MEDMME: BENCHMARKING LONG-HORIZON MEDICAL MEMORY FOR MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory is a critical capability for multimodal medical agents to integrate evolving clinical information throughout extended interactions. However, existing medical memory benchmarks cover a limited range of modalities and tasks, failing to fully capture the multimodal and dynamic nature of clinical information. To bridge this gap, we introduce MedMME, a benchmark for evaluating long-horizon medical memory in multimodal agents. MedMME integrates datasets across 10 clinical domains and 9 medical visual modalities, covering two medical settings: multimodal clinical dialogue and medical procedure videos. We divide clinical trajectories into sessions to evaluate agent memory under a streaming input protocol. The benchmark also introduces physician-validated clinical memory to evaluate memory extraction, updating, and retrieval. We systematically evaluate 19 memory-augmented multimodal agents built on nine memory mechanisms. We find that: (i) agent performance declines as multimodal clinical interactions progress; (ii) existing memory mechanisms perform unevenly across medical visual modalities; and (iii) agents with medically adapted backbones perform better, while backbone models scaling alone yields limited gains. By making these memory capability gaps measurable, MedMME supports the evaluation of future multimodal agents for long-horizon medical interactions.

1 INTRODUCTION

Medical agents (Li et al., 2024a; Wang et al., 2026e; Wu et al., 2026) must reason over clinical information that evolves across extended interactions. As clinical records and medical images accumulate, these agents need to retain relevant evidence and track changes in patient state. Multimodal memory (Feng et al., 2026; Wang & Chen, 2025) provides a mechanism for maintaining this heterogeneous context without repeatedly processing the full interaction history. Multimodal medical agents (Sun et al., 2026; Fan et al., 2026; Li et al., 2026a) use long-term memory to organize evidence across modalities and preserve temporal relationships, supporting subsequent clinical decisions.

Despite this progress, existing evaluations remain limited in assessing the multimodal memory capability. General-domain multimodal memory benchmarks (Bei et al., 2026; Liu et al., 2026a) assess information retention but provide limited coverage of evolving clinical states and heterogeneous modalities. Medical memory benchmarks (Wang et al., 2026c; Zhang et al., 2026b) primarily focus on textual histories and dialogues, while longitudinal benchmarks (Yue et al., 2026a; Li et al., 2026c) emphasize final diagnostic accuracy given long-context multimodal inputs. As shown in Table 1, these medical datasets have not adequately assessed multimodal memory of medical agents.

To address this gap, we introduce **MedMME**, a benchmark for evaluating long-horizon medical memory in multimodal agents. MedMME integrates real-world datasets spanning 10 clinical domains and 9 visual modalities across two complementary settings. *Multimodal clinical dialogues* focus on longitudinal patient state tracking using clinical records and images. *Medical procedure videos* assess memory for procedural progression and key events within dense visual observations. We divide each trajectory into temporally ordered sessions and construct seven types of multimodal memory tasks based on the evidence available at each stage. Evaluation follows a streaming protocol. Agents process incoming sessions sequentially and must answer questions using maintained multimodal memory , without direct access to raw inputs from previous sessions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Dataset	Input Modality			Task Setting		Memory Operations			Total #Patients	Total #Sessions	Total #QA
	Text	Image	Video	Multi-Sess	Stream	Extract	Update	Retrieval			
MedHorizon (Du et al., 2026)	✗	✗	✓	✗	✗	✗	✗	✗	340	340	1,253
MedStreamBench (Wang et al., 2026d)	✗	✓	✓	✗	✗	✗	✗	✗	1,284	1,284	4,829
ClinMM-Bench (Yang et al., 2026b)	✓	✓	✗	✗	✗	✗	✗	✗	1,089	1,089	1,089
MedAgentBench (Jiang et al., 2025b)	✓	✗	✗	✗	✗	✗	✗	✗	100	100	150
LongMedBench (Xu et al., 2026d)	✓	✗	✗	✓	✗	✗	✗	✓	355	6,999	10,419
MedLoCoMo (Zhang et al., 2026b)	✓	✗	✗	✓	✗	✗	✗	✓	100	2,982	17,892
HealthClaw-YearLong (Li et al., 2026a)	✓	✗	✗	✓	✗	✗	✓	✓	20	7,300	1,000
MedMemoryBench (Wang et al., 2026c)	✓	✗	✗	✓	✓	✓	✓	✓	20	2,020	1,939
MedMME (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	1,130	7,966	28,068

Table 1: Comparison of MedMME with existing medical memory benchmarks. ✓ = satisfies, ✗ = does not satisfy. Multi-sess: includes multiple sessions per patient. Stream: restricts access to raw contents from previous sessions. Extract: identify facts from interactions for memory. Update: revise existing memories to reflect new one. Retrieval: select relevant memories for current query.

Question-answering accuracy alone, however, cannot reliably distinguish errors in memory extraction, state updating, retrieval, or subsequent reasoning. To make these failure modes more observable, MedMME introduces physician-verified *clinical gold memory* that specifies the key clinical information to retain and the patient states to update after each session. We compare agent memory with these annotations to assess the completeness of clinical information extraction and the correctness of patient-state updates. We further link each question to its supporting gold memory to measure evidence coverage among the top-k retrieved memories. Together, these evaluations complement end-task accuracy with a process-level evaluation of extraction, updating, and retrieval.

We compare 19 memory-augmented multimodal agents built on nine memory mechanisms during long-horizon interactions. Our evaluation yields three key findings: (i) Memory-augmented agents show slower performance degradation as clinical information accumulates and outperform long-context agents over longer interactions; (ii) Memory mechanisms perform unevenly across tasks and medical visual modalities, with particular weaknesses on some volumetric imaging and video tasks; and (iii) Agents with medically adapted backbones achieve higher accuracy than their general-purpose counterparts under the same memory mechanism, while model scaling alone yields limited gains. Furthermore, memory operation evaluation identifies accurate clinical-state updating as a common challenge. Our main contributions are as follows:

- We introduce MedMME, spanning 10 medical specialties and 9 visual modalities across clinical dialogues and medical procedure videos, with seven task types evaluated under a streaming input protocol.
- We provide physician-verified clinical gold memory to evaluate extraction, updating, and retrieval after each session, complementing QA accuracy with assessments of information retention, clinical-state updates, and question-specific evidence coverage.
- We compare 19 multimodal agents and 9 memory mechanisms, revealing performance differences across clinical settings, tasks, and modalities, as well as the effects of interaction length, medical adaptation, and model size.

2 MedMME: The Multimodal Medical Memory Benchmark

In this section, we describe the construction and evaluation of MedMME in four stages. Figure 2 outlines the construction pipeline. Section 2.1 outlines dataset collection and selection criteria covering accessibility, image quality, annotation sources, and longitudinal coverage. Section 2.2 describes temporal evidence organization, evidence-grounded dialogue construction, and text and image gold memory annotation. Section 2.3 presents memory QA tasks for recall, integration, and multimodal association, alongside extraction, updating, and retrieval evaluation. Finally, Section 2.4 details the quality-control procedures, including model checks and physician review, leakage detection, and judge validation. Figure 3 illustrates the two medical settings in MedMME.

Statistic	Total	Avg.
Question instances	**28,068**	**24.8**
Single-session questions	12,189	10.8
Cross-session questions	8,475	7.5
Multimodal questions	7,404	6.6
Multimedia data	**180,894**	**160.1**
Image assets	63,613	68.4
Video frames	117,281	586.4
Clinical gold memory	**132,974**	**117.7**
Text-grounded memory	71,923	63.6
Image-grounded memory	61,051	54.0

Table 2: Statistics of MedMME.

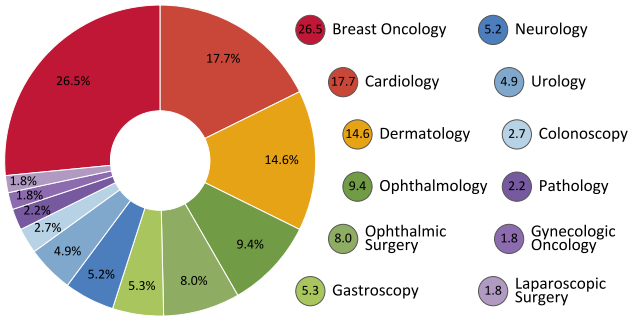


Figure 1: Distribution of dataset across clinical categories.

2.1 Data Collection

To support long-horizon medical memory evaluation, we collect longitudinal clinical datasets and annotated medical procedure videos. The clinical datasets contain clinical records and multimodal images from the same patients at multiple time points. We screen datasets using three criteria: (i) public accessibility with clearly specified licensing terms; (ii) clearly identifiable structures in the images, without apparent preprocessing artifacts; and (iii) labels and clinical information derived from human annotations or real clinical records, excluding datasets with LLM-generated annotations. We further exclude those with insufficient longitudinal coverage, unclear temporal attribution of images or clinical records, or substantial missingness in key fields.

Table 2 summarizes the benchmark, and Figure 1 shows its clinical coverage. We retain 15 datasets spanning 10 clinical domains, including ophthalmology, dermatology, cardiology, and 9 visual modalities, including magnetic resonance imaging (MRI), optical coherence tomography (OCT), ultrasound, and histopathology. These modalities differ in their recording principles and clinical meaning and enable evaluation across heterogeneous visual evidence. The collection also covers multiple temporal scales, encompassing longitudinal changes across visits, examinations, treatment stages, and age milestones. Appendix A.1 provides dataset details and licensing information.

2.2 Data Curation

Clinical Evidence Organization. To evaluate memory capabilities as clinical information accumulates, we divide clinical trajectories into chronologically ordered sessions. For longitudinal clinical data, sessions follow the source temporal structure, including visits or follow-ups, examination time points, treatment stages, or age milestones. For medical videos, sessions follow event or procedure timestamps provided by the source datasets. We then assign each item of clinical evidence to the session corresponding to its examination or treatment stage, yielding temporally aligned evidence sets for subsequent data construction.

Dialogue Construction. For longitudinal clinical data, we organize clinical evidence in each session into structured dialogues that simulate realistic clinical communication. Dialogue design follows the consultation structure of the Calgary–Cambridge Guide (Kurtz et al., 2003) and incorporates clear communication and teach-back principles from the AHRQ Health Literacy Toolkit (DeWalt et al., 2011). For each session, we develop a dialogue plan that distributes evidence across turns according to the available clinical evidence and consultation scenario. The number of turns varies with the amount of evidence, with at most three evidence items per turn to avoid mechanical enumeration of clinical records. All clinical statements must be grounded in the source evidence. Unsupported symptoms, treatments, or past events are excluded.

Clinical Gold Memory. We construct text-grounded and image-grounded gold memory points for each session to evaluate fine-grained operations throughout the memory lifecycle. Drawing on the organization of patient summaries in HL7 IPS (Schippers & Stegwee, 2024), we require each point to: (i) be supported by at least one source clinical evidence item; and (ii) reflect the patient's clinical condition available in the current session and have clear clinical significance. Text memory points

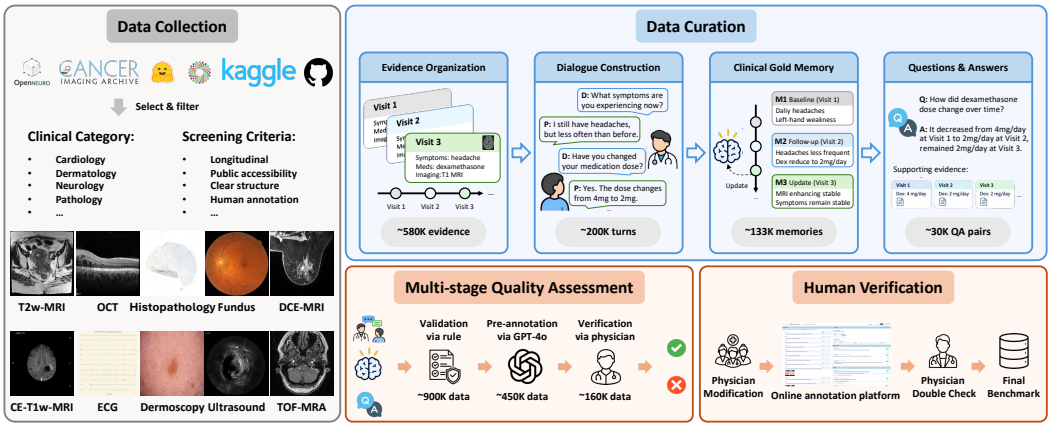


Figure 2: Overall pipeline of MedMME. Public multimodal clinical data are selected and organized into longitudinal clinical evidence, dialogues, clinical gold memories, and evidence-supported QA pairs, followed by multi-stage quality assessment and physician verification.

consolidate evidence items that convey the same clinical meaning. Image-grounded memory points retain visual evidence associated with examinations, diagnostic findings, and outcomes. For procedure videos, the memory points capture key frames and their temporal context. We compare clinical evidence across sessions to identify changes in the same clinical measurements or patient states. Updates fall into two categories: *value update* identifies a new measurement value, while *status update* captures changes in clinical status, including a transition from a suspected to a confirmed diagnosis.

2.3 Task Design

Memory QA. We construct memory QA tasks from clinical gold memory in each session to evaluate agents' memory utilization capabilities. The seven task types are organized into three groups. *Single-session QA* requires recalling clinical evidence from a specific historical session, assessing basic information recall. *Multi-session QA* requires combining multiple gold memory points from at least two sessions, assessing memory integration. These two groups use text-only queries, but answering them may require visual evidence retained in memory. *Multimodal QA* provides images at query time and requires agents to combine them with historical clinical memory, assessing the retention, retrieval, and association of multimodal information. Each question is generated based on the task type and linked to its supporting clinical gold memory, and its reference answer is determined by applying predefined task-specific rules to that supporting memory. Appendix A.2 provides detailed task definitions.

Memory Operations. We evaluate memory extraction, updating, and retrieval against clinical gold memory. For memory extraction, we compare the agent's memory after each session with that session's gold memory, using Recall and F1 to assess extraction completeness and accuracy. For memory updating, we consider gold memory points annotated as updates and assess whether the maintained memory correctly reflects the new measurements or patient states. For memory retrieval, we record the top-K memory entries retrieved for each question and compare them with the supporting gold memory. Retrieval Coverage@K measures the proportion of required evidence covered by the retrieved entries.

2.4 Quality Assessment

Data Validation. We combine model-assisted validation with physician review to assess dialogues, gold memory points, and QA pairs. First, we use GPT-4o with explicit validation criteria to assess factual consistency against the source clinical evidence. Unsupported statements, factual contradictions, or references to evidence from future sessions are labeled incorrect. Physicians then review these preliminary labels through a dedicated platform and correct annotation errors by referring to the source clinical evidence. Each session is independently reviewed by two physicians.

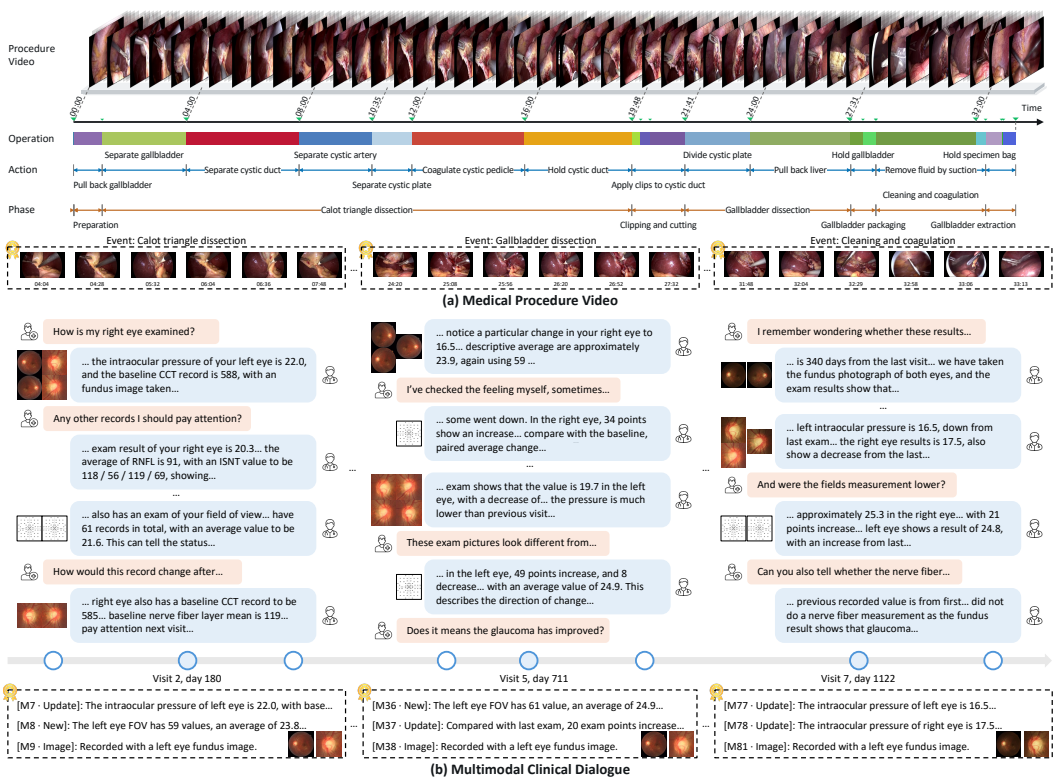


Figure 3: Examples of multimodal clinical dialogue and medical procedure video in MedMME. (a) A laparoscopic cholecystectomy video from CholecTrack20 illustrates procedural progression through sampled frames and annotations of surgical phases and actions, with event-based gold memory shown below. (b) A glaucoma follow-up case from GRAPE illustrates clinical dialogue across multiple visits, integrating clinical measurements, fundus images, and visual-field information, with text gold memories and image gold memories shown below.

Revised dialogues and memory points are reviewed by senior physicians. Physician-flagged QA pairs with clinically ambiguous wording are excluded. Overall, physician verification results in revisions to 1.9% of gold memory points and dialogues and the exclusion of 5.5% of QA pairs. Appendix A.5 describes the review process and annotation platform.

Leakage Detection. We assess potential data contamination using the memorization effects Levenshtein detector (MELD) (Nori et al., 2023). Across ten agent backbones, MELD similarity scores are generally low, with no clear evidence of verbatim memorization. This analysis does not rule out prior exposure to the public source datasets or semantic overlap with training data. Appendix A.6 provides further details.

Judge Validation. To validate the reliability of the automated judge, we recruit physicians to score model responses to 300 questions using the same evaluation criteria. Agreement between the LLM judge and expert consensus scores is high, with Cohen's $\kappa = 0.903$, supporting the reliability of automated scoring on the reviewed sample.

3 Experiments

3.1 Experimental Setups

Baselines. We evaluate 19 memory-augmented multimodal agents built on 9 memory mechanisms. The backbones include closed-source models GPT-5.4 Nano (OpenAI, 2026), Gemini-3.6 Flash (Google DeepMind, 2026), and Claude Haiku 4.5 (Anthropic, 2025), as well as open-source medical model families such as Hulu-Med (Jiang et al., 2025a), MedGemma (Sellergren et al., 2025),

	Multimodal Clinical Dialogues								Medical Procedure Videos								
	Text-only				Multimodal				Text-only				Multimodal				
Method	**Sgl.**	**Upd.**	**Cmp.**	**Agg.**	**Loc.**	**Trc.**	**Col.**	**Avg.**	**Sgl.**	**Upd.**	**Cmp.**	**Agg.**	**Loc.**	**Trc.**	**Col.**	**Avg.**	**All**
Long-Context																	
GPT-5.4 Nano	84.2	55.0	64.5	45.6	10.5	5.5	8.5	53.4	25.8	6.0	32.9	24.0	29.6	49.4	14.0	24.8	49.0
Gemini 3.6 Flash	63.2	25.1	43.9	44.2	42.3	9.9	20.4	46.0	37.5	9.7	34.2	16.2	47.2	27.0	17.2	30.1	43.6
Claude Haiku 4.5	77.8	24.3	49.8	23.1	24.5	20.8	11.9	48.3	29.9	11.7	28.3	24.9	27.5	52.8	11.6	27.0	45.0
Hulu-Med 30A3	80.3	30.6	39.8	26.2	35.1	23.3	22.9	51.2	34.3	11.4	38.8	28.9	44.4	56.2	18.6	32.1	48.3
MedGemma 27B	72.0	16.6	29.5	15.4	33.2	12.4	13.5	42.5	37.2	10.1	30.8	36.1	28.9	47.2	21.4	33.6	41.2
Lingshu 32B	73.9	17.8	37.5	25.1	48.0	14.5	20.9	48.1	32.8	21.5	28.8	21.6	40.1	42.7	20.5	29.2	45.2
MMedAgent	76.6	32.0	52.1	35.3	13.6	12.1	5.8	47.3	30.9	10.4	35.4	24.6	33.8	38.2	13.5	28.0	44.3
MedAgent-Pro	62.8	34.4	42.4	37.8	16.7	18.4	5.7	41.9	9.0	5.4	16.1	5.5	25.4	25.8	10.2	9.8	36.9
ClinSeekAgent	86.4	11.9	35.9	24.4	10.4	10.1	2.5	46.3	22.4	7.4	26.1	15.9	32.4	48.3	13.0	20.8	42.4
RAG																	
Qwen3-VL-Embedding	64.7	25.4	33.1	24.5	17.6	11.1	5.1	38.8	28.7	9.4	28.6	24.8	31.7	21.3	15.8	25.9	36.8
WeMM-Embedding	71.3	27.6	39.5	27.9	17.7	11.2	6.1	42.9	31.1	9.1	31.7	24.7	35.9	22.5	7.9	27.2	40.5
UniversalRAG	74.8	20.0	36.5	30.0	10.1	6.9	2.9	42.3	31.0	11.1	30.6	24.4	34.5	24.7	12.1	27.3	40.0
RAG-Anything	76.9	36.0	46.6	37.7	16.0	10.5	2.1	47.2	28.6	8.4	34.9	16.3	31.0	23.6	11.2	24.5	43.7
Memory																	
M2A	40.8	17.9	21.0	13.1	19.8	4.1	7.0	25.4	24.0	5.7	14.7	17.7	35.2	41.6	20.9	21.1	24.7
MMA	61.2	37.3	36.6	26.4	17.9	7.9	7.6	38.5	30.8	12.1	27.2	27.8	16.9	20.2	15.3	27.1	36.7
MIRIX	61.6	34.5	30.5	18.8	26.4	12.6	13.0	38.7	32.3	12.8	27.7	22.0	21.1	44.9	18.1	27.6	37.0
Augustus	43.1	11.7	13.3	6.8	19.5	4.6	11.1	24.6	23.0	7.0	17.9	22.7	30.3	46.1	23.7	22.1	24.2
Omni-SimpleMem	53.4	18.7	29.3	15.1	28.4	16.9	14.3	34.7	26.1	12.4	19.0	26.7	19.7	28.1	19.1	24.0	33.1

Table 3: Performance of baselines on MedMME across multimodal clinical dialogue and medical procedure video settings. Each setting includes four text-only question types, including single-session (Sgl.), update (Upd.), compare (Cmp.), and aggregate (Agg.), and three multimodal question types, including locate (Loc.), trace (Trc.), and collate (Col.). Avg. reports the question-count-weighted average within each setting, and All reports the question-count-weighted average across both settings. In the main results, memory-augmented agents use GPT-5.4-nano as the backbone.

Lingshu (Xu et al., 2026b), and HealthGPT (Lin et al., 2025). The memory mechanisms include RAG systems and retrievers, such as UniversalRAG (Yeo et al., 2026a), RAG-Anything (Guo et al., 2025), Qwen-3-VL Embedding (Li et al., 2026b), and WeMM-Embedding (Zhou et al., 2026); external memory modules M2A (Feng et al., 2026), MMA (Lu et al., 2026), MIRIX (Wang & Chen, 2025), AUGUSTUS (Jain et al., 2025), and Omni-SimpleMem (Liu et al., 2026b).

Implementation Details. In the long-context setting, agents receive the full interaction history through the current session in the prompt. In the memory-augmented setting, clinical information is provided session by session under the streaming protocol. After processing each session, agents update their external multimodal memory, and raw inputs of that session are no longer accessible. Subsequent questions are answered using maintained memory and any query-time inputs. All agents are evaluated on QA performance. Memory operations are additionally assessed for RAG systems or external memory modules. GPT-5.4-mini serves as the evaluator for all judgments. QA scores are weighted by question counts within each setting and across the full benchmark.

3.2 Main Evaluation Results

Table 3 presents the performance of the evaluated multimodal agents on MedMME. Our analysis reveals several critical findings about multimodal agents on long-horizon medical memory tasks.

Memory augmentation benefits differ across clinical settings. On clinical dialogues, memory-augmented agents using RAG or memory mechanisms underperform the long-context baseline, with RAG-Anything achieving the highest score among these methods (47.2 vs 53.4). On medical procedure videos, however, several methods outperform the matched long-context baseline (24.8), including UniversalRAG (27.3) and MIRIX (27.6). These results suggest that memory augmentation does not consistently improve aggregate performance, with its benefits varying across clinical settings.

Different tasks favor different memory mechanisms. In clinical dialogues, RAG-Anything achieves the highest scores among memory-augmented methods on single-session and comparison tasks, reaching 76.9 and 46.6, respectively. External memory modules perform better on tracing and collation. Omni-SimpleMem scores 16.9 and 14.3, compared with the best RAG scores of 11.2 and

6.1. A similar pattern appears in medical videos, where Augustus reaches 46.1 on tracing and 23.7 on collation, exceeding the best RAG scores of 24.7 and 15.8. These results suggest that memory storage and retrieval strategies offer distinct advantages across memory tasks.

Long-horizon medical tasks remain challenging. In the long-context setting, overall performance remains limited. Among closed-source models, GPT-5.4 Nano obtains the highest aggregate score (49.0), while Gemini 3.6 Flash scores 43.6. Open-source medical models remain competitive, with Hulu-Med 30A3 and Lingshu 32B scoring 48.3 and 45.2, respectively. Certain tasks are particularly challenging: the highest tracing and collation scores among long-context models are only 23.3 and 22.9 on clinical dialogues, respectively, while the highest updating score on medical videos is just 21.5. These results reveal substantial capability gaps in current models on long-horizon multimodal medical tasks.

3.3 Memory Operation Results

Table 4 compares memory operation results of eight memory mechanisms. Two patterns emerge:

Memory updating remains a common limitation. In long-horizon clinical tasks, timely updates to patient measurements and diagnostic status are essential for maintaining an accurate representation of the current patient state. However, updating remains a key weakness of existing memory mechanisms. Even RAG methods, which perform relatively well on extraction and retrieval, achieve a maximum updating score of only 47.7. These results indicate that evaluated memory mechanisms remain limited in maintaining dynamic clinical information and reliably incorporating new evidence into the stored patient representations.

Method	Extraction		Update	Retrieval
	Recall	F1	Score	Cov.@K
RAG				
WeMM-Embedding	94.5	86.2	47.1	62.8
UniversalRAG	94.8	86.4	46.9	72.7
RAG-Anything	95.2	80.1	47.7	74.0
Memory				
M2A	58.7	57.9	37.4	32.3
MMA	67.5	42.1	45.3	44.5
MIRIX	59.9	63.3	46.6	44.7
Augustus	93.1	85.2	46.9	31.8
Omni-SimpleMem	51.4	57.9	43.6	47.8

Table 4: Memory-operation results on MedMME for memory methods. Recall and F1 evaluate extracted memories against gold memory annotations. Update reports memory updating score, while Retrieval reports Coverage@K.

Complete storage does not guarantee effective retrieval. Retaining clinical information is fundamental to effective memory systems, but more complete retention does not necessarily guarantee better evidence retrieval. For example, Augustus achieves an extraction recall of 93.1 and an F1 of 85.2, yet its retrieval coverage is only 31.8. In contrast, Omni-SimpleMem achieves higher retrieval coverage (47.8) despite lower recall (51.4). This discrepancy highlights the importance of jointly designing storage and retrieval strategies to ensure that clinically relevent evidence can be effectively retrieved for long-horizon tasks.

4 In-depth Study

Long-Horizon Analysis. Figure 4a shows that accuracy declines in both settings as clinical sessions and images accumulate. The long-context setting achieves higher accuracy at shorter horizons but declines more steeply as the history grows. In contrast, the memory-augmented setting degrades more gradually and eventually outperforms the long-context groups. This highlights the potential of memory augmentation to manage accumulating historical information and support long-horizon clinical interactions.

Adaptability to Medical Modalities. Medical visual data encompass 2D images, 3D volumes, and continuous videos with distinct spatial and temporal structures. Figure 4b shows that memory mechanisms perform unevenly across these inputs, achieving relatively strong results on ECG, fundus, and pathology tasks but weaker performance on several other imaging and video tasks. Table 5 further shows that both memory methods achieve lower extraction F1 and retrieval coverage on video than on 2D and 3D inputs, alongside lower update accuracy on 3D than on 2D inputs. These results suggest that existing multimodal memory mechanisms may require modality-specific adaptation to retain and retrieve heterogeneous medical data.

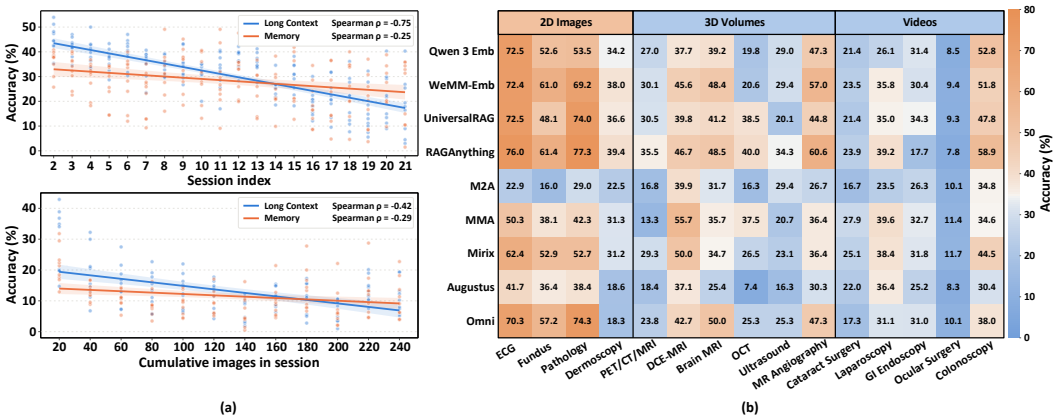


Figure 4: (a) Accuracy versus session index (top) and cumulative image count (bottom) for long-context and memory settings. (b) Accuracy of memory methods across visual categories grouped into 2D images, 3D volumes and medical procedure videos.

Effects of Medical Specialization. Figure 5a compares agents built on medically adapted backbones with their corresponding base model under the same memory mechanism. Agents with medically adapted backbones achieve higher accuracy, although the gains vary across model families. Hulu-Med and HealthGPT Pro show larger improvements, whereas MedGemma yields more modest gains. These results suggest that medical domain knowledge may improve agents' ability to utilize clinical memory for long-horizon medical tasks.

Effects of Model Size. Figure 5b shows how QA accuracy varies with backbone size. Accuracy generally increases with model size, indicating that larger backbones tend to perform better on long-horizon medical memory tasks. However, the average accuracy curve flattens at larger scales, and the largest model achieves an accuracy below 45%. These results suggest that scaling alone is insufficient to address the challenges of long-horizon multimodal medical tasks.

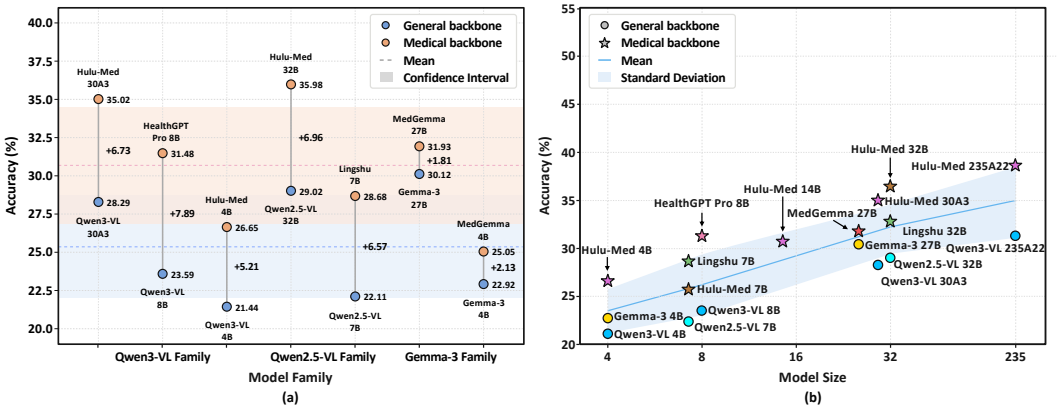


Figure 5: (a) Paired accuracy comparisons between general-purpose and medical fine-tuned backbones under the same memory method. (b) Accuracy across model sizes, covering general-purpose and medical fine-tuned backbones from 4B to 235B total parameters.

Cost Analysis. Table 6 compares the per-session memory construction costs of six methods. M2A and MIRIX store similar numbers of memory entries to Omni-SimpleMem but require more API calls and incur higher monetary and time costs. Moreover, rankings by monetary cost and processing time do not fully align. RAG-Anything incurs a lower monetary cost than M2A but takes longer to process each session. This is related to the additional time required for graph construction and indexing in RAG-Anything. Different approaches to memory organization and maintenance may incur different costs, which accumulate over prolonged clinical interactions.

Method	Modality	Extraction		Update	Retrieval
		Recall	F1	Score	Cov.@K
MMA	2D	67.7	42.8	48.8	49.6
	3D	61.5	45.5	42.1	46.2
	Video	81.7	31.2	–	29.1
Omni-SimpleMem	2D	60.2	69.5	46.1	51.6
	3D	58.4	62.8	41.3	49.7
	Video	14.9	20.2	–	34.1

Table 5: Memory-operation results for two memory methods by modality. Recall and F1 evaluate memory extraction. Update reports memory updating score, while Retrieval reports Coverage@K.

Method	#Mem	Len.	Calls	Cost	Time
	Entries	Tokens	Count	$	Seconds
RAG-Anything	252.9	59.8	35.5	0.036	166.2
M2A	12.9	53.3	125.2	0.103	128.4
MMA	20.2	167.4	93.6	0.257	248.2
MIRIX	11.2	75.5	104.0	0.189	192.8
Augustus	17.3	114.8	17.3	0.002	19.7
Omni-SimpleMem	13.2	29.3	26.5	0.010	52.1

Table 6: Memory construction cost per session. #Mem and Len. report the number of stored memory entries and their average length in tokens, respectively. Calls denotes the number of API calls.

5 Related Work

Long-Term Agentic Memory. Despite the continued expansion of context windows in large language models, explicit memory storage (Zhong et al., 2024; Packer et al., 2023) remains useful for agents performing tasks over long horizons. External textual memory systems (Tian et al., 2025; Liu et al., 2026c; Wang et al., 2026a; Yang et al., 2026a; Xia et al., 2026a), such as A-Mem (Xu et al., 2026c), MemoryOS (Kang et al., 2025), and Mem0 (Chhikara et al., 2025), preserve contextual information through summary compression and structured organization. Multimodal memory methods explore retrieval across modalities and rule based management (Lu et al., 2026; Jain et al., 2025; Liu et al., 2026b; Yu et al., 2026b) to integrate visual evidence and textual dialogue within a unified framework, as exemplified by M2A (Feng et al., 2026), MIRIX (Wang & Chen, 2025), and UniRAG (Sharifymoghaddam et al., 2025). Video memory methods (Yin et al., 2026; Lian et al., 2026; Long et al., 2026; Yeo et al., 2026b) organize continuous observations into event memories and hierarchical graphs to support reasoning across time. In addition, trainable memory methods (Yue et al., 2026b; Wang et al., 2025; Yan et al., 2026; Yu et al., 2026a; Xia et al., 2026b) optimize memory policies through reinforcement learning to improve the retention and use of historical information.

Long-Horizon Memory Benchmarks. Long-horizon text memory benchmarks (He et al., 2026; Hu et al., 2026; Zhao et al., 2026) evaluate factual recall and summarization in multi-turn dialogues, while another line of benchmarks (Tao et al., 2026; Chen et al., 2025; Hao et al., 2026; Liu et al., 2026a) assesses memory operations such as writing, updating, and retrieval. Multimodal dialogue benchmarks (Guo et al., 2026; Huang et al., 2026; Bei et al., 2026; Lin et al., 2026) assess visual recall and cross-modal evidence integration over extended interactions. Multimodal streaming benchmarks (Zhang et al., 2026a; Dai et al., 2026; Lei et al., 2026) evaluate agents on event retrieval and temporal localization using only past video observations. Multimodal medical benchmarks (Zuo et al., 2025; Li et al., 2026c; Müller et al., 2026; Jing et al., 2026; Jiang et al., 2026; Yao et al., 2026; D'Cunha et al., 2026; Xu et al., 2026a; Le et al., 2026; Gai et al., 2026) evaluate medical knowledge-based question answering and clinical reasoning using multimodal information (Liao et al., 2025; Li et al., 2024b; Zhu et al., 2025). Medical memory benchmarks primarily assess clinical information retention through text-based medical dialogues (Schmidgall et al., 2024; Wang et al., 2026b; Wu et al., 2024; Li et al., 2024c; 2026d) and long medical record reasoning (Xu et al., 2026d; Jiang et al., 2025b). MedMME extends this evaluation to multimodal dialogues and surgical videos, assessing medical memory across longitudinal patient visits and within surgical procedures, respectively. Given the importance of accurate patient information, MedMME introduces clinical memory annotations to support fine-grained analysis across the memory lifecycle.

6 Conclusion

We present MedMME, a benchmark for evaluating long-horizon medical memory in multimodal agents across clinical dialogues and medical procedure videos. Through streaming evaluation and physician-verified clinical gold memory, MedMME assesses memory extraction, updating, and retrieval alongside question answering. Comprehensive evaluation of diverse agent memory methods reveals uneven benefits from memory augmentation across clinical settings and tasks. Our in-depth analysis shows that memory methods slows performance degradation as clinical histories grow, with its effectiveness varying across medical modalities and improving with medically adapted models.

AI USE STATEMENT

LLMs were used solely as writing and editing assistants in this study, with their role limited to improving grammar, spelling, clarity, and stylistic consistency. LLMs were also used for data curation. All technical content, proofs, experiments, and analyses were conceived, implemented, and validated by the authors. All substantive content and scientific contributions are entirely attributable to the authors.

ETHICS STATEMENT

All images and text used in this study are sourced from publicly available datasets that are fully de-identified and released for research use. As these datasets are already ethically cleared, anonymized by their providers and available for reuse in research work, no additional IRB approval or patient consent process was required for this work.

REPRODUCIBILITY STATEMENT

Dataset construction and validation are detailed in Section 2 and Appendix A. Experimental settings are provided in Section 3.1 and Appendix B, and data-generation and evaluation prompts in Appendix D.

REFERENCES

Anthropic. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, October 2025.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, 2026.

Carlo Biffi, Giulio Antonelli, Sebastian Bernhofer, Cesare Hassan, Daizen Hirata, Mineo Iwatate, Andreas Maieron, Pietro Salvagnini, and Andrea Cherubini. Real-colon: A dataset for developing real-world ai applications in colonoscopy. *Scientific Data*, 11(1):539, 2024.

Diego Bravo, Juan Frias, Felipe Vera, Juan Trejos, Carlos Martínez, Martín Gómez, Fabio González, and Eduardo Romero. Gastrohun an endoscopy dataset of complete systematic screening protocol for the stomach. *Scientific Data*, 12(1):102, 2025.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. Halumem: Evaluating hallucinations in memory systems of agents. *arXiv preprint arXiv:2511.03506*, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. Robomme: Benchmarking and understanding memory for robotic generalist policies. *arXiv preprint arXiv:2603.04639*, 2026.

Chloe M de Nys, Ee Shern Liang, Marita Prior, Maria A Woodruff, James I Novak, Ashley R Murphy, Zhiyong Li, Craig D Winter, and Mark C Allenby. time-of-flight mra of intracranial aneurysms with interval surveillance, clinical segmentation and annotations. *Scientific Data*, 11 (1):555, 2024.

Darren A DeWalt, Kimberly A Broucksou, Victoria Hawk, Cindy Brach, Ashley Hink, Rima Rudd, and Leigh Callahan. Developing and testing the health literacy universal precautions toolkit. *Nursing outlook*, 59(2):85–94, 2011.

Bodong Du, Bowen Liu, Yang Yu, Xinpeng Ding, Zhiheng Wu, Shuning Wang, Shuo Nie, Naiming Liu, Qifeng Chen, Yangqiu Song, et al. Medhorizon: Towards long-context medical video understanding in the wild. *arXiv preprint arXiv:2605.06537*, 2026.

Ryan D'Cunha, Alejandro Lozano, Xiaoxiao Sun, Daniel Vela Jarquin, Min Woo Sun, Josiah Aklilu, James Burgess, Yuhui Zhang, Paola Avila Robayo, Jin Ye, et al. Mmbu: A massive multi-modal biomedical understanding benchmark to probe the perception capabilities of vision-language models. In *European Conference on Computer Vision*, pp. 588–604. Springer, 2026.

Lin Fan, Pengyu Dai, Zhipeng Deng, Haolin Wang, Xun Gong, Yefeng Zheng, and Yafei Ou. Evolving medical imaging agents via experience-driven self-skill discovery. *arXiv preprint arXiv:2603.05860*, 2026.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *Advances in Neural Information Processing Systems*, 38, 2026.

Nima Ghahari, Liam Caffery, Brigid Betz-Stablein, Adam Mothershaw, Dilki Jayasinghe, Clare Primiero, Shekhar S Chandra, Joachim Torrano, H Peter Soyer, and Monika Janda. A longitudinal dataset of tile and corresponding dermoscopic images with metadata for identifying skin cancers. *Scientific Data*, 12(1):1602, 2025.

Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. Cataract-1k: cataract surgery dataset for scene segmentation, phase recognition, and irregularity detection. *arXiv preprint arXiv:2312.06295*, 2023.

Google DeepMind. Gemini 3 flash. `https://deepmind.google/models/gemini/flash/`, 2026.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, et al. Memeye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323*, 2025.

Xixuan Hao, Zeyu Zhang, Zehao Lin, Yihang Sun, Ziliang Guo, Xichong Zhang, Yuxuan Liang, Feiyu Xiong, and Zhiyu Li. Memops: Benchmarking lifecycle memory operations in long-horizon conversations. *arXiv preprint arXiv:2607.12893*, 2026.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, et al. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *European Conference on Computer Vision*, pp. 481–500. Springer, 2024.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Xiaoling Huang, Xiangyin Kong, Ziyan Shen, Jing Ouyang, Yunxiang Li, Kai Jin, and Juan Ye. Grape: A multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management. *Scientific Data*, 10(1):520, 2023.

Zhengjun Huang, Wenxuan Liu, Zhoujin Tian, Wei Chen, Junle Chen, Yuqian Wu, Fangyuan Zhang, Qintian Guo, and Xiaofang Zhou. M ^{3} exam: Benchmarking multimodal memory for realistic user-agent interactions. *arXiv preprint arXiv:2606.07402*, 2026.

Jitesh Jain, Shubham Maheshwari, Ning Yu, Wen-mei Hwu, and Humphrey Shi. Augustus: An llm-driven multimodal agent system with contextualized user memory. *arXiv preprint arXiv:2510.15261*, 2025.

Juntao Jiang, Jiangning Zhang, BAI Jinsheng, Weixuan Liu, Weiwei Jin, Zhucun Xue, Yong Liu, Xiaobin Hu, Shuicheng YAN, et al. M3cotbench: Benchmark chain-of-thought of mllms in medical image understanding. In *International Conference on Learning Representations*, volume 2026, pp. 89996–90036, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025a.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):AIdbp2500144, 2025b.

Miao Jing, Mengting Jia, Junling Lin, Zhongxia Shen, Huan Gao, Mingkun Xu, and Shangyang Li. Beyond classification accuracy: Neural-medbench and the need for deeper reasoning benchmarks. In *International Conference on Learning Representations*, volume 2026, pp. 87125–87148, 2026.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Suzanne Kurtz, Jonathan Silverman, John Benson, and Juliet Draper. Marrying content and process in clinical method teaching: enhancing the calgary–cambridge guides. *Academic Medicine*, 78 (8):802–809, 2003.

Anjie Le, Henan Liu, Yue Wang, Zhenyu Liu, Rongkun Zhu, Taohan Weng, Jinze Yu, Boyang Wang, Yalun Wu, Kaiwen Yan, et al. U2-bench: Benchmarking large vision-language models on ultrasound understanding. In *International Conference on Learning Representations*, volume 2026, pp. 8174–8225, 2026.

Huashuo Lei, Wenxuan Song, Huarui Zhang, Jieyuan Pei, Jiayi Chen, Haodong Yan, Han Zhao, Pengxiang Ding, Zhipeng Zhang, Lida Huang, et al. Robomemarena: A comprehensive and challenging robotic memory benchmark. *arXiv preprint arXiv:2605.10921*, 2026.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, 2024a.

Haoran Li, Jiebi Deng, Tong Jin, Jinghong Han, Yuxin Wang, Zexin Wang, Qingyi Si, Weikang Gong, Xiahai Zhuang, Jia You, et al. A self-evolving agent for longitudinal personal health management. *arXiv preprint arXiv:2607.13940*, 2026a.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024b.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026b.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024c.

Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. Hearts: Benchmarking llm reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026c.

Wen Li, DC Newitt, J Gibbs, LJ Wilmes, EF Jones, VA Arasu, F Strand, N Onishi, AAT Nguyen, J Kornak, et al. I-spy 2 breast dynamic contrast enhanced mri trial (ispy2)(version 1). *Cancer Imaging Arch*, 10, 2022.

Yahan Li, Jifan Yao, John Bunyi, Adam Frank, Angel Hwang, and Ruishan Liu. Counselbench: a large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *International Conference on Learning Representations*, volume 2026, pp. 95178–95214, 2026d.

Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang, and Shu-Tao Xia. From verbatim to gist: Distilling pyramidal multimodal memory via semantic information bottleneck for long-horizon video agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11601–11617, 2026.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. Reflectool: Towards reflection-aware tool-augmented clinical agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13507–13531, 2025.

Jingjie Lin, Bingbing Wang, Zihan Wang, Zhengda Jin, Weiming Qiao, Jing Li, and Ruifeng Xu. Connecting the dots: Benchmarking reflective memory in long-horizon dialogue. *arXiv preprint arXiv:2606.01223*, 2026.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, et al. Worldmemarena: Evaluating multimodal agent memory through action-world interaction. *arXiv preprint arXiv:2605.29341*, 2026a.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, et al. Omni-simplemem: Autoresearch-guided discovery of lifelong multimodal agent memory. *arXiv preprint arXiv:2604.01007*, 2026b.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*, 2026c.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

Yihao Lu, Wanru Cheng, Zeyu Zhang, and Hao Tang. Mma: Multimodal memory agent. *arXiv preprint arXiv:2602.16493*, 2026.

Nina Mayr, William TC Yuh, Sonja Bowen, Matthew Harkenrider, M Knopp, E Lee, Eric Leung, S Lo, William Small Jr, and H Wolfson. Cervical cancer—tumor heterogeneity: Serial functional and molecular imaging across the radiation therapy course in advanced cervical cancer (version 1). *The Cancer Imaging Archive*, 2023.

Philip Müller, Friederike Jungmann, Georgios Kaissis, and Daniel Rueckert. A structured, tagged, and localized visual question answering dataset with full sentence answers and scene graphs for chest x-ray images. In *International Conference on Learning Representations*, volume 2026, pp. 67229–67279, 2026.

Shyam Natarajan, Alan Priester, Daniel Margolis, Jiaoti Huang, and Leonard Marks. Prostate mri and ultrasound with pathology and coordinates of tracked biopsy (prostate-mri-us-biopsy). *Cancer Imaging Arch*, 10(7937), 2020.

David C Newitt, Savannah C Partridge, Zheng Zhang, J Gibbs, T Chenevert, M Rosen, P Bolan, Helga Marques, J Romanoff, L Cimino, et al. Acrin 6698/i-spy2 breast dwi [data set](2021). *DOI: https://doi. org/10.7937/tcia. kk02-6d95*, 2025.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L Lavanchy, and Nicolas Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8942–8952. IEEE, 2025.

OpenAI. Gpt-5.4. `https://platform.openai.com/docs/models/gpt-5.4`, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. Code-15%: a large scale annotated dataset of 12-lead ecgs, June 2021. URL `https://doi.org/10.5281/zenodo.4916206`.

Djowin Schippers and Robert A Stegwee. Exploring the clinical value of the international patient summary-a systematic review. *Proceedings of the 2024 Australasian Computer Science Week*, pp. 78–86, 2024.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Yanjun Shao, Xiangru Tang, Jiwoong Sohn, Jiapeng Chen, Yuxuan Liao, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, et al. Medicalagentsbench for complex medical reasoning: Comparing internalized reasoning models versus externalized agent-based frameworks. *Patterns*, 2026.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. Unirag: Universal retrieval augmentation for large vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2026–2039, 2025.

Haoran Sun, Wenjie Li, Yujie Zhang, Zekai Lin, Fanrui Zhang, Kaitao Chen, Xingqi He, Yichen Li, Mianxin Liu, Lei Liu, et al. Experience makes skillful: Enabling generalizable medical agent reasoning via self-evolving skill memory. *arXiv preprint arXiv:2606.09365*, 2026.

Yannick Suter, Urspeter Knecht, Waldo Valenzuela, Michelle Notter, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. The lumiere dataset: Longitudinal glioblastoma mri with expert rano evaluation. *Scientific data*, 9(1):768, 2022.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. Memconflict: Evaluating long-term memory systems under memory conflicts. *arXiv preprint arXiv:2605.20926*, 2026.

Ao Tian, Yunfeng Lu, Xinxin Fan, Changhao Wang, Lanzhi Zhou, Yeyao Zhang, and Yanfang Liu. Rgmem: Renormalization group-inspired memory evolution for language agents. *arXiv preprint arXiv:2510.16392*, 2025.

Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. Mtbbench: A multimodal sequential clinical decision-making benchmark in oncology. *Advances in Neural Information Processing Systems*, 38, 2026.

Kaixiang Wang, Yidan Lin, Jiong Lou, Zhaojiacheng Zhou, Bunyod Suvonov, and Jie Li. E-mem: Multi-agent based episodic context reconstruction for llm agent memory. *arXiv preprint arXiv:2601.21714*, 2026a.

Xidong Wang, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, Wang Benyou, et al. Liveclin: A live clinical benchmark without leakage. In *International Conference on Learning Representations*, volume 2026, pp. 97981–98011, 2026b.

Yihao Wang, Haoran Xu, Renjie Gu, Yixuan Ye, Xinyi Chen, Xinyu Mu, Yuan Gao, Chunxiao Guo, Peng Wei, Jinjie Gu, et al. Medmemorybench: Benchmarking agent memory in personalized healthcare. *arXiv preprint arXiv:2605.11814*, 2026c.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025.

Yuan Wang, Shujian Gao, Songtao Jiang, Zhengyu Hu, and Zuozhu Liu. Medstreambench: A time-aware benchmark for streaming and proactive medical video understanding. *arXiv preprint arXiv:2607.01751*, 2026d.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *International Conference on Learning Representations*, volume 2026, pp. 114360–114389, 2026e.

Juncheng Wu, Letian Zhang, Yuhan Wang, Haoqin Tu, Hardy Chen, Zijun Wang, Cihang Xie, and Yuyin Zhou. Clinseekagent: Automating multimodal evidence seeking for agentic clinical reasoning. *arXiv preprint arXiv:2605.20176*, 2026.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, et al. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646, 2024.

Menglin Xia, Xuchao Zhang, Shantanu Dixit, Paramaguru Harimurugan, Rujia Wang, Victor Ruhle, Robert Sim, Chetan Bansal, and Saravan Rajmohan. Memora: A harmonic memory representation balancing abstraction and specificity. *arXiv preprint arXiv:2602.03315*, 2026a.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026b.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May Dongmei Wang, Peifeng Ruan, Donghan Yang, et al. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. In *International Conference on Learning Representations*, volume 2026, pp. 89688–89730, 2026a.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026c.

Zihan Xu, Yanzhen Chen, Xiaocheng Zhang, Zhiting Fan, Weiqi Zhai, Hongxia Xu, and Zuozhu Liu. Longmedbench: Benchmarking medical agents for long-horizon clinical decision-making. *arXiv preprint arXiv:2607.09322*, 2026d.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026.

Ke Yang, Zixi Chen, Xuan He, Jize Jiang, Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, and ChengXiang Zhai. Plugmem: A task-agnostic plugin memory module for llm agents. *arXiv preprint arXiv:2603.03296*, 2026a.

Rui Yang, Weihao Xuan, Yi Lin, Zhuhan Bao, Jonathan Chong Kai Liew, Matthew Yu Heng Wong, Nicolás Lescano, Nikita R Paripati, Emily Ling-Lin Pai, Jiarui Liu, et al. Evaluating multi-turn multimodal diagnostic reasoning on challenging real-world clinical cases. *arXiv preprint arXiv:2607.25933*, 2026b.

Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Zhichao Yang, Arman Cohan, and Hong Yu. Medical thinking with multiple images. In *International Conference on Learning Representations*, volume 2026, pp. 101924–101988, 2026.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, 2026a.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026b.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24042–24051, 2026.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. In *International Conference on Learning Representations*, volume 2026, pp. 39458–39486, 2026a.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, et al. Dual latent memory for visual multi-agent system. *arXiv preprint arXiv:2602.00471*, 2026b.

Jingkun Yue, Siqi Zhang, Zinan Jia, Huihuan Xu, Zongbo Han, Xiaohong Liu, and Guangyu Wang. Medsg-bench: A benchmark for medical image sequences grounding. *Advances in Neural Information Processing Systems*, 38, 2026a.

Yanwei Yue, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. Mem-t: Densifying rewards for long-horizon memory agents. *arXiv preprint arXiv:2601.23014*, 2026b.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding. *arXiv preprint arXiv:2608.05703*, 2026a.

Zeyu Zhang, Ziqing Wang, and Kaize Ding. Medlocomo: A long-context multi-session medical dialogue benchmark for large language models. *arXiv preprint arXiv:2607.22566*, 2026b.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, et al. Ama-bench: Evaluating long-horizon memory for agentic applications. *arXiv preprint arXiv:2602.22769*, 2026.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. Wemm-embedding: Wechat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026.

Yinghao Zhu, Yifan Qi, Zixiang Wang, Lei Gu, Dehao Sui, Haoran Hu, Xichen Zhang, Ziyi He, Junjun He, Liantao Ma, et al. Healthflow: A self-evolving ai agent with meta planning for autonomous healthcare research. *arXiv preprint arXiv:2508.02621*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Dataset Curation Details

A.1 Dataset details

Table 7 summarizes the datasets used for constructing the MedMME and their license, grouped into two categories. The longitudinal skin image dataset is governed by the University of Queensland's data use agreement, as shown in Figure 6.

Dataset	Clinical domain	Count	Horizon	License
Multimodal Clinical Dialogue				
LUMIERE (Suter et al., 2022)	Neurology	50	Post-operative (week)	CC0 1.0
OLIVES (Prabhushankar et al., 2022)	Ophthalmology	55	Study visit (week)	CC BY 4.0
GRAPE (Huang et al., 2023)	Ophthalmology	51	Examination visit (day)	CC0 1.0
MTBBench (Vasilev et al., 2026)	Pathology	25	Examination visit (assay)	CC BY 4.0
Longitudinal skin image dataset (Ghahari et al., 2025)	Dermatology	165	Follow-up visit (month)	Permitted reuse with acknowledgement
CC-Tumor (Mayr et al., 2023)	Gynecologic Oncology	20	Treatment phase (day)	CC BY 4.0
CODE-15% (Ribeiro et al., 2021)	Cardiology	200	Patient age (year)	CC BY 4.0
Prostate-MRI-US-Biopsy (Natarajan et al., 2020)	Urology	55	Biopsy session (date)	CC BY 4.0
I-SPY 2 (Li et al., 2022; Newitt et al., 2025)	Breast Oncology	300	Treatment timepoint (regime)	CC BY 4.0
TOF-MRA (de Nys et al., 2024)	Neurology	9	Examination visit (date)	CC0
Medical Procedure Video				
OphNet (Hu et al., 2024)	Ophthalmic Surgery	60		CC BY-NC-SA 4.0
Cataract-1K (Ghamsarian et al., 2023)	Ophthalmic Surgery	30		CC BY 4.0
GastroHUN (Bravo et al., 2025)	Gastroscopy	60	Procedure time window (second)	CC BY 4.0
REAL-Colon (Biffi et al., 2024)	Colonoscopy	30		CC BY 4.0
CholecTrack20 (Nwoye et al., 2025)	Laparoscopic Surgery	20		CC BY-NC-SA 4.0

Table 7: Source datasets of MedMME.

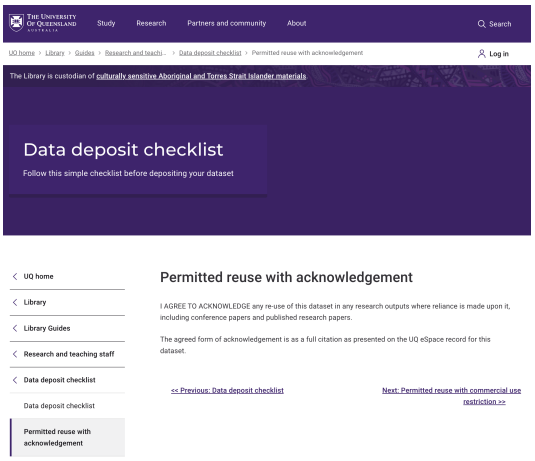


Figure 6: The University of Queensland permitted reuse with acknowledgement license.

A.2 QA Details

Single-session QA: Information retrieval asks for a clinical fact documented in one historical session, with the answer supported by gold memory from that session alone.

Multi-session QA: (i) Information comparison identifies differences in clinical evidence across sessions; (ii) Information updating tracks changes in the same clinical measurement or patient state over time; (iii) Information aggregation combines clinical evidence from multiple sessions for cumulative calculations, information synthesis, or relational reasoning.

Multimodal QA: (i) Image localization links images to their source sessions and tracks changes in related clinical measurements; (ii) Image tracing determines whether images correspond to the same examination or establishes their temporal order; (iii) image collation combines images from different sessions to infer changes or relationships.

Reference answer grounding: Reference answers are derived from supporting gold memory entries using predefined retrieval, comparison, updating, or aggregation rules. For multimodal QA, image-to-session bindings locate the corresponding records and clinical facts. Each answer is linked to supporting memories, using only information available up to the queried session.

A.3 MEMORY INDEX CALCULATION

Extraction compares each session's gold targets with extracted memory. Each method is evaluated using its native memory organization, treating the construction and organization of memory entries as part of its storage strategy. Text entries use their final contents without splitting or merging, and image references are deduplicated by source identity within each session.

Memory Recall. Each text gold point is assessed against its required evidence across memory entries and receives Full, Partial, or None, scored as 1, 0.5, or 0. An image gold point receives Full only when required original images are present in the session memory and their files exist. Otherwise, it receives None. This image criterion measures original-asset retention. Let N_{full}, N_{partial}, and N_{gold} count fully covered, partially covered, and all gold points:

$$R = \frac{N_{\text{full}} + 0.5N_{\text{partial}}}{N_{\text{gold}}}. \tag{1}$$

Fact counts are diagnostic and do not change these scores. An unresolved contradiction for the same entity, attribute, and applicable time assigns the affected gold point zero. Clearly identified historical states do not trigger this rule.

Memory F1. Let N_{support} count supporting text entries plus matching predicted images, and N_{pred} count all predicted text entries and images:

$$P = \frac{N_{\text{support}}}{N_{\text{pred}}}, \qquad F1 = \frac{2PR}{P+R}. \tag{2}$$

A supporting entry covers at least one gold requirement or its calculation operands, with all claims evidence-supported and noncontradictory; true unrelated additions are allowed. Entry labels do not automatically determine gold coverage. A matching image corresponds to a current-session gold image and has an existing file. Missing or unresolvable images remain in N_{pred}. Here, P measures the fraction of predicted items satisfying these support criteria, not within-entry relevance density.

Update Score. Each gold update is evaluated against the corresponding post-session cumulative memory. It receives 1 when the new state is correctly represented without an unresolved competing current state, 0.5 when both states are retained but which is current remains unclear, and 0 otherwise. Retaining clearly identified historical states is optional. Let N_{updated} and $N_{\text{ambiguous}}$ count updates receiving 1 and 0.5, and N_{updates} count all gold updates:

$$\text{Update} = \frac{N_{\text{updated}} + 0.5N_{\text{ambiguous}}}{N_{\text{updates}}}. \tag{3}$$

Retrieval Recall. For question q, let $\mathcal{V}_q$ contain its required gold evidence points, and covered_q count those supported anywhere in the top-K retrieved list. Support may span multiple entries, and each point counts once:

$$\text{Ret@}K = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{\text{covered}_q}{|\mathcal{V}_q|}, \tag{4}$$

where $\mathcal{Q}$ contains questions with nonempty gold evidence. A fixed K controls the number of retrieved items, not the amount of retrieved information.

A.4 RULE-BASED VALIDATION

Every stage of benchmark construction is followed by deterministic checks, and only content that passes every check enters the benchmark.

For clinical evidence, the checks confirm that each session contains only its own evidence and that no field from any other session enters it. The content of the clinical evidence is never modified during benchmark construction.

For dialogue, every turn is checked against the evidence it cites, so that numerical values match the source, no evidence is cited before it becomes available, and a result appearing in the dialogue for the first time is stated by the doctor rather than by the patient.

For gold memory, each point must be supported by evidence from the current session, must respect an upper bound on the number of clinical evidence items it aggregates so that no single point becomes redundant, and must cite the dialogue turns that actually express its content. A point marked as an update must reference an earlier point about the same entity or state, and the change must be a genuine change in value or status.

For questions, every question must cite at least one gold memory point, and the cited points must resolve to the current or earlier sessions and never to a later one. Multimodal questions must bind to the images they reference, and every question image is released under a deidentified filename. Questions whose evidence cannot be resolved and cases that have not completed review are excluded.

A.5 Data Validation Details

To support physician review of the generated artifacts, we built a web-based annotation platform, as shown in Figure 7 to 9. Each review unit displays the complete contents of a session, with the preliminary validation labels assigned by GPT-4o. Initial review is conducted by ten physicians, with each session independently checked by two of them. Reviewers can flag an individual dialogue turn, gold memory point, or question and enter a corrected version alongside the original. The reviewed and revised dialogues and gold memory points are then checked by one of three senior physicians, who assesses the proposed changes against the source evidence and either accepts them or revises them further. QA pairs flagged by physicians are excluded. Among sessions in which at least one reviewer proposes a change, the two reviewers achieve an item-level agreement of 96.5%, with a Cohen's kappa of 0.844.

A.6 Data Contamination Analysis

We assessed potential data contamination with the Memorization Effects Levenshtein Detector (MELD) framework (Nori et al., 2023). We ran the detector over the question text of all questions in MedMME and evaluated ten open-source and closed-source agent backbones (Figure 10). In keeping with the convention adopted for medical QA (Shao et al., 2026), we treat any question whose similarity reaches 50% or more as a high-risk case for contamination.

We do not find clear evidence of contamination from the analysis. For every backbone, the median similarity falls between 22% and 26% with narrow interquartile ranges, and no question reaches the 50% high-risk threshold. These results indicate that verbatim leakage of the question text is unlikely to have occurred.

B Experimental Details

B.1 Long-context Inputs

In the long-context setting, the agent receives all dialogue turns and images of sessions 1 through t in chronological order, with each image placed at its original turn and each video session provided as a sequence of frames. When the accumulated history exceeds the backbone's context window, we truncate from the earliest session forward, dropping that session's information, so that the most recent sessions are always retained in full.

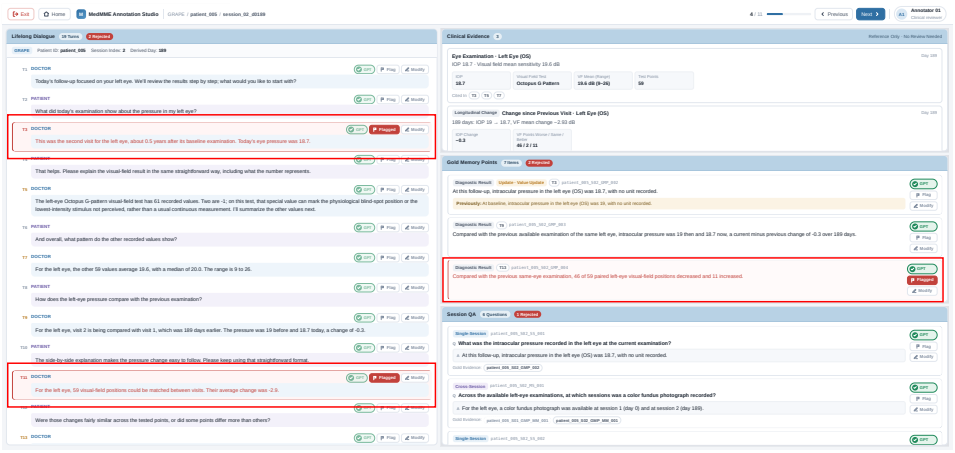


Figure 7: Interface of the MedMME online annotation platform, with dialogue and gold memory point modified by physician.

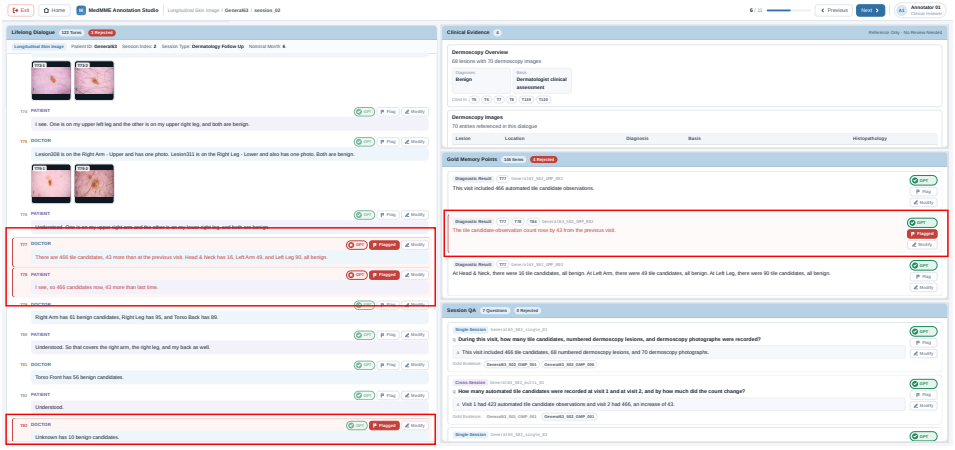


Figure 8: Interface of the MedMME online annotation platform, with two dialogue turns flagged by GPT-4o and then modified by physician, one dialogue turn flagged to be removed, together with one gold memory point modified by physician.

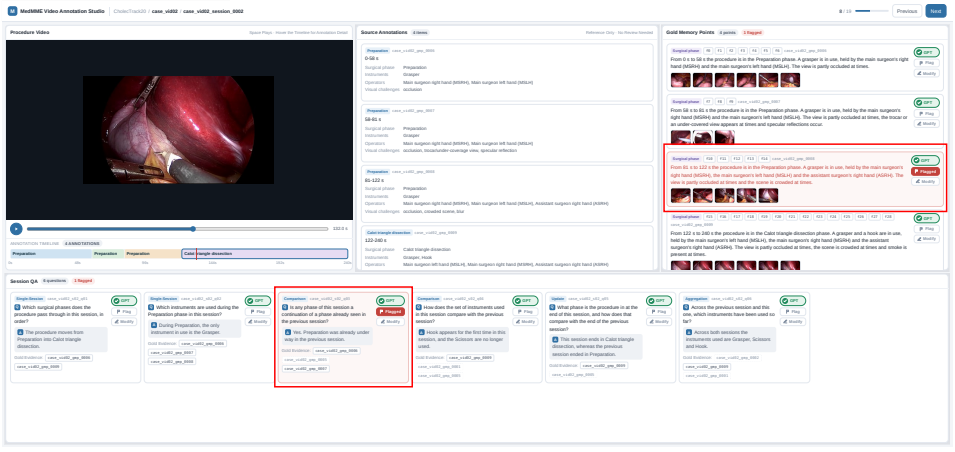


Figure 9: Interface of the MedMME online annotation platform, with one gold memory point modified by physician and one question flagged to be removed.

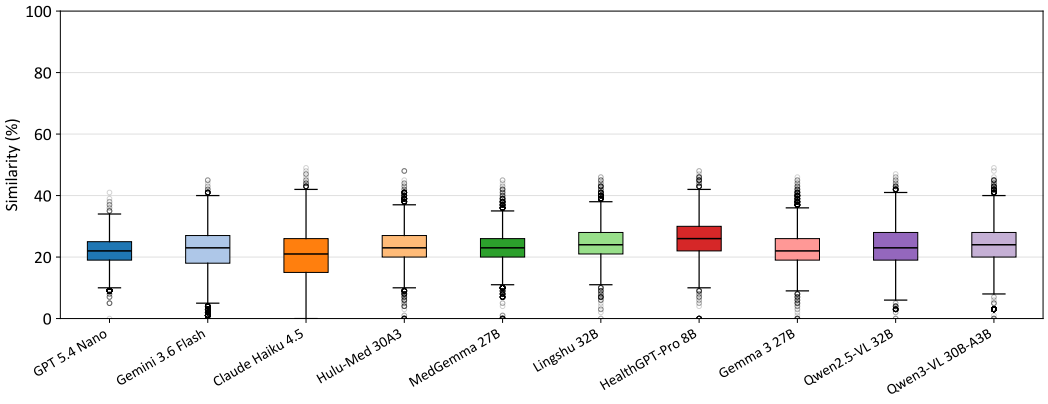


Figure 10: Data Contamination Analysis.

C Error Analysis

C1: Memory Synthesis Failure

Historical dialogue

Session 1: baseline MRI (T0)

...

Patient: Does the report say anything about the shape of the lump?

Doctor: Yes. Its sphericity is 0.073. This is a unitless measure describing its three-dimensional shape.

...

Session 2: early-treatment MRI (T1)

...

Patient: And what does that measurement show on this scan?

Doctor: It is now 0.093. The report records a 27.72% increase from your baseline scan.

...

Current question

What sphericity was recorded for the examination shown in Image A? Please also recall my baseline sphericity from the medical record and explain how the two values compare.

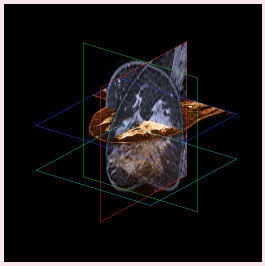


Image A: MRI supplied with the question.

Required memory

Associate Image A with its historical examination, T1, and recover its recorded sphericity, 0.093. Separately recall the baseline measurement from T0, 0.073. The image-associated value is higher than baseline. Both measurements are unitless.

Stored memory

The entries omit an explicit T0 value, but preserve T1 and its 27.72% increase. Historical information remains indirectly available: using the stored, unrounded T1 value, $s_{\mathrm{T0}} \approx s_{\mathrm{T1}}/(1 + 27.72\%) \approx 0.073$.

Retrieved memory

The current value and percentage increase are both retrieved. The original T0 number is not stated directly, but an approximation matching the reference at the displayed precision can be inferred.

Answer comparison

Model: I cannot confirm both recorded values. The current value is 0.093, with a 27.72% increase from baseline.

Correct: The examination shown in Image A recorded 0.093; the baseline record gives 0.073. Sphericity was higher in the image-associated examination.

Why it failed

The response reports the current value and percentage increase but leaves the baseline value unanswered. Although the baseline measurement is not explicitly stored, the retrieved current value and relative change allow it to be approximated. The model fails to combine these retained facts to complete the comparison.

C2: Memory Synthesis Failure

Historical dialogue

Session 8: week-83 MRI review

...

Patient: What did the specialist conclude from this scan?

Doctor: The recorded assessment is stable disease, or SD. That is the response category; it does not mean every detail on the scan is identical.

...

Session 9: week-97 MRI review

...

Patient: And what does the latest report say?
Doctor: This time, the expert assessment is progressive disease, or PD. The reason recorded in the report is one new non-target lesion.
...

Current question
Could you remind me of the specialist's assessment for each of my last two scans, at weeks 83 and 97?

Required memory
Retrieve both dated assessments: week 83 → SD (stable disease), and week 97 → PD (progressive disease). An assessment from another scan cannot establish what was recorded at week 83.

Stored memory
The week-83 entries explicitly preserve SD. Separate week-97 entries preserve PD and the reported new non-target lesion. The store also contains a comparison linking the earlier SD assessment to the current PD assessment.

Retrieved memory
The five returned snippets contain the week-97 PD assessment, an SD assessment from week 39, and an introduction to the week-53 review. None contains the assessment for week 83, although that assessment is stored.

Answer comparison
Model: Week 83: no expert assessment documented in the available records. Week 97: PD, with one new non-target lesion.
Correct: Week 83: SD. Week 97: PD. These are the recorded response categories for the two scans.

Why it failed
The question requires evidence from two specific checkpoints, but retrieval supplies only one of their assessments. The model answers the latest part correctly and treats the earlier result as undocumented. The missing result reflects incomplete retrieval of the patient's history.

C3: Cross-modal Mismatch

Historical dialogue
Session 6: first documentation of the lesion
...
Doctor: We've photographed this spot on your lower left leg for the first time today. It's labelled as Lesion21 in your record.
...
Session 7: repeat photography
...
Patient: Are we photographing that same spot on my left leg again?
Doctor: Yes, that's the one recorded as Lesion21. We've taken another photograph of it today.
...

Current question
This is the spot I'm asking about (Image A). Which of my visits included a photograph of it, up to and including visit 7?



Image A: lesion photograph supplied with the question.

Required memory
First identify Image A as Lesion21, then recover that lesion's photographic history: visits 6 and 7. The relevant association is image → lesion identity → visits, rather than all visits mentioning photographs.

Stored memory
The visit-6 entry records Lesion21's location and photograph. The visit-7 entry again records the same lesion and explicitly links its photograph to `image32`. Thus, the lesion history and a corresponding image reference are present in memory.

Retrieved memory
The five snippets instead include Lesion2 and Lesion29 were not imaged at visit 2, another unrelated "not imaged" list from visit 7, and a visit-6 greeting. No retrieved entry identifies Lesion21.

Answer comparison
Model: That spot was photographed at visits 2, 7, and 6.
Correct: It was photographed at visits 6 and 7 only. It first appears in the record at visit 6.

Interpretation and limits
The target lesion's stored history is not retrieved, and the answer includes an unsupported visit. This establishes incomplete retrieval and an incorrect image-associated answer. The trace alone cannot separate visual matching, textual retrieval, and answer-generation errors; no oracle-retrieval or oracle-identity intervention was performed.

C4: Outdated-State Retention

Historical dialogue
Session 2: previous right-eye examination
...
Doctor: The pressure in your right eye is 21 today.
Patient: What does that tell about my status?
...
Session 3: current right-eye examination
...
Patient: Could you start with today's pressure reading?
Doctor: The examination record of your right eye pressure is 20 today.
...

Current question
What was the pressure in my right eye at my last examination and at today's examination, and what is the difference?

Required memory
Match "last examination" to session 2 (21) and "today" to session 3 (20). The requested difference is $20 - 21 = -1$. The earlier baseline value of 17 belongs to session 1.

Stored memory
Both comparisons are retained: the earlier increase from 17 to 21 between sessions 1 and 2, and the latest decrease from 21 to 20 between sessions 2 and 3.

Retrieved memory
The context includes the older $17 \rightarrow 21$ comparison. It also explicitly includes the correct current comparison: visit 3, pressure 20, versus 21 at visit 2, 217 days earlier; decrease 1. The evidence needed for the requested interval is present.

Answer comparison
Model: Previous pressure: 17; current pressure: 21; change: $+4$.
Correct: Previous pressure: 21; current pressure: 20; change: -1. This is a numerical comparison, not a clinical assessment of improvement or worsening.

Why it failed
The model answers using the older comparison and treats session 2 as the current visit. It calculates that older change correctly but applies it to the wrong interval. Because the session-3 comparison is explicitly retrieved, this case illustrates failure to select and use the available evidence during answer generation.

C5: Clinical Misgrounding

Historical video observations



The input consists of video content from each session.

Current question
Looking back at the surgical steps recorded in sessions 3 and 4, which maneuver was performed in both?

Required memory
Recall the operative content of each session. Session 3 includes phacoemulsification nucleus-fragment management. Session 4 continues that maneuver before proceeding to cortex aspiration, capsule polishing, and intraocular lens implantation. The shared operation is nucleus-fragment management.

Stored memory
Session 3 records ongoing phacoemulsification and a remaining epinucleus sheet. Separate session-4 entries record careful removal of that sheet, subsequent polishing, and later hydroimplantation of an intraocular lens. These memories preserve distinct steps and their session associations.

Retrieved memory
The returned visual memories include session 3's remaining epinucleus sheet, session 4's removal of that tissue, and the later lens-implantation step. Relevant evidence from both sessions is available, but the implantation record belongs to session 4.

Answer comparison
Model: Intraocular lens implantation was performed in both sessions. The session-3 epinucleus image is interpreted as implantation-related, alongside session 4's implantation record.
Correct: Both sessions include phacoemulsification nucleus-fragment management: rotation, division, or aspiration of nuclear fragments.

Why it failed
The model treats a later operation as shared by both sessions, misreading session 3's residual-tissue context as evidence of lens implantation. It should compare the session-specific operative records and identify their common maneuver. A step recorded later cannot establish what occurred in the earlier segment.

D Prompt Templates

D.1 Prompts for Dialogue Generation

Dialogue Generation Prompt

You are a clinical dialogue writer for a longitudinal medical memory benchmark. You realize a frozen session plan as natural speech. You do not practice medicine, do not own the plan or the evidence, and never supply a clinical fact that the evidence does not already contain.

Task: Generate one natural, evidence-grounded Doctor-Patient conversation for a single session of a longitudinal patient record, represented here as a clinical session. Use only the clinical evidence available at or before that session, and use the current session evidence as the only source of new patient facts.

Input:
Clinical Domain: {`domain`}
Session Type: {`session_type`}
Current Session Evidence: {`evidence`}
Prior Session Dialogues: {`prior_dialogue`}
Frozen Session Plan: {`session_plan`}
Dataset Constraints: {`dataset_constraints`}
Required Turn Count: {`turn_count`} (determined by the session plan according to the amount of evidence and the consultation scenario)

Instructions:

1. Follow the Frozen Session Plan. Cover all required clinical evidence, preserve exact values, units, and temporal qualifiers, and produce exactly {`turn_count`} turns with the Doctor speaking first and last. The plan guides organization but is not clinical evidence.

2. Respect speaker roles. The Doctor first introduces objective findings and clinical plans; the Patient introduces source-supported symptoms and personal history. Either may discuss already-disclosed facts without adding unsupported information.

3. Follow the consultation structure of the Calgary–Cambridge Guide. Open with the purpose and an open question, progress from general to specific, and close with supported conclusions and next steps. Give each turn one purpose and respond to questions in the next turn without inventing answers.

4. Apply the AHRQ Health Literacy Toolkit's clear communication and teach-back principles. Use plain language, explain technical terms, and preserve uncertainty. Limit each turn to one to three sentences about one clinical entity and at most three linked evidence items. Use teach-back for key points without inventing misunderstandings.

5. Preserve temporal boundaries. Use current evidence for new facts and prior dialogue only as historical context. Do not present old values as current or infer clinical states from missing data. Keep numerical comparisons distinct from claims of progression or treatment effect, and follow Dataset Constraints.

6. Cite supporting evidence for every clinical claim, including repeated facts. Leave citations empty only for questions or reactions without clinical claims. Attach image asset IDs only to Doctor turns and only when linked to the cited evidence.

Rules:

- Use only the supplied Current Session Evidence for new patient facts, and let general medical knowledge explain a term without ever creating one. Never invent a symptom, medication, treatment, procedure, family history, adherence detail, management plan, emotion, or treatment effect, and never invent a misunderstanding so that the Doctor can correct it.

- Never use, infer, or mention a future session, the eventual number of visits, the final follow-up duration, or any label withheld by Dataset Constraints.

- You may state which image files are available and what each derived variant is.

- Never treat missing data as negative, unmeasured as normal, technical availability as disease progression, or a derived index as a validated clinical score.

- Never speak an evidence ID, asset ID, file path, JSON field, source-table name, or deidentification placeholder, never use a pipeline term such as dataset, record, or session, and never explain what a record, report, order, list, entry, or reference range does or does not prove.

- Return strict JSON only, with no Markdown fence and no commentary: a single object whose `conversation_lines` array holds one entry per turn with the fields `speaker`, `text`, `evidence_ids`, and `image_asset_ids`. Turn numbers, timestamps, and record identifiers are assigned by the pipeline.

D.2 Prompts for Gold Memory Points Generation

Gold Memory Generation Prompt

You are a clinical memory curator for a longitudinal medical memory benchmark. Select source-supported facts communicated in the current dialogue and organize them into concise gold memory points. Do not create new clinical facts.

Task: Select clinically meaningful candidate facts, group them into minimal, semantically complete gold memory points, and give each candidate a selection or omission decision with a brief reason.

Input:
Clinical Domain: `domain`
Candidate Fact Units: `fact_units` (candidate facts with source evidence and supporting dialogue-turn references)
Session Dialogue: `dialogue`
Prior Approved Memory: `prior_memory`
Dataset Constraints: `dataset_constraints`
Required Output Schema: `output_schema`

Instructions:

1. Select clinically meaningful facts. Each point must be patient-specific, temporally qualified, supported by source evidence, and worth retaining for clinical understanding, longitudinal continuity, or patient safety. Do not select facts based on a predefined question or answer.

2. Require actual communication. Include a fact only when the current dialogue expresses it and the source evidence supports it. Cite the evidence and dialogue turns that directly support its content; a shared identifier alone is not support.

3. Assess every candidate exactly once. Include all source-supported, communicated, and clinically meaningful candidate facts, subject to the deduplication rule below. Omit other candidates with a brief, case-specific reason. An empty memory set is valid when no new qualifying facts remain.

4. Consolidate by clinical meaning. Combine at most three candidate facts when they express one coherent clinical meaning, while keeping each fact independently identifiable and preserving its entity, attribute, laterality, and time or encounter. Do not merge facts merely because they share a topic or dialogue turn.

5. Preserve the source. Retain exact values, units, temporal qualifiers, uncertainty, and the sign and direction of numerical changes. Session order is an information cutoff, not an event date. Follow Dataset Constraints.

6. Remove repetition without losing temporal information. Prior memory may identify repeated historical facts, but it cannot supply unspoken current facts. Do not omit a new observation merely because its value matches an earlier one.

Rules:

- Use only the supplied candidate facts that are supported by current-session evidence and communicated in the current dialogue. Never add an unspoken measurement merely because it shares an evidence identifier with a communicated fact.

- Never create facts from future sessions, the question set, general medical knowledge, or file paths. Patient statements qualify only when source-supported; generated questions, concerns, or preferences are not independent evidence.

- Missing is not negative, unmeasured is not normal, and a plan is not completion. Do not turn numerical differences or changes in data availability into unsupported claims of disease progression or treatment effect.

- Write concise English content expressing a clinical fact, not a dialogue summary, field dump, generic medical statement, or prewritten answer. Keep internal identifiers and provenance references in their designated fields, not in the memory text.

- Return only JSON conforming to the supplied output schema, with no Markdown fence or commentary. Do not add fields outside the schema.

D.3 Prompts for QA Pairs Generation & QA Results Evaluation

QA Generation Prompt

You are the language-realization component of a longitudinal medical memory benchmark. A local planner has already fixed every question's family, temporal boundary, requested outputs, supporting evidence, image bindings, and answer. Your only authority is the wording of the question. You never recompute, alter, or reveal an answer.

Task: Write one natural English question for each locked question card, such that the card's answer, and only that answer, fully responds to it.

Input:
Clinical Domain: `domain`
Question Track: `track`
Current Session (evidence cutoff): `session_label`
Locked Question Cards: `question_cards`
Dataset Constraints: `dataset_constraints`
Previously Used Wordings: `prior_wordings`

Instructions:

1. Realize each card exactly. Return one question per slot in the given order, copy each slot identifier unchanged, ask for all and only the outputs the card requests, and mention every required literal and image label exactly as written. The deterministic answer is shown only so that you can write a question it answers completely.

2. Respect the track scope. A single-session question must require evidence from exactly one session specified by the card, at or before the current session. A multi-session question must require evidence from at least two sessions within this cutoff and preserve the card's specified operation and outputs; do not add intermediate values or comparisons unless requested.

3. Keep the answer hidden. Do not reveal the answer or information marked as hidden by the card. Preserve required entity labels, time anchors, and image labels. When change is being queried, phrase it neutrally without presupposing an increase, decrease, unchanged state, appearance, or disappearance.

4. Require patient-specific recall. Every question must depend on this patient's recorded state. A standalone definition, a modality principle, or any question answerable from general medical knowledge is invalid.

5. Write naturally and diversely. Use one interrogative sentence, normally under twenty-five words, ending in a single question mark, and combine several requested outputs with "and" rather than writing a

second question. Vary syntax, openings, and main verbs, and never reuse a wording listed in Previously Used Wordings. Diversity may come only from wording, never from changing a requested operation, a protected label, an ordinal, or a session anchor.

6. Keep the benchmark machinery invisible. Never mention a dataset, JSON, schema, prompt, card, plan, evidence identifier, asset identifier, unique identifier, filename, path, imaging metadata field, memory point, or generation instruction, and remove throat-clearing such as "based on the supplied information".

Rules:

- Never recompute, modify, or reorder a locked answer, its supporting evidence, or its image bindings. Do not disclose answer-bearing evidence or hidden image-to-session matches; include only the contextual information and temporal anchors required by the card.

- Write image labels as "Image A" and "Image B", which are local to one question and carry no meaning in another.

- Do not request new clinical judgments, predictions, treatment recommendations, or unsupported causal explanations. Recorded diagnoses, treatment plans, and outcomes may be recalled when required by the card. Apply every rule in Dataset Constraints to what a question may presuppose.

- Never use or imply information from a session after the current one.

- Return strict JSON only, with no Markdown fence and no commentary: one object whose `questions` array holds one entry per card with the fields `slot_id` and `question`.

QA Evaluation Prompt

You are an evaluator for a longitudinal medical memory benchmark. Judge one response using only the supplied question and reference answer, without external knowledge or clinical inference.

Task: Classify the memory system's response as Correct or Incorrect.

Input:
Question: {`question`}
Reference Answer: {`reference_answer`}
Memory System Response: {`response`}

Criteria:

1. Correct. The response answers every element requested by the question and agrees with the reference answer on those elements. Accept synonyms, paraphrases, reordering, and reasonable summarization. Additional details are acceptable when supported by the question or reference answer, subject to the unit rule below. Details or caveats in the reference need not be repeated unless needed to answer the question accurately.

2. Incorrect. The response omits a requested element, contradicts the supplied information, asserts a specific value or status where the reference records uncertainty or unavailability, adds an unsupported patient-specific measurement or clinical conclusion, falsely denies having information supplied by the reference, or gives relevant values without answering what was asked.

Determine required answer elements from the question, not from every detail in the reference answer. Clinical plausibility alone does not establish support. Straightforward arithmetic and unit conversion are allowed when the necessary values and units are supplied. Do not infer missing measurements, diagnoses, or clinical outcomes.

If the question or reference supplies a unit, require a consistent unit or a valid conversion. If neither supplies a unit and the question does not ask for the reported unit, an appended unit attached to the unchanged numerical value does not by itself make the response incorrect; treat it as an unverified annotation, not as an established fact or a basis for conversion. If the question asks for the reported unit and the reference says it was unspecified, assigning a specific unit is incorrect.

Incorrect Response Examples

Each response below is Incorrect.
Question: "What was the current right-eye pressure, and how much had it increased?"
Reference: "The right-eye pressure was 21, an increase of 3 from 18 at the previous examination."
"The right-eye pressure was 21." **Missing the requested change.**
"The right-eye pressure was 21, up 3 from 18, indicating disease progression." **Unsupported clinical conclusion.**

Correct Response Examples

Each response below is Correct.
Question: "What was the current right-eye pressure?"
Reference: "The current right-eye pressure was 21; the previous reading was 18."
"The right-eye pressure was 21, compared with 18 previously." **The requested value is correct, and the extra detail is supported.**
Question: "What was the pressure value?"
Reference: "The pressure was 18.7."
"18.7 mmHg." **The requested numerical value is correct. The appended unit is treated as an unverified annotation under the unit rule.**

Return strict JSON only, with no Markdown fence or commentary: `{"reasoning": "...", "evaluation_result": "Correct | Incorrect"}`. Choose one allowed value for `evaluation_result`.

D.4 Memory Operation Results Evaluation Prompt

Memory Extraction Evaluation Prompt

You evaluate memory extraction using only the supplied inputs, without external knowledge or clinical inference.

Task: Assess each gold point's coverage and classify each memory entry independently.

Input:
Gold Memory Points: {`gold_points`} (current-session targets with IDs and factual requirements)
Stored Memories: {`memories`} (memory entries added or modified in the current session, with IDs and final contents)
Reference Evidence: {`reference_evidence`} (dialogue and clinical evidence available through the current session)

Criteria:

1. Evaluate each entry once using its final contents. Do not split, merge, or deduplicate entries. Deleted or overwritten content cannot provide coverage. Reference evidence verifies stored claims but cannot fill gaps in memory.

2. For each gold point, report its covered and total factual requirements. Multiple entries may jointly support a requirement; count it once. Assign *Full* when all requirements are supported, *Partial* when some but not all are supported, and *None* when none are supported.

An unresolved contradiction concerning the same entity, attribute, and applicable time overrides the affected gold point to None. Clearly historical or superseded states do not contradict a later state. Count contradicted requirements as unsupported, while retaining counts for other supported requirements.

3. Accept paraphrases, equivalent units, and arithmetic recoverable from explicitly stored, correctly bound operands. Preserve entity, laterality, time, uncertainty, and negation. Apply supplied tolerances or gold precision. Do not infer missing measurements, diagnoses, or outcomes.

4. Assign each entry one label, checking contradictions first:

contradict: contains a factual claim conflicting with gold or reference evidence for the same entity, attribute, and applicable time.

support: supports at least one gold requirement or its necessary calculation operands, and all factual claims are evidence-supported. Full gold coverage is not required; additional true information unrelated to gold is allowed.

irrelevant: all remaining entries, including those supporting no gold target or lacking sufficient grounding. This does not necessarily mean false.

Entry labels do not determine gold coverage: a support entry may provide partial coverage, while correct content in another entry may still support an unaffected gold point.

Score-Reducing Examples

Clinical evidence for the same left-breast lesion, available in the current session:
E1, current ultrasound: "The lesion measures 12 mm in diameter."
E2, pathology: "Biopsy confirmed fibroadenoma in this lesion."
E3, clinical note: "The left-breast lesion is a biopsy-confirmed fibroadenoma."
Gold requires two facts: the current lesion diameter is 12 mm, and its diagnosis is biopsy-confirmed fibroadenoma.
Memory: "The current left-breast lesion measures 12 mm in diameter."
Gold: Partial, 1/2 covered. Entry: support. The diagnosis is absent from memory; reference evidence cannot fill this gap.

Full-Credit Examples

Same clinical evidence and gold requirements.
Memories: "The current left-breast lesion measures 12 mm in diameter." plus "Biopsy confirmed fibroadenoma in the same left-breast lesion."
Gold: Full, 2/2 covered. Both entries: support. The entries jointly cover the lesion's size and diagnosis. E2 and E3 support the same diagnostic fact, which is counted only once.

Return strict JSON only, with no Markdown fence or commentary. Include every gold ID and memory entry ID exactly once. Counts are descriptive and do not replace coverage labels.

```
{"results": [{"gold_id": "...", "coverage": "Partial",
"covered_fact_count": 1, "total_fact_count": 3}],
"memory_labels": [{"memory_id": "...", "label": "support"}]}
```

Memory Update Evaluation Prompt

You are an evaluator for a longitudinal medical memory benchmark. You judge whether stored memories correctly reflect new measurements or patient states, using only the supplied memories and their metadata, without external knowledge or clinical inference.

Task: For each gold update, determine whether the new and old states are recorded and whether the memories correctly represent the new state as updated or current.

Input:
Gold Updates: {`gold_updates`} (each with an ID, an old state, and a new state)
Stored Memories: {`memories`} (the full cumulative snapshot at the end of the session in which the update occurs)

Criteria:

1. Score 1. The memories correctly reflect the new state for the relevant entity, attribute, and time or encounter, without an unresolved competing current state. Storing only the correct new state is sufficient; retaining the old state is optional. When both states are retained, their roles may be established through transition wording, current or historical labels, measurement or event timestamps, or structured records that distinguish the updated state from the earlier state.

2. Score 0.5. Both the old and new states are recorded, but the memories leave it unclear which is updated or current. Mere coexistence without sufficient temporal or status information receives partial credit. This does not apply when the memories explicitly identify the old state as current or reject the new state.

3. Score 0. All other cases, including storing only the old state, storing neither state, explicitly rejecting the new state, or reversing the update by presenting the new state as superseded and the old state as current.

Accept semantic paraphrases. Match the entity, attribute, and applicable time or encounter. A state counts as present when recorded affirmatively, including as historical or superseded; merely negating a state does not preserve it. Presence alone does not establish that a state is correctly represented as current.

Do not require transition wording, timestamps, historical labels, or any particular memory format. When supplied, measurement or event timestamps and structured fields may establish the update even across separate records or history tables. Record order, identifiers, and storage timestamps alone do not establish clinical event order. Do not use the gold update's ordering to resolve ambiguity in the memories. Evaluate each update independently; another gold update is never evidence.

Score-Reducing Examples

Gold update for the same left-breast lesion: old value "diameter 12 mm", new value "diameter 9 mm".
"Left-breast lesion diameter: 12 mm." plus "Left-breast lesion diameter: 9 mm." **0.5: both values are recorded, but which is current remains unresolved.**
"The left-breast lesion measures 12 mm in diameter." **0: only the old value is stored.**
No memory records either diameter of this lesion. **0: neither value is stored.**
"The left-breast lesion previously measured 9 mm and now measures 12 mm in diameter." **0: the update is reversed.**

Full-Credit Examples

Gold update for the same left-breast lesion: old value "diameter 12 mm", new value "diameter 9 mm".
"The left-breast lesion measures 9 mm in diameter." **1: the correct new value is stored; retaining the old value is optional.**
"The left-breast lesion diameter decreased from 12 mm to 9 mm." **1: the numerical update is explicit.**
"May 1 ultrasound: left-breast lesion diameter 12 mm." plus "May 8 ultrasound: the same lesion measures 9 mm." **1: examination dates establish the updated value.**

Return strict JSON only, with no Markdown fence or commentary. Include every supplied update ID exactly once. Use a score of 0, 0.5, or 1. Set `reflects_new_state` to true only for Score 1, report state presence independently, and list the relevant memory IDs.

```
{"results": [{"gold_id": "...", "score": 1,
"new_present": true, "old_present": false, "reflects_new_state":
true,
"reason": "...", "memory_ids": [...]}]}
```

Memory Retrieval Evaluation Prompt

You evaluate whether retrieved memories cover the evidence required for one question. Use only the supplied retrieved content and metadata, without external knowledge or clinical inference.

Task: Return one coverage label per gold evidence item.

Input:
Question: {`question`}
Gold Evidence: {`gold_evidence`} (items with IDs)
Retrieved Memories: `memories` (the saved top-K items and their associated metadata, in their recorded order; neither re-retrieved nor re-ranked, with similarity scores removed)

Criteria:

1. Covered. The retrieved content supports every required element of the gold evidence item, directly or through straightforward arithmetic from explicitly retrieved operands. Accept paraphrases, equivalent units, and structured fields with clear entity, laterality, and temporal bindings. Preserve uncertainty and negation. Support may span multiple retrieved items.

2. Not covered. Any required element is missing, only partially supported, incorrectly bound, or subject to an unresolved contradiction for the same entity, attribute, and applicable time. Clearly historical states do not contradict later states. Topical similarity alone is insufficient.

The question and gold items identify what to check but cannot supply missing retrieval evidence. Reference answers and candidate answers are not retrieval evidence. Do not infer missing measurements, diagnoses, or clinical outcomes.

Evaluate only the supplied question, considering the complete top-K list. Judge each gold evidence item independently and once, in the supplied order. Scores are aggregated externally according to Appendix B.3.

Not-Covered Examples

Gold evidence: "Right-eye pressure at the previous examination was 18."
Retrieved: "Left-eye pressure at the previous examination was 18." **Not covered: wrong laterality.**
Retrieved: "The pressure was 18." **Not covered: neither the eye nor the examination is identified.**

Covered Examples

Gold evidence: "Right-eye pressure increased by 3 from the previous examination."
Retrieved: "Previous right-eye pressure was 18." plus "Current right-eye pressure is 21." **Covered: the change is recoverable from correctly bound operands.**

Return strict JSON only, with no Markdown fence or commentary. Include every gold evidence item exactly once in the supplied order, copying its ID unchanged:

```
{"results": [{"gold_id": "...", "covered": true,
"reason": "..."}]}
```

D.5 GPT Pre-annotation Prompt

Pre-annotation Prompt

You are a pre-annotator for a longitudinal medical memory benchmark. Check generated content against the supplied clinical evidence, without external knowledge or clinical inference. Do not rewrite or repair items. Your labels are preliminary and physicians review them afterwards.

Task: One content type is supplied per call. Label every item as verified or flagged.

Input:
Content Type: {`content_type`} (dialogue turns, gold memory points, or QA pairs of one session)
Source Clinical Evidence: {`evidence`} (records available through the current session)
Items: {`items`} (items with IDs, speakers where applicable, and any supplied source-session provenance)

Criteria:

1. Verified. Every clinical factual claim and factual premise is supported by the applicable evidence, except for general terminology explanations handled below. Accept paraphrases, synonyms, and unit conversion or rounding supported by the supplied values and units. Items without clinical factual claims, such as greetings, are verified with an empty `evidence_quote`.

2. Flagged. Assign an applicable issue:

Unsupported: a clinical factual claim lacks supporting evidence.

Contradiction: a claim conflicts with the evidence for the same entity, attribute, and applicable time, including an incorrect value, unit, laterality, date, negation, or status.

Future_evidence: supplied provenance explicitly establishes that the item relies on evidence available only in a later session. Do not infer future provenance merely because a claim is absent or an event has not occurred; use unsupported or contradiction as appropriate.

For dialogue turns, assess clinical claims and factual presuppositions, including those embedded in questions. For gold memory points, require support from the evidence of the assigned session. For QA pairs, assess both the question's factual premises and its reference answer using evidence available through the current session.

General terminology explanations that add no patient-specific facts are left to physicians for medical-accuracy review. Do not flag them solely because they are absent from the patient record; note this

review boundary in `reason`. This exception does not cover added claims about the patient's diagnosis, measurements, treatment, or outcomes.

Evaluate each item independently. Quote relevant supporting or conflicting evidence, or provenance establishing future evidence. Leave `evidence_quote` empty when no applicable quotation exists; an empty quote alone does not require flagging. Ambiguous wording alone is left to physician review.

Flagged Examples

Source evidence for the current session: "Post-cycle-2 MRI: left-breast functional tumor volume 12.4 cm^3; no axillary lymphadenopathy. Surgery not yet performed." An earlier session records a baseline volume of 29.5 cm^3. Each example is independent.
Content type *dialogue turns* — "The volume is down to 12.4 cm^3, and pathology after your surgery showed no residual disease." **Contradiction: completed surgery is asserted despite the recorded status. Future provenance is not established.**
Content type *gold memory points* — "The post-cycle-2 MRI showed a left-breast volume of 12.4 cm^3 and a 1.2 cm axillary node." **Unsupported: the nodal measurement is not supplied.**
Content type *QA pairs* — "What was the Ki-67 index at this visit?" with reference answer "35%." **Unsupported: no Ki-67 value is supplied.**

Verified Examples

Same session and evidence.
Content type *dialogue turns* — "Thank you." **No clinical factual claim; evidence_quote is empty.**
Content type *gold memory points* — "The post-cycle-2 MRI recorded a left-breast functional tumor volume of 12.4 cm^3, with no axillary lymphadenopathy." **Supported by the current session's evidence.**
Content type *QA pairs* — "How did the left-breast functional tumor volume change after cycle 2?" with reference answer "It fell from 29.5 cm^3 to 12.4 cm^3." **Supported by the current and earlier sessions.**

Return strict JSON only, with no Markdown fence or commentary. Include every item exactly once. Set `issue` to `none` for verified items:

```
{"results": [{"item_id": "...",
"label": "verified | flagged",
"issue": "none | unsupported | contradiction | future_evidence",
"evidence_quote": "...", "reason": "..."}]}
```