gap again for Qwen3.5-27B. Together, Table 2 and Figure 4 show that the evaluated models use local evidence or abstain more reliably than they ground answers in evidence distributed across encounters.

This separation matters because a useful long-context clinical system should extract local facts, connect them across encounters, and reject unsupported questions without letting one behavior substitute for the others.

5 Conclusion

We introduced `MedLoCoMo`, a patient-level benchmark for evaluating long-context LLMs on longitudinal clinical reasoning. Unlike short-context medical QA or aggregate long-context scores, `MedLoCoMo` turns multi-admission EHR-derived timelines into evidence-linked questions that separately test local evidence use, cross-admission integration, and principled abstention when the record does not support an answer. This structure makes the benchmark useful for diagnosing where a system fails: whether it loses relevant details in long inputs, cannot connect evidence across encounters, or improves apparent performance by abstaining more often. In the evaluated suite, larger context windows, medical specialization, retrieval baselines, and external memory each help in specific ways, but none eliminates the gap between local evidence use and cross-admission reasoning. `MedLoCoMo` therefore provides a focused testbed for measuring whether long-context clinical systems can reliably link evidence across patient timelines.

Limitations

MedLoCoMo has several limitations. First, it uses a text-only design: generated doctor–patient conversations do not include clinical images or other non-text media, so radiology evidence is represented through reports rather than image pixels. Second, the conversations and QA items are synthetic transformations of deidentified EHR evidence. This makes large-scale, evidence-linked evaluation possible, but it may introduce artifacts in patient phrasing, clinical dialogue flow, and question style that differ from real patient–clinician interactions. Third, the benchmark intentionally selects patients with many eligible admissions to stress long-context reasoning, which can overrepresent admission-rich trajectories and underrepresent shorter or less documented care histories. Future work should develop MedLoCoMo beyond this text-only setting, add clinician-facing audits of generated dialogue and QA, and broaden patient sampling while preserving the benchmark's focus on longitudinal evidence use.

Ethics Statement

MedLoCoMo is intended as a research benchmark for evaluating long-context clinical reasoning, not as a clinical decision-support system or patient-facing tool. The released MedLoCoMo benchmark contains synthetic doctor–patient conversations, summaries, QA items, and evaluation metadata, and is downloadable for research review; it does not redistribute raw MIMIC-IV or MIMIC-IV-Note tables or notes. In the code release, the folders reserved for MIMIC-IV and MIMIC-IV-Note inputs are empty placeholders. Users who want to rerun the construction pipeline must obtain credentialed access to the underlying MIMIC resources, comply with their data-use requirements, and place the authorized files locally in those folders. Although the source records are deidentified and the released conversations are synthetic, the benchmark is derived from real clinical documentation; users should treat it as sensitive medical data, avoid reidentification attempts, and interpret model outputs as benchmark behavior rather than clinical advice.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint*.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2024. MedOdyssey: A medical domain benchmark for long context evaluation up to 200k tokens. *arXiv preprint*.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, Jinjie Gu, and Junwei Liu. 2026. MedDialogRubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Hebin Hu, Renke Dai, Ah-Hwee Tan, and Yilin Kang. 2026. Synthesis and evaluation of long-term history-aware medical dialogue. *arXiv preprint*. Accepted by AAMAS 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

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.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023b. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. EHRNoteQA: An LLM benchmark for real-world clinical practice using discharge summaries. *arXiv preprint arXiv:2402.16040*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*. Oral presentation.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

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.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and 3 others. 2015. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. 2025. Mem-α: Learning Memory Construction via Reinforcement Learning. *arXiv preprint*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhumin Chen, Zhaochun Ren, and Huasheng Liang. 2021. M^2-MedDialog: A dataset and benchmarks for multi-domain multi-service medical dialogues. *arXiv preprint*.

Lin Yang, Yuancheng Yang, Xu Wang, Changkun Liu, and Haihua Yang. 2026. MedMT-Bench: Can LLMs memorize and understand long multi-turn conversations in medical scenarios? *arXiv preprint*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. ACI-Bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

A Details of MedLoCoMo

MedLoCoMo is released as a patient-level benchmark directory rather than as a single flat file. This appendix describes the released artifact layout and the fields needed to run or audit evaluation. Section 3 describes how the artifacts are generated, and Appendix C provides detailed prompt templates and concrete examples.

Generation model. Conversation, admission-summary, and QA artifacts are generated with Qwen3-235B-A22B-Instruct-2507.[1] This model is used only for benchmark construction. It is separate from the evaluated models and belongs to a different Qwen model family than the Qwen3.5 baselines used throughout the main evaluation suite.

A.1 Generation Prompt Structure

The conversation-generation prompt presents clinical evidence in three progressively richer blocks: a discharge-note storyline, an ordered problem list derived from ICD diagnosis titles, and a supporting-context block containing truncated radiology notes and selected procedure and microbiology entries. The implementation includes up to 8 radiology notes and up to 10 procedure and microbiology entries to keep prompts bounded while preserving diverse clinical evidence. The full raw packet is included only as backup context. The prompt also provides a recommended turn-count range based on length of stay, allowing longer admissions to produce richer conversations while keeping outputs predictable. Appendix C gives the full generation prompt templates and examples.

[1] Hugging Face model page.

A.2 Released Dataset Layout

The benchmark release is organized hierarchically. The benchmark root contains one folder per patient, named by patient identifier. Each patient folder contains patient-level files and one subfolder per processed admission. Admission folders are named by admission identifier and contain the generated conversation, admission summary, admission-level QA, and audit artifacts for that admission. At a high level, the release contains:

- Patient-level evaluation files: combined conversation, patient summary, benchmark QA, and cross-admission QA.
- Admission-level evaluation files: conversation, summary, and QA.
- Admission-level audit files: formed packet and prompt record, which support traceability but are not required for model evaluation.

A.3 Patient-Level Files

The combined conversation file stores the full patient timeline as a chronological list of admission sessions. Each session records the admission identifier, admission start and end times, and generated conversation lines. This is the primary long-context input for full-history model evaluation.

The patient summary file records patient-level metadata, including eligible and processed admission counts, processed admission identifiers, LLM call statistics, total and mean conversation turns, total and mean token counts, and tokenizer name. The benchmark QA file contains the shuffled final QA set for that patient, while the cross-admission QA file keeps the cross-session subset separate for split-level analysis.

A.4 Admission-Level Files

Each admission subfolder contains the generated conversation, summary, and QA for one hospital stay. The conversation consists of ordered turns with turn number, timestamp, speaker, and text. The summary contains the admission window, a one-paragraph hospital-course summary, and a problem list. The admission-level QA file contains the two answerable single-admission questions and

one adversarial unanswerable question generated for the target admission.

The formed packet and prompt record are included for reproducibility and auditing. They are not required for model evaluation, but they make it possible to trace how a generated admission session was constructed from the underlying evidence and how the generation request was issued.

A.5 Timestamp Normalization

MIMIC contains both precise timestamps and date-only fields. We normalize times to make ordering deterministic and validation reproducible. Date-only procedure and microbiology events are normalized to noon. Diagnoses often lack reliable within-stay timing, so we assign pseudo-timestamps near discharge, anchored to the end of the stay. If discharge time is unavailable, admission end falls back to the latest discharge-note chart time or store time. These rules preserve temporal order without claiming unsupported precision. Accordingly, QA construction avoids fine-grained within-stay temporal comparisons whose answer would depend on normalized pseudo-times; temporal questions rely on admission order or explicit conversation evidence in the dialogue.

A.6 QA and Evidence Format

Each QA item contains id, scope, question type, question, answer, and evidence. Single-admission answerable questions cover medical reasoning and care plan rationale; cross-admission questions cover longitudinal progression, cross-admission comparison, and frequency pattern. Evidence cites supporting turns for single-admission items and the required admissions for cross-admission items. Adversarial items use plausible but unsupported questions with the stored gold answer "the question is not answerable."

Adversarial questions are designed to be close to regular answerable questions rather than visibly out of scope. They preserve the clinical topic and reasoning frame while minimally changing the question so the record no longer supports an answer, forcing models to distinguish semantic similarity from actual evidence support.

A.7 Benchmark QA Statistics

The full benchmark contains 17,892 QA items across 100 patients, or 178.9 questions per patient on average. Of these, 11,928 are answerable and 5,964 are adversarial unanswerable. Single-admission QA contributes 8,946 items, and cross-admission QA contributes another 8,946 items. Table 4 reports the benchmark-level evidence-span distribution. The one-admission row corresponds to single-admission QA, while the remaining rows quantify how cross-admission questions are distributed across two, three, or at least four admissions in total.

Table 4: **Evidence-admission span distribution in `MedLoCoMo`.** Percent of all QA is computed over all 17,892 questions; percent of cross-admission QA is computed over the 8,946 questions whose evidence spans multiple patient admissions.

Evidence admissions	Questions	All QA	Cross QA
1	8,946	50.00%	–
2	5,282	29.52%	59.04%
3	2,497	13.96%	27.91%
4+	1,167	6.52%	13.04%

A.8 Quality Control Checks

Because the full benchmark is too large for complete manual inspection, the sampled human audit in Section 3 is complemented by automated checks over the released artifacts. The generation pipeline applies these checks before patient-level artifacts are released. Conversation checks enforce grounded clinical facts from the EHR-derived packet, valid schema, exact speaker labels, contiguous turns, nonempty text, monotonic timestamps, and admission-window consistency. QA checks validate each answer with its cited evidence: single-admission items must cite supporting turns from the target admission, cross-admission items must cite valid patient admissions, and adversarial items must use the standard abstention gold answer in the released QA file. QA construction also avoids fine-grained within-stay temporal comparisons that would depend on normalized pseudo-times. Together, these checks keep generated dialogue and evidence-linked answers traceable to the admission packets.

B Additional Experimental Details and Results

B.1 Metric Construction

Let S denote an evaluation split such as overall, single-admission, or cross-admission. We write A_S for the answerable questions in the split and U_S for adversarial unanswerable questions. All

table values are reported as percentages in every reported result table.

Answerable F1. For each answerable question i, the prediction $\hat{a}_i$ and gold answer a_i are normalized by lowercasing, removing punctuation, removing articles, and collapsing whitespace. Let $T(\cdot)$ be the resulting multiset of tokens. Whole-answer precision, recall, and F1 are:

$$P_i = \frac{|T(\hat{a}_i) \cap T(a_i)|}{|T(\hat{a}_i)|}, \quad R_i = \frac{|T(\hat{a}_i) \cap T(a_i)|}{|T(a_i)|},$$

$$F_i^{\text{whole}} = \begin{cases} \frac{2P_iR_i}{P_i+R_i}, & P_i + R_i > 0, \\ 0, & P_i + R_i = 0. \end{cases}$$

When either answer contains comma-separated spans, the evaluator also computes a comma-aware score. For prediction parts $\{\hat{p}_{ik}\}$ and gold parts $\{g_{i\ell}\}$,

$$P_i^{\text{comma}} = \frac{1}{K} \sum_k \max_\ell F^{\text{whole}}(\hat{p}_{ik}, g_{i\ell}),$$

$$R_i^{\text{comma}} = \frac{1}{L} \sum_\ell \max_k F^{\text{whole}}(\hat{p}_{ik}, g_{i\ell}),$$

and F_i^{comma} is the harmonic mean of P_i^{comma} and R_i^{comma}. The reported per-question F1 is

$$F1_i = \max(F_i^{\text{whole}}, F_i^{\text{comma}}),$$

with F_i^{whole} used when no comma-aware score is available. Split-level F1 is the mean over answerable questions:

$$F1_S = \frac{1}{|A_S|} \sum_{i \in A_S} F1_i.$$

LLM-judge accuracy. For each answerable question, the judge returns $j_i \in \{0, 1\}$. The LLM-judge score reported as J is

$$J_S = \frac{1}{|A_S|} \sum_{i \in A_S} j_i.$$

Adversarial abstention accuracy. Adversarial gold answers use the canonical phrase "the question is not answerable," but scoring is not exact-string. Let $\mathcal{C}_{\text{abs}}$ be the accepted set of normalized abstention phrasings, including common equivalents such as "not answerable from the record," "cannot be determined," and "not mentioned." After case folding, punctuation removal, article removal, and whitespace normalization,

$$\text{Acc}_i = \mathbf{1}\{\text{norm}(\hat{a}_i) \in \mathcal{C}_{\text{abs}}\}.$$

The split-level adversarial accuracy is

$$\text{Acc}_S = \frac{1}{|U_S|} \sum_{i \in U_S} \text{Acc}_i.$$

Combined Score. The combined Score uses LLM-judge accuracy for answerable questions and abstention accuracy for adversarial questions:

$$\text{Score}_S = \frac{\sum_{i \in A_S} j_i + \sum_{i \in U_S} \text{Acc}_i}{|A_S| + |U_S|}.$$

B.2 LLM Judge Prompt

For answerable predictions, we grade all candidate answers with a single fixed LLM judge, Gemini 3 Flash Preview, using the model code `gemini-3-flash-preview`.[2] Before final evaluation, we tested several judge choices and found completely consistent binary decisions together with human review, which is expected because the task is a constrained 0/1 correctness judgment. We therefore use one fixed judge for consistency across all reported results. The judge is used only for binary grading after model answers are produced; it is not used to generate `MedLoCoMo` or to answer benchmark questions. Its decisions were checked by human review for consistency with the grading rubric. Adversarial questions are scored directly by the abstention-phrase matcher above rather than by the judge. The judge prompt structure is:

Answerable QA Judge Prompt

```
=== System Message ===
You are grading candidate answers for short-answer
medical benchmark questions.

Judge only from the provided question, gold_answer,
and candidate_answer.

Score 1 when the candidate answer is correct.

Score 0 when the candidate answer is false, incorrect,
unsupported, incomplete enough to be wrong, or only
says it is not answerable.

Return strict JSON with the schema
{"judgments": [{"qa_id": "...", "score": 1}]}.

Each score must be exactly one of: 0, 1.
Return exactly one judgment per provided qa_id.

=== User Message ===
{
  "items": [
    {
      "qa_id": "...",
      "question": "...",
      "gold_answer": "...",
      "candidate_answer": "..."
    }
  ]
}
```

[2] Google AI for Developers model documentation.

B.3 Question-Type Results

Tables 5 and 6 break down results by question type. For answerable regular question types, each block reports token-level F1 and LLM-judge accuracy (J). For adversarial question types, the tables report abstention accuracy (Acc). All values are reported as one-decimal percentage values.

Table 5 shows that GPT-5.1 leads the local answerable judge scores among general-purpose models and is also strong on adversarial abstention. Qwen models remain strongest on several longitudinal question types: Qwen3.5-9B leads longitudinal progression, cross-admission comparison, and frequency pattern, while Qwen3.5-27B leads adversarial abstention. Medical-specialized models are weaker overall, but Lingshu-32B is comparatively strong on longitudinal and cross-admission type judge scores, while MedGemma-27B is the strongest medical model on adversarial abstention accuracy.

Table 6 shows that memory augmentation helps most clearly through adversarial Acc and selected Gemma3 answerable J scores, with BM25 and Embedding included as additional baseline results. The strongest memory cells are concentrated in mem0 and mem0*, while memα remains competitive for Gemma3-27B on adversarial Acc but does not consistently dominate. The retrieval baselines are competitive on some local or pattern-oriented judge scores, but they remain weaker on the adversarial and longitudinal columns where the best memory methods are strongest, reinforcing that the benchmark stresses cross-admission evidence selection and linking rather than only context-window length by itself.

B.4 Memory Methods and Retrieval Baselines

Memory construction. The memory-augmented experiments use the same QA evaluation protocol and metrics as the main baseline results. For each patient, the conversation is processed chronologically by admission-local chunks. Each chunk is passed to a memory-construction call together with a compact rolling patient summary and a small window of recent chunk summaries. The rolling summary is capped at 1,200 characters and is used only to preserve longitudinal context for later memory extraction; answer prompts are built from retrieved memories rather than rereading the rolling patient summary during answer generation.

Store size and update policy. The memory store is data-dependent rather than fixed to a prespecified final size. Each chunk can propose only a bounded number of candidate memories, but the final store may grow or shrink depending on the update policy; every run logs total, active, deleted, duplicate-skipped, added, updated, and no-op memory counts. In *mem0*, the extractor first proposes atomic timestamp-aware memories. A dense local retriever then finds similar active memories, and an update call chooses among ADD, UPDATE, DELETE, and NOOP actions. Invalid or missing update decisions fall back to ADD, while DELETE and UPDATE must target an existing active memory. In *mem0**, the same extraction interface is used but the policy is write-only: nonduplicate candidates are appended as new memories and exact normalized duplicates are skipped, so there are no delete or merge decisions. In *mem*α, extracted memories are also add-only but are typed into core, episodic, and semantic stores following the Mem-α organization (Wang et al., 2025). This separates durable patient facts, timestamped events, and reusable patient-specific clinical relations while keeping the same downstream evaluation pipeline.

Answer-time retrieval and budgets. For each benchmark question, memory methods query the active memory store using the question text. The dense retriever returns the top ranked memories, the answer prompt renders at most the configured maximum number of memories, and retrieved memories are shown in chronological order when visible timestamps are available. If the estimated prompt exceeds the effective budget, defined by model context length minus answer-generation tokens and a safety margin, the lowest-ranked memories are pruned until the answer prompt fits.

Retrieval baselines. The retrieval baselines use the same Gemma3 backbones but do not maintain a memory store. They build one document per admission from the generated doctor–patient conversation, score all admissions for each question, render the selected admissions chronologically, and prune lower-ranked admissions if the prompt exceeds the same token budget. BM25 uses lexical retrieval with $k_1 = 1.5$ and $b = 0.75$ (Robertson and Zaragoza, 2009). Embedding uses the same local dense Hugging Face retrieval stack as the memory retriever in a retrieval-augmented generation setup (Lewis et al., 2020).

Table 5: **Baseline question-type results on MedLoCoMo.** MR = medical reasoning, CPR = care plan rationale, LP = longitudinal progression, CAC = cross-admission comparison, FP = frequency pattern, Single Adv. = single-admission adversarial, and Cross Adv. = cross-admission adversarial. For regular question types, F1 and J evaluate answerable QA; for adversarial question types, Acc evaluates abstention. Colored J and Acc cells mark top-three ranks within each model group: **first**, second, and third place.

Model	Context Length	MR		CPR		LP		CAC		FP		Single Adv.	Cross Adv.
		F1	J	F1	J	F1	J	F1	J	F1	J	Acc	Acc
General-Purpose LLMs													
GPT-5.1	400K	31.2	**87.2**	25.2	**89.7**	16.8	45.1	9.8	35.7	13.3	34.1	89.0	91.4
Qwen3.5-4B	256K	27.1	61.7	23.8	65.2	18.9	46.5	10.2	31.5	8.3	29.3	40.5	48.7
Qwen3.5-9B	256K	34.3	71.6	27.4	72.7	22.4	**51.5**	12.8	**41.2**	14.6	**34.9**	38.6	39.0
Qwen3.5-27B	256K	45.2	80.8	35.7	80.8	15.7	36.3	6.8	25.6	9.7	27.5	**92.3**	**92.3**
Gemma3-4B	128K	9.0	18.9	10.3	23.8	15.1	22.2	8.3	7.1	5.8	12.7	0.2	1.1
Gemma3-12B	128K	22.2	40.3	19.1	47.4	17.4	26.8	8.5	10.9	14.1	27.1	10.1	70.9
Gemma3-27B	128K	23.9	41.4	24.5	46.7	18.6	30.2	8.5	11.8	13.1	24.0	33.3	72.7
Medical-Specialized LLMs													
MedGemma-4B	128K	12.0	29.5	12.0	28.0	16.4	21.8	10.2	11.8	6.6	12.2	7.9	11.7
MedGemma-27B	128K	15.2	**36.8**	14.6	**37.7**	7.1	16.3	5.1	9.2	3.1	10.0	**56.4**	**72.9**
MedMO-8B	256K	11.1	18.5	11.6	21.4	14.8	18.1	9.3	13.0	8.1	15.3	0.0	4.4
MedMO-4B	256K	6.2	15.0	8.8	14.5	11.0	14.1	7.8	7.1	4.0	7.9	0.0	0.9
MediPhi	128K	7.8	14.8	8.2	16.5	9.7	13.2	6.5	11.3	5.9	10.9	0.0	0.0
Lingshu-32B	128K	9.0	28.2	9.8	26.9	13.6	**25.4**	8.3	**18.5**	6.2	**19.2**	0.0	0.0

Table 6: **Memory-augmented question-type results on MedLoCoMo, with BM25 and Embedding included as additional baseline results.** MR = medical reasoning, CPR = care plan rationale, LP = longitudinal progression, CAC = cross-admission comparison, FP = frequency pattern, Single Adv. = single-admission adversarial, and Cross Adv. = cross-admission adversarial. For regular question types, F1 and J evaluate answerable QA; for adversarial question types, Acc evaluates abstention. Colored J and Acc cells mark top-three ranks across rows: **first**, second, and third place.

Context Method	Model	MR		CPR		LP		CAC		FP		Single Adv.	Cross Adv.
		F1	J	F1	J	F1	J	F1	J	F1	J	Acc	Acc
mem0	Gemma3-4B	20.1	42.7	14.5	32.8	19.5	32.9	9.8	16.4	13.1	25.8	10.6	29.7
mem0	Gemma3-12B	30.8	**59.9**	19.9	44.3	17.3	39.7	10.5	**27.7**	12.0	26.2	49.6	80.4
mem0	Gemma3-27B	31.3	56.6	19.5	48.5	16.5	**45.8**	5.2	18.9	7.5	21.4	87.7	87.5
mem0*	Gemma3-4B	15.7	30.6	9.7	15.6	17.3	27.4	8.9	10.1	11.0	21.4	11.7	35.9
mem0*	Gemma3-12B	26.3	48.2	15.9	30.4	15.5	38.8	9.6	19.8	10.8	28.4	50.2	83.0
mem0*	Gemma3-27B	30.6	57.1	19.2	46.9	17.4	41.7	5.0	16.8	7.8	22.3	**91.4**	**89.9**
memα	Gemma3-4B	11.8	16.1	8.5	9.0	15.0	17.5	8.8	3.8	9.0	12.2	20.9	28.6
memα	Gemma3-12B	22.1	41.2	10.9	18.3	15.0	32.9	9.5	16.4	8.8	21.8	62.6	81.5
memα	Gemma3-27B	22.7	43.0	13.3	33.0	16.6	40.4	5.2	17.7	7.6	18.8	89.0	**89.9**
BM25	Gemma3-4B	15.2	33.0	14.4	37.9	14.7	26.5	9.6	8.4	5.5	12.7	0.2	1.3
BM25	Gemma3-12B	29.1	51.8	22.7	54.6	18.1	33.3	10.3	14.3	13.4	28.0	8.6	66.3
BM25	Gemma3-27B	29.5	54.2	26.1	**56.0**	17.9	35.4	10.2	14.7	12.8	25.8	33.5	70.9
Embedding	Gemma3-4B	14.1	28.6	12.5	30.6	14.8	24.5	10.1	9.2	5.6	13.1	0.4	1.8
Embedding	Gemma3-12B	25.9	48.0	20.2	46.7	17.3	31.3	10.2	16.8	12.8	**29.3**	8.6	68.9
Embedding	Gemma3-27B	26.6	48.9	22.5	44.5	18.8	34.9	10.5	16.8	12.7	27.1	39.4	72.5

C Generation Prompts and Examples

Admission Conversation and Summary Prompt

```
System:
You are generating a synthetic but clinically grounded conversation for exactly one hospital admission.

Rules:
1. Use only facts supported by the provided admission context.
2. The discharge notes are the backbone of the hospital-course narrative. Use them first.
3. The ordered problem list should shape the main clinical focus.
4. Radiology, procedures, microbiology, and the previous-admission summary are supporting context, not a checklist.
5. Not every packet item needs to appear in the conversation.
6. Speakers must be exactly: Doctor and Patient.
7. Every conversation line must contain:
   - turn number
   - exact timestamp
   - speaker
   - text
8. Timestamps must be monotonically increasing and stay within the admission start and admission end times.
9. Keep the conversation medically plausible and understandable to a patient.
10. Use short natural exchanges, patient questions, clarifications, explanations, reassurance, and follow-up discussion.
11. Do not invent unsupported diagnoses, procedures, test findings, or outcomes.
12. Do not mention the dataset, packet, note types, or that the conversation is synthetic.
13. Generate the conversation first, then write the summary strictly from the generated conversation lines.
14. Every claim in the summary paragraph and every listed problem must be explicitly supported by the conversation lines.
15. Return output that exactly matches the required JSON schema.

User:
Generate one rich doctor-patient conversation for this admission and one admission summary.

Generation guidance:
- Approximate stay length: {stay_days} day(s).
- Recommended turn range: {recommended_turn_range} turns.
- Exact admission window: {admission_start} to {admission_end}.

Return JSON with this shape:
{
  "conversation_lines": [
    {
      "turn_number": 1,
      "time": "YYYY-MM-DD HH:MM:SS",
      "speaker": "Doctor",
      "text": "..."
    }
  ],
  "summary": {
    "admission_start": "YYYY-MM-DD HH:MM:SS",
    "admission_end": "YYYY-MM-DD HH:MM:SS",
    "summary_paragraph": "...",
    "problems": ["...", "..."]
  }
}
```

Single-Admission Answerable QA Prompt

```
System:
You are generating benchmark-quality hard short-answer question-answer pairs from one hospital-admission conversation
between a doctor and a patient.

Rules:
1. Use the provided admission conversation as the evidence source.
2. Generate only answerable questions.
3. Outside medical knowledge is allowed only if it is common, stable, and clinically basic.
4. Use outside knowledge to interpret the context, not to invent unsupported facts.
5. Every question must be hard and should usually require synthesis across multiple turns.
6. Questions must sound natural and be written for benchmark evaluators, not for the patient.
7. Do not use second-person wording like "you" or "your" in the question text.
8. Use third-person phrasing such as "the patient", "the patient's symptoms", or "the doctor" when needed.
9. Every answer must be a short open answer, not yes/no, not multiple choice, and not more than 10 words.
10. Evidence must cite the admission id plus supporting turn_ids.
11. Output valid JSON only.

User:
Generate exactly {question_count} hard answerable short-answer question-answer pairs for this single admission.

Allowed question_type values:
- medical_reasoning
- care_plan_rationale

Return JSON with this exact shape:
{
  "qas": [
    {
      "qa_id": "...",
      "scope": "single_admission",
      "question_type": "medical_reasoning",
      "question": "During the hospitalization from YYYY-MM-DD to YYYY-MM-DD, which condition most likely drove the
antibiotic escalation?",
      "answer": "worsening pneumonia",
      "evidence": {
        "admissions": ["..."],
        "turn_ids": [1, 5, 9]
      }
    }
  ]
}
```