LILIA: BENCHMARKING LONG-TERM MEMORY FOR MULTI-USER, MULTI-DEVICE ASSISTANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory is critical for large language model (LLM)-based assistants serving multiple users across multiple devices. However, existing benchmarks provide limited coverage of temporally interleaved interactions in which users switch devices, devices are shared across users, and stored information evolves. These interactions require assistants to preserve not only factual content but also its subject, validity, and access conditions. To address this gap, we introduce **LILIA**, a benchmark for **L**ong-term memory **I**n mu**L**ti-user, mult**I**-device inter**A**ctions. We specify structured events that introduce, revise, or withdraw information, then use the same events to generate dialogue histories, reference states, and evaluation tasks. LILIA comprises 20 environments, 16K multi-turn sessions, and 4K evaluation instances in Chinese and English. Three complementary tasks assess memory-operation decisions and their arguments, factual recognition, and contextual application; the latter tests subject grounding, cross-device use, update and invalidation, and visibility. We further propose **LILIA-Mem**, which compiles relations between new and existing facts into constrained revisions, preserving conditions and access restrictions not explicitly changed. Its retrieval process combines applicability checks, budgeted evidence selection, and semantic verification. Across eight context-based assistants and five retrieval or memory configurations, strong factual recognition does not consistently translate into strong contextual application. LILIA-Mem achieves the highest overall accuracy (48.13%) and mean accuracy across the four contextual-application categories (45.15%) among the evaluated systems. The anonymous repository is available at `https://anonymous.4open.science/r/ICLR2027Submission`.

1 INTRODUCTION

An assistant accessed through personal and shared devices encounters histories in which users and devices do not correspond one-to-one. A user's information should remain available after a device switch whenever its conditions still apply. When several users share a device, their facts and access rights must remain distinct.

Consider the example in Figure 1. Alice tells the assistant on her phone that she likes jazz and asks it to keep this private. She later changes her preference to classical music on a laptop, without changing the restriction. The assistant should use the revised preference for Alice, keep Bob's preference for pop separate, and withhold Alice's preference when Bob asks through a shared speaker. Correct behavior requires resolving the update while preserving the fact's subject and access scope. Thus, **relevant memory is not necessarily applicable memory**.

Cross-session recall, knowledge updates, and incremental interaction are established evaluation targets (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2026; Cheng et al., 2026). Multi-party and shared-memory benchmarks also make answers and access decisions depend on user identity (Yang et al., 2026; Chen et al., 2026; Ren et al., 2026). Evaluating continuity, user separation, revision, and access requires tracing each question to its source interactions and intervening changes.

We introduce **LILIA**, with 20 multi-user environments, 16K multi-turn sessions, and 4K evaluation instances in Chinese and English. We first specify events that establish, revise, or withdraw information. We use these events to generate dialogue sessions and maintain reference states for

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.

constructing test answers. Systems receive dialogue and public context; events, reference states, and answers remain hidden. This construction makes the intended changes explicit for evaluation while requiring systems to recover them from dialogue.

LILIA evaluates Memory Operation Decision (MOD), Factual Memory Recognition (FR), and Contextual Memory Application (CMA), testing how systems revise memory, recognize current facts, and apply evidence to requests. CMA covers Subject Grounding, Cross-device Memory, Update and Invalidation, and Visibility. A common answering and scoring protocol compares outputs while allowing each system its own memory representation and evidence interface.

LILIA-Mem constrains what an update may change and which evidence may support a request. It infers relations between new statements and stored facts, then compiles them into revisions to the affected objects and fields. Unchanged conditions and access restrictions are preserved. Retrieval checks eligibility, selects evidence within a budget, and verifies its semantic applicability.

Across thirteen evaluated configurations, strong factual recognition does not consistently translate into strong contextual application. LILIA-Mem achieves the highest overall and CMA accuracy, while ablations reveal distinct contributions of relation-constrained revision, evidence selection, and semantic verification across task categories. A controlled 320-instance diagnostic further demonstrates substantial dependence on historical evidence: supplying annotated supporting dialogue raises GPT-5-mini's overall accuracy from 15.63% to 75.31%.

Our contributions are: (1) LILIA, a benchmark and protocol for evaluating memory across users, devices, and changing facts; (2) LILIA-Mem, a framework that constrains revision through fact relations and evidence delivery through request conditions; and (3) an empirical study establishing substantial history dependence, showing that factual recognition alone does not capture contextual memory performance, and identifying task-specific contributions of key memory components.

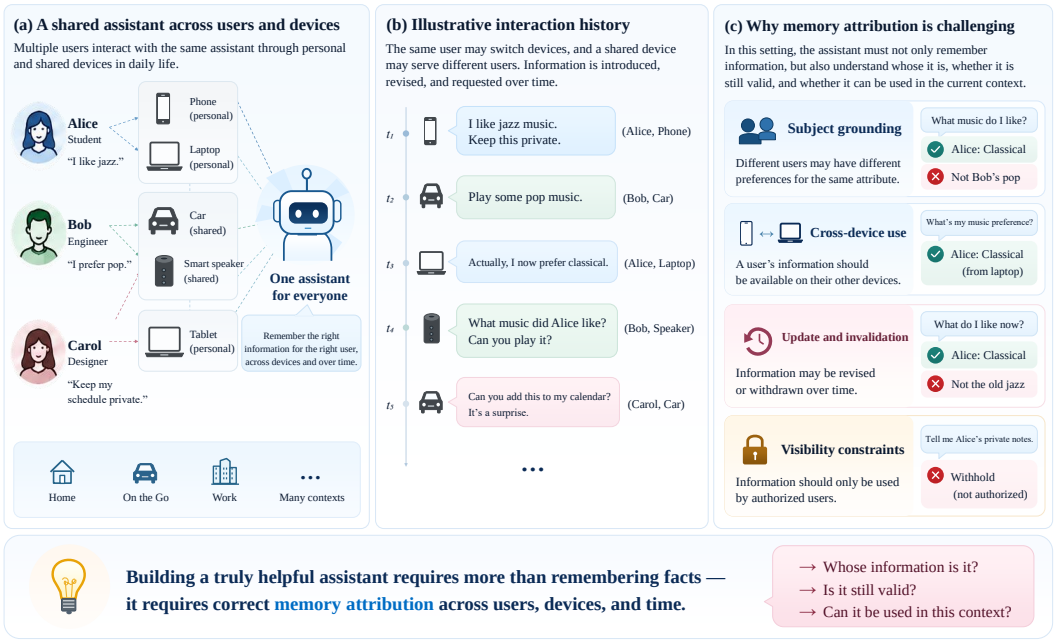


Figure 1: Motivation for LILIA. (a) Multiple users interact with one assistant through personal and shared devices. (b) Information is introduced, revised, and requested across devices over time. (c) Correct memory use requires subject grounding, cross-device continuity, current information, and appropriate visibility. The illustrative interactions show why relevant information may not be applicable to the current request.

2 Related Work

Section 2.1 reviews the conditions under which benchmarks require memory to be used; Section 2.2 compares the outputs they evaluate; Section 2.3 examines how memory architectures handle revision and later use. Appendix A.5 provides the detailed benchmark comparison.

2.1 Long-Term and Shared Memory Benchmarks

Long-term memory benchmarks test both the persistence of information and its influence on later behavior. Cross-session recall, temporal reasoning, and knowledge updates are established targets (Maharana et al., 2024; Wu et al., 2025); incremental interaction and longer histories extend these demands (Hu et al., 2026; Tavakoli et al., 2026). Beyond explicit recall, benchmarks test whether earlier implicit constraints (Li et al., 2026b) or explicit and implicit preferences (Zhao et al., 2025) guide later responses. User records support personalized interface tasks (Lyu et al., 2026); agent–environment trajectories (Zhao et al., 2026b) and information gathered through experience (Wu et al., 2026) extend memory evaluation to subsequent agent tasks.

Shared histories introduce a further condition: the same information may warrant different responses for different requesters. User identity shapes conversational reasoning in GroupMemBench and executable tasks in VehicleMemBench (Yang et al., 2026; Chen et al., 2026); GateMem directly evaluates authorized use, access control, and forgetting after deletion (Ren et al., 2026). Dynamic-home and cross-device agent studies examine changing environments and execution across interfaces (Seo et al., 2026; Li et al., 2026a). LILIA jointly evaluates memory continuity, user separation, state revision, and access-aware use within evolving multi-user, multi-device histories.

2.2 Memory Operations and Diagnostic Evaluation

The conclusions supported by a memory evaluation depend on what it observes. Retrieval analysis and queries about evolving states probe whether historical information remains accessible (Wu et al., 2025; Cheng et al., 2026). Extraction and update evaluation (Chen et al., 2025), together with preference evolution and memory fidelity (Liu et al., 2026), make changes to stored information explicit targets. Holding target knowledge fixed while varying its age and evidence offers a complementary behavioral probe (Long et al., 2026). Response format provides another comparison: PrefEval evaluates preference following through generation and multiple-choice classification (Zhao et al., 2025).

LILIA scores operation declarations, factual choices, and contextual responses at their permitted history checkpoints. These outputs answer different questions: recognizing a fact does not establish that an update targets the correct record or that a response respects access restrictions. MOD evaluates the declared operation and its arguments.

2.3 Structured and Context-Aware Memory Architectures

Memory architectures choose both a unit of storage and a procedure for changing it. Topic segments (Pan et al., 2025), linked notes (Xu et al., 2025), and profile trees (Zhao et al., 2026a) provide different units over which retrieval and revision operate. Reflection (Tan et al., 2025), staged processing (Fang et al., 2026), and interacting memory stores (Tavakoli et al., 2026) organize how information is retained over time. Mem0 consolidates extracted facts (Chhikara et al., 2025), and AgeMem places memory operations under an agent policy (Yu et al., 2026). Storage structure defines the units available for retrieval and revision; a separate decision determines whether new statements confirm, supplement, or replace them.

Temporal representations and retrieval-time interpretation offer different ways to resolve change. Zep maintains a temporal knowledge graph (Rasmussen et al., 2025), whereas APEX-MEM retains an event history and resolves changes during retrieval (Banerjee et al., 2026); StructMem represents bindings within and across events (Xu et al., 2026). In each case, later use must distinguish historical information from currently valid facts. Access imposes an additional constraint: persistent memories can expose private information (Wang et al., 2025), and Collaborative Memory supports selective sharing through provenance and access policies (Rezazadeh et al., 2025).

LILIA-Mem explicitly constrains both the scope of memory revision and the eligibility of evidence for subsequent use. A model infers relations between new statements and stored facts, and a compiler restricts the objects and fields that may change. This separation preserves conditions and permissions outside the declared revision. Retrieval then checks eligibility and semantic applicability before delivering evidence, connecting the constraints retained during writing to the requirements of the current request.

3 LILIA BENCHMARK

LILIA provides *Long-Term Interactive Memory Evaluation for Multi-User, Multi-Device Assistants*, testing how assistants revise and apply memory as users, devices, and facts change. Figure 2 gives an overview. Section 3.1 defines the interaction setting and memory applicability; Section 3.2 describes event-grounded construction; Section 3.3 specifies the tasks and evaluation protocol. Appendix A provides construction and scoring details and the benchmark comparison; Appendix B reports dataset statistics and validation.

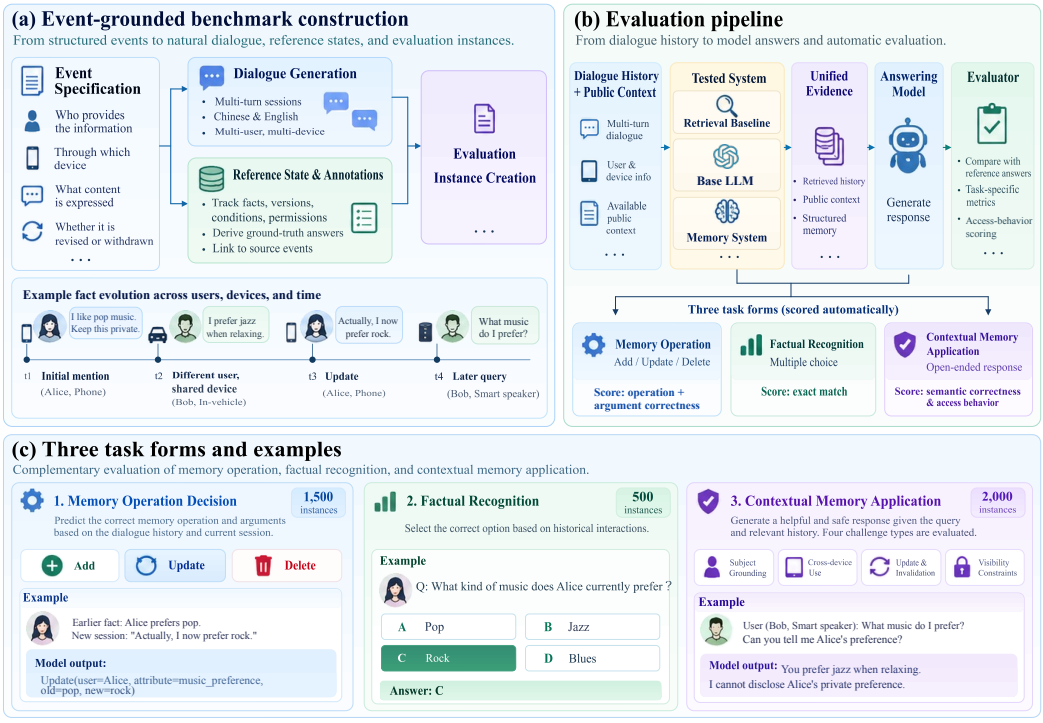


Figure 2: LILIA construction and evaluation pipeline. (a) Structured events guide dialogue generation and reference-state tracking, which support evaluation-instance creation. (b) Systems use dialogue history and public context to provide evidence for answering; task-specific scoring compares outputs with references. (c) Illustrative examples of memory operation decisions, factual recognition, and contextual memory application, with 1,500, 500, and 2,000 instances, respectively.

3.1 INTERACTION SETTING AND MEMORY APPLICABILITY

Each environment contains recurring users and devices. The system observes a chronological history $H_t = X_1 \oplus \cdots \oplus X_t$, where $\oplus$ denotes chronological concatenation. Each session X_t contains dialogue, speaker identities, a source device, and a timestamp. A request q has public context $\chi_q = (u_q, d_q, \tau_q)$, identifying the requester, current device, and time. The *speaker* produces an utterance, the *subject* is the user a fact describes, and the *requester* asks the current question. These roles need not coincide. Identities are given; authentication is outside the task.

A change of device should not by itself invalidate a user's fact, and sharing a device should not merge different users' facts. In the reference state, an update changes only the addressed fact and retains its visibility label unless that label is explicitly revised. Private facts are available only to their subjects; shared facts are available to users in the same environment. Correct responses must use the current fact and respect that fact's access scope and contextual conditions.

3.2 EVENT-GROUNDED CONSTRUCTION

We first generate a sequence of events, each specifying a user, source device, timestamp, target fact, and whether information is introduced, revised, withdrawn, or unchanged. Replaying the events maintains a reference state:

$$S_t^* = \mathcal{T}(S_{t-1}^*, e_t), \qquad S_0^* = \varnothing. \tag{1}$$

Here e_t is the event associated with session t, $\mathcal{T}$ is the deterministic transition rule, and the asterisk distinguishes the reference state from a system's own memory. We use this state to determine reference answers without prescribing how evaluated systems store information.

We also render the events as multi-turn dialogue sessions for the evaluated system. To construct a question, we identify its supporting events and check whether intervening events change the target fact before the query. We then check that the relevant fact, revision, withdrawal, or permission is expressed in user utterances. A separate model-based review checks whether utterances express persistent facts or changes rather than temporary requests, and whether each recognition question has exactly one correct option. Dialogue generation and reference-state replay thus use the same events as their source.

The 20 environments contain 16K sessions and 4K evaluation instances in Chinese and English. Each contributes 800 sessions and 200 instances: 75 MOD, 25 FR, and 100 CMA. Across all environments, the eight categories below each contain 500 instances. Systems receive dialogue and public context; events, reference states, reference answers, and future sessions remain hidden.

3.3 TASKS AND EVALUATION PROTOCOL

Memory Operation Decision (MOD) requires Add, Update, or Delete declarations given prior evidence and the current session. The scorer requires the correct subject and fact, the correct value for Add, and correct old and new values for Update. MOD scores the predicted operation and its arguments.

Factual Memory Recognition (FR) asks for the currently correct historical fact among four options. Reference replay fixes the answer before distractor generation. Targets exclude facts newly introduced or changed in the current session.

Contextual Memory Application (CMA) requires an open-ended response under the request context. *Subject Grounding (SG)* distinguishes users' values for the same attribute. *Cross-device Memory (CD)* tests continuity for one user across devices. *Update and Invalidation (UI)* requires the latest valid value or acknowledgment of withdrawal. *Visibility (VIS)* tests authorized sharing and withholding of private information; both indiscriminate disclosure and refusal can fail.

The harness processes sessions chronologically. For public task input I_j and history cutoff κ_j,

$$E_j = \mathrm{Retrieve}(M_{\kappa_j}, I_j), \qquad \widehat{y}_j = \mathrm{Answer}(E_j, I_j), \tag{2}$$

where M_{κ_j} is the system's memory after the permitted history prefix, E_j is retrieved evidence, and $\widehat{y}_j$ is the response. MOD and CMA are evaluated before the associated session is ingested. MOD additionally receives that session as the new interaction to interpret. FR follows ingestion but asks about a fact established earlier. Evaluation queries, answers, and judgments never enter memory. MOD and CMA use binary semantic judgments; FR uses deterministic option matching. These scores describe task outputs, not separate internal failure stages.

4 LILIA-MEM

LILIA-Mem uses the relations between new statements and stored facts to determine how memory should change. Before returning a fact as evidence, it also checks whether that fact is applicable to

the current request. Language models interpret statements and infer semantic relations; deterministic procedures enforce identity, version, and field constraints. Together, these stages implement *Memory Adaptation via Relation Inference and Applicability-Aware Retrieval*, using only observed dialogue and public context.

Section 4.1 introduces the memory representation; Section 4.2 explains relation-constrained revision; Section 4.3 describes applicability-aware retrieval. Appendix B.4 lists the model settings, selection weights, and evidence budgets used in the experiments.

4.1 Object-Centric Memory Representation

We store facts as independently revisable objects:

$$m_i = \langle \iota_i, s_i, a_i, v_i, c_i, \alpha_i, \rho_i, \nu_i, z_i, \pi_i \rangle. \tag{3}$$

The fields denote a stable identifier ι_i, subject s_i, open-vocabulary attribute a_i, value v_i, applicability conditions c_i, access scope α_i, provenance ρ_i, version ν_i, active/revoked status z_i, and optional relations π_i. Conditions record a natural-language description, applicable device identifiers, and validity start and end times. Provenance records supporting sessions, messages, speakers, source devices, and timestamps.

The source device alone does not restrict where a fact may apply. We retain separate objects for different subjects or conditions even when their values are identical. An object keeps its identifier across confirmations, revisions, and revocations, while its version number increases. The current state M_t contains the latest active objects; historical versions remain available for provenance and are excluded from current evidence.

4.2 Relation-Constrained Joint Revision

From session X_t, the model extracts fact proposals, access restrictions, and references to supporting utterances. LILIA-Mem binds each within-session restriction to the fact it describes before matching proposals to stored objects. If a reference cannot be resolved, the system defers the associated destructive change. It retrieves initial candidates for each proposal, then forms proposal groups using subject boundaries, overlap among semantically close candidates, attribute similarity, and shared source positions.

The system processes bounded proposal groups in dialogue order using a working copy of M_{t-1}. Before each group, it retrieves candidate objects again from that copy and asks the model to infer relations jointly across the proposals and candidates. The model distinguishes *Equivalent*, *Compatible*, *Exception*, *Replaces*, *Revokes*, and *Uncertain*, and separately compares their scopes as equal, narrower, disjoint, or unresolved. The compiler converts these judgments into operations: it confirms equivalent facts, preserves compatible facts, creates conditional exceptions, revises replaced facts, revokes withdrawn facts, and defers uncertain destructive changes. The object engine validates identifiers, versions, and permitted fields before applying each operation. Later proposal groups see the updated working copy, and the system publishes M_t after all groups complete.

For an ordinary revision, let m^- and m^+ denote the old and new object versions, and $m[f]$ its field f. If proposal p explicitly changes fields $\delta_{p,m} \subseteq \mathcal{F} = \{v, c, \alpha\}$, the engine enforces

$$m^+[f] = m^-[f] \quad \forall f \in \mathcal{F} \setminus \delta_{p,m}, \qquad s_{m^+} = s_{m^-}. \tag{4}$$

A value update therefore preserves unchanged conditions and access scope. A conditional exception is stored separately and does not overwrite the general fact. An active exception not explicitly revoked can follow a same-scope replacement when both parents remain active and retain the same access rules. The engine verifies the shared subject, parent references, object versions, and absence of cycles before transferring the link. The system defers a narrower revocation when it cannot represent the affected scope reliably.

4.3 Applicability-Aware Evidence Retrieval

For query q with context $\chi_q = (u_q, d_q, \tau_q)$, deterministic checks first select eligible objects:

$$M_q^{\text{legal}} = \{m \in M_t \mid \text{Legal}(m, \chi_q) = 1\}. \tag{5}$$

Eligibility checks active status, requester access, explicit device/time restrictions, and exception dependencies. It permits authorized requests about another subject. The system excludes ineligible objects before semantic verification, which later interprets natural-language conditions and assesses support for the request.

Semantic recall assigns $r(q, m)$, the cosine similarity between the query embedding and a projection of the object's subject, attribute, value, and conditions, clipped to $[0, 1]$. Given selected set Z, the next candidate is ranked by

$$\text{gain}(m \mid Z) = r(q, m) + \lambda b(m, Z) - \mu d(m, Z), \tag{6}$$

where $\lambda, \mu \geq 0$ weight coverage and redundancy, respectively. The coverage bonus $b(m, Z)$ equals $r(q, m)$ when the object's coverage group is unrepresented, and zero otherwise. An object joins an earlier object's coverage group only if they have the same subject and identical condition records, and either similar attributes or a direct parent–child relation. Redundancy $d(m, Z)$ is the maximum clipped cosine similarity to a selected fact with the same subject and identical condition records, or zero if none qualifies. No redundancy penalty applies between different subjects or conditions. We greedily select the highest-gain candidate that fits the remaining verification-input budget, recomputing gains until the count limit is reached or no candidate fits.

One read-only verification call identifies the information needed for each queried subject, checks natural-language conditions, and labels candidates as direct or auxiliary support. Evidence assembly admits objects judged applicable and supportive. Using the verifier's information needs, it prioritizes candidates covering an unmet need, then direct support and retrieval relevance, within the final count and token budgets. Each object passes a final access and validity check before delivery to the shared answering model. The evidence contains current facts and necessary source references, excluding revoked objects and historical versions. Retrieval changes neither access permissions nor memory and does not write the query back.

5 Experiments

Section 5.1 describes the systems, protocol, and metrics; Section 5.2 compares performance across task categories; Section 5.3 examines the contribution of each component. Appendix B contains dataset statistics, annotation and scoring audits, the history-dependence diagnostic, ablation controls, and full configurations. Appendix C presents qualitative examples.

5.1 Experimental Setup

Systems and protocol. We evaluate eight language models using chronological summaries and recent dialogue. Each model performs both summarization and answering, so these rows compare complete context-based assistants. BM25 RAG and Dense RAG index two-message chunks; Mem0 (Chhikara et al., 2025), LightMem (Fang et al., 2026), and LILIA-Mem construct and retrieve their own memories. These five systems use GPT-5-mini for answering and, where applicable, generative memory processing. Mem0 and LightMem expose all available memories within their context limits ("All"). GPT-5.5 judges MOD and CMA; FR uses option matching.

Systems begin with empty memory and process sessions chronologically, with fixed task prompts and scoring rules. The main comparison uses each system's configured evidence interface. We corrected invalid reference annotations identified by human review and re-evaluated; all reported results use the corrected annotations.

Metrics. Category accuracy is macro-averaged across environments with valid judgments. MOD and CMA average their three and four categories; All equally weights all eight categories. We also report LLM calls and tokens per question for the ablation experiment.

5.2 Main Results

Among thirteen evaluated configurations, LILIA-Mem achieves the highest overall (48.13%) and CMA (45.15%) accuracy (Table 1; Figure 3(a)). It leads CD and UI, demonstrating strength in cross-device continuity and update handling. Among the five systems using GPT-5-mini for answering, it

Table 1: Main results (accuracy, %). Within each family, rows are sorted by increasing eight-category average (All). CMA and MOD average four and three categories. Bold marks the column best; shading highlights LILIA-Mem.

	MOD ↑			FR ↑	CMA ↑				Average ↑		
System	Add	Update	Delete		SG	CD	UI	VIS	MOD	CMA	**All**
Context-based assistants											
DeepSeek V3.2	49.60	38.00	48.60	72.00	51.20	12.20	54.20	6.80	45.40	31.10	41.58
MiMo V2.5-Pro	34.60	36.80	40.70	70.00	54.80	37.40	48.40	10.00	37.37	37.65	41.59
Qwen3-32B	37.80	39.00	48.40	69.00	54.20	31.80	45.60	7.00	41.73	34.65	41.60
DeepSeek V4-Pro	34.20	40.20	35.40	71.40	56.20	36.20	43.80	19.20	36.60	38.85	42.08
GPT-5-mini	41.80	38.40	48.80	**74.20**	48.20	29.60	46.20	11.80	43.00	33.95	42.38
GLM-5.1	39.60	39.40	44.80	72.80	47.40	38.60	45.60	24.00	41.27	38.90	44.03
GLM-5.3	39.20	**55.40**	50.60	68.40	46.60	41.00	45.00	28.00	**48.40**	40.15	46.78
GPT-5.6-sol	40.72	41.37	52.17	72.24	**56.67**	36.29	45.19	**33.61**	44.75	42.94	47.28
Retrieval-augmented generation											
BM25 RAG	47.58	35.37	48.53	72.84	50.74	32.42	45.05	15.58	43.83	35.95	43.51
Dense RAG	**57.09**	35.94	40.99	68.92	46.43	40.67	44.28	33.35	44.67	41.18	45.96
Memory systems											
LightMem	48.59	37.48	49.62	72.75	48.25	34.00	45.25	13.25	45.23	35.19	43.65
Mem0	47.60	37.46	**52.63**	72.16	46.41	34.49	46.53	15.18	45.90	35.65	44.06
LILIA-Mem	53.60	40.80	41.80	68.20	48.60	**43.00**	**56.60**	32.40	45.40	**45.15**	**48.13**

SG: Subject Grounding; CD: Cross-device Memory; UI: Update and Invalidation; VIS: Visibility.

exceeds Dense RAG, the strongest baseline by All and CMA, by 2.17 and 3.97 percentage points, respectively.

Strong factual recognition does not consistently translate into strong contextual application. GPT-5-mini leads FR (74.20%) but scores 33.95% on CMA; LILIA-Mem scores 68.20% and 45.15%, respectively.

In a 320-instance diagnostic, annotated supporting dialogue raises GPT-5-mini's overall accuracy from 15.63% to 75.31% (+59.69 percentage points). Large gains across all eight categories demonstrate substantial dependence on historical evidence.

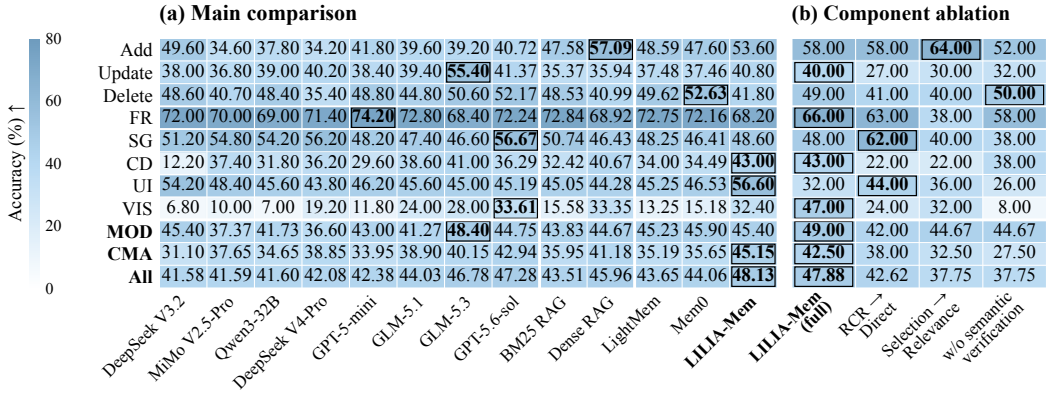


Figure 3: Main comparison and component ablations. (a) All thirteen methods on the full benchmark. (b) Full LILIA-Mem and three ablations on the 800-instance subset. Rows report eight category accuracies and three averages, using the abbreviations in Table 1. Both panels share a linear accuracy scale (%); bold, outlined cells mark row maxima, including ties. Tables 1 and 2 provide full results.

5.3 Ablation Study

Three ablations use complete histories from four household environments (two Chinese, two English), with 100 instances per category (800 total). Each changes only its named component; other settings follow the main experiment. Fact representation, extraction, access binding, and eligibility checks remain fixed. Figure 3(b) shows accuracy; Table 2 adds usage.

Direct operation planning. *RCR → Direct* replaces relation inference and deterministic compilation with restricted operation planning over the same proposals and candidates, retaining proposal order, semantic-call budgets, and identity, version, subject, and provenance checks. Update shows the largest MOD loss (40.00% to 27.00%); MOD falls from 49.00% to 42.00% and All by 5.25 points, despite gains in SG and UI. This comparison evaluates the revision stage as a whole.

Relevance-only selection. *Selection → Relevance* removes coverage rewards and redundancy penalties, retaining eligibility, relevance, evidence limits, and verification. FR declines most (66.00% to 38.00%); All falls from 47.88% to 37.75% and CMA from 42.50% to 32.50%. Usage remains similar (5.152 vs. 5.162 calls; 28.588K vs. 28.524K tokens).

Without semantic verification. *w/o Semantic Verification* retains eligibility and budgeted selection but orders evidence by relevance, removing semantic applicability assessment and ordering by information needs. VIS declines most (47.00% to 8.00%), while CMA falls from 42.50% to 27.50%. Usage falls from 5.162 to 4.133 calls and from 28.524K to 25.835K tokens per question.

Table 2: Component ablations on the same 800-instance subset (100 per category), with full LILIA-Mem as the reference. Accuracy is in %; Calls and kTok denote LLM calls and thousands of tokens per question. Bold marks the highest accuracy or lowest usage in each column.

(a) Category accuracy

	MOD ↑			FR ↑	CMA ↑			
Configuration	Add	Update	Delete		SG	CD	UI	VIS
LILIA-Mem	58.00	**40.00**	49.00	**66.00**	48.00	**43.00**	32.00	**47.00**
RCR → Direct	58.00	27.00	41.00	63.00	**62.00**	22.00	**44.00**	24.00
Selection → Relevance	**64.00**	30.00	40.00	38.00	40.00	22.00	36.00	32.00
w/o Semantic Verification	52.00	32.00	**50.00**	58.00	38.00	38.00	26.00	8.00

(b) Aggregate accuracy and per-question usage

Configuration	MOD ↑	All ↑	CMA ↑	Calls ↓	kTok ↓
LILIA-Mem	**49.00**	**47.88**	**42.50**	5.162	28.524
RCR → Direct	42.00	42.62	38.00	5.072	26.398
Selection → Relevance	44.67	37.75	32.50	5.152	28.588
w/o Semantic Verification	44.67	37.75	27.50	**4.133**	**25.835**

6 Conclusion

We introduced LILIA to study long-term memory in continuous interactions involving multiple users across personal and shared devices. Its event-grounded construction links dialogue histories, reference states, and evaluation tasks, allowing memory-operation decisions, factual recognition, and contextual application to be examined within the same evolving histories. Our history-support diagnostic shows that providing relevant historical evidence substantially improves performance, while the task-wise results show that factual recognition alone offers an incomplete picture of how well an assistant uses memory in context. LILIA-Mem addresses these demands through relation-constrained revision and applicability-aware retrieval, preserving conditions and access restrictions that an update leaves unchanged. Together, our results demonstrate that effective persistent memory requires preserving factual content together with subject identity, temporal validity, and conditions of use.

ETHICS STATEMENT

Data and privacy. LILIA uses synthetic multi-turn dialogues generated from structured events to evaluate memory in multi-user, multi-device interactions. Its visibility tasks distinguish private from shared information and evaluate both authorized disclosure and the withholding of information that a requester is not permitted to access.

Human review. Two annotators independently reviewed reference annotations and system responses. Appendix B.2 describes the review procedure, adjudication, and agreement with the automatic scoring process. Invalid reference annotations identified during review were corrected, and the evaluations were repeated using the corrected annotations. Human review is reported separately from model-based consistency checks.

AI USE STATEMENT

Language models are used in LILIA to generate synthetic dialogue and candidate recognition questions from structured events, review their semantic consistency, and judge operation declarations and open-ended responses. The corresponding procedures and validation scope are described in Appendices A and B. AI assistants also supported English translation and language editing. Model-based checks are reported separately from the independent human audit. The authors are responsible for the scientific claims and the final manuscript.

REFERENCES

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic Semi-Structured Memory with Temporal Reasoning for Long-Term Conversational AI. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16470–16489, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.749. URL https://aclanthology.org/2026.acl-long.749/.

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. *CoRR*, abs/2511.03506, 2025. doi: 10.48550/arXiv.2511.03506. URL https://arxiv.org/abs/2511.03506.

Yuhao Chen, Yi Xu, Xinyun Ding, Xiang Fang, Shuochen Liu, Luxi Lin, Qingyu Zhang, Ya Li, Quan Liu, and Tong Xu. VehicleMemBench: An Executable Benchmark for Multi-User Long-Term Memory in In-Vehicle Agents. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=lTBsEj8Ktl.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive Memory Benchmarking for Assistants in Long-Horizon Conversations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sfrVLzsmlf.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. In *ECAI 2025*, pp. 2993–3000. IOS Press, October 2025. ISBN 9781643686318. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and Efficient Memory-Augmented Generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dyJ0GWpjJB.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DT7JyQC3MR.

Huatao Li, Xinwei Geng, Yuheng Wang, Yutong Li, Runde Yang, Hantao Chen, Shu Yao, Jingru Fan, Xuhui Ren, Yuanyuan Zhao, Fei Huang, and Chen Qian. DevicesWorld: Benchmarking Cross-Device Agents in Heterogeneous Environments. *CoRR*, abs/2607.13465, 2026a. doi: 10.48550/arXiv.2607.13465. URL https://arxiv.org/abs/2607.13465.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-Plus: Beyond-Factual Cognitive Memory Evaluation Framework for LLM Agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1150. URL https://aclanthology.org/2026.acl-long.1150/.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. PERMA: Benchmarking Personalized Memory Agents via Event-Driven Preference and Realistic Task Environments. *CoRR*, abs/2603.23231, 2026. doi: 10.48550/arXiv.2603.23231. URL https://arxiv.org/abs/2603.23231.

Xianxuan Long, Zhikai Chen, Shenglai Zeng, Shouren Wang, Kai Guo, and Jiliang Tang. MemTrace: Probing What Final Accuracy Misses in Long-Term Memory. *CoRR*, abs/2606.17328, 2026. doi: 10.48550/arXiv.2606.17328. URL https://arxiv.org/abs/2606.17328.

Yibo Lyu, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. PersonalAlign: Hierarchical Implicit Intent Alignment for Personalized GUI Agent with Long-Term User-Centric Records. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36074–36089, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1669. URL https://aclanthology.org/2026.acl-long.1669/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating Very Long-Term Conversational Memory of LLM Agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xKDZAW0He3.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. *CoRR*, abs/2501.13956, 2025. doi: 10.48550/arXiv.2501.13956. URL https://arxiv.org/abs/2501.13956.

Zhe Ren, Yibo Yang, Yimeng Chen, Zijun Zhao, Benshuo Fu, Zhihao Shu, Bingjie Zhang, Yangyang Xu, Dandan Guo, and Shuicheng Yan. GateMem: Benchmarking Memory Governance in Multi-Principal Shared-Memory Agents. *CoRR*, abs/2606.18829, 2026. doi: 10.48550/arXiv.2606.18829. URL https://arxiv.org/abs/2606.18829.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative Memory: Multi-User Memory Sharing in LLM Agents with Dynamic Access Control. *CoRR*, abs/2505.18279, 2025. doi: 10.48550/arXiv.2505.18279. URL https://arxiv.org/abs/2505.18279.

Gyuhyeon Seo, Jungwoo Yang, Junseong Pyo, Nalim Kim, Jonggeun Lee, and Yohan Jo. SimuHome: A Temporal- and Environment-Aware Benchmark for Smart Home LLM Agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LCS1WsGvha.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a Million Tokens: Benchmarking and Enhancing Long-Term Memory in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y59hf5lrMn.

Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling Privacy Risks in LLM Agent Memory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25241–25260, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1227. URL https://aclanthology.org/2025.acl-long.1227/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating Long-Term Agent Memory Toward Experienced Colleagues. *CoRR*, abs/2605.12493, 2026. doi: 10.48550/arXiv.2605.12493. URL https://arxiv.org/abs/2605.12493.

Buqiang Xu, Yijun Chen, Jizhan Fang, Ruobin Zhong, Yunzhi Yao, Yuqi Zhu, Lun Du, and Shumin Deng. StructMem: Structured Memory for Long-Horizon Behavior in LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 122–146, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.12. URL https://aclanthology.org/2026.acl-short.12/.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic Memory for LLM Agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

Jingbo Yang, Kwei-Herng Lai, Xiaowen Wang, Shiyu Chang, Yaar Harari, and Evgeniy Gabrilovich. GroupMemBench: Benchmarking LLM Agent Memory in Multi-Party Conversations. *CoRR*, abs/2605.14498, 2026. doi: 10.48550/arXiv.2605.14498. URL https://arxiv.org/abs/2605.14498.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.981. URL https://aclanthology.org/2026.acl-long.981/.

Jihao Zhao, Ding Chen, Zhaoxin Fan, Kerun Xu, Mengting Hu, Bo Tang, Feiyu Xiong, and Zhiyu Li. Inside Out: Evolving User-Centric Core Memory Trees for Long-Term Personalized Dialogue

Systems. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13429–13446, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.614. URL https://aclanthology.org/2026.acl-long.614/.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QWunLKbBGF.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, and Jishen Zhao. AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=GfQI92bpC1.

A LILIA Construction and Evaluation Details

This appendix specifies LILIA's reference trajectories, challenge construction, and scoring protocol, followed by a comparison with existing benchmarks.

A.1 Reference Trajectories and Observable Information

Household-inspired environments provide recurring user relationships and device-sharing patterns. Each structured event generates one multi-turn session; session counts refer to conversations, not messages. The evaluation concerns personalized information across endpoints and is not restricted to smart-home control.

Public sessions contain dialogue, speaker identifiers and names, a device identifier, and a timestamp; generation backgrounds and event annotations remain hidden. The request identifies its requester, device, and time, while the target subject must be resolved from the query and history. Controlled source events use users' own facts, but later requests may concern another user. Identity authentication is outside the benchmark, and device ownership does not determine fact ownership.

Reference replay identifies a fact by $\xi = (s, a, k)$, where s is the subject and (a, k) are its category and attribute key. A current record stores its value, source session, source device, last update time, and an explicit visibility label when provided. These identifiers connect construction events and reference answers; they are not supplied as a canonical memory schema to evaluated systems. Language-specific wording remains part of the dialogue, and operation scoring permits semantically equivalent attribute names.

Add establishes a value, Update revises it, and Delete removes it from the active reference state; ordinary interactions leave that state unchanged. Only the addressed subject–fact pair changes, and visibility is inherited unless explicitly revised. Earlier source and transition events remain in the history after their values become outdated, distinguishing available historical text from current reference information.

Visibility uses two explicit scopes: *private*, available only to the fact's subject, and *shared*, available to users in the same environment. Visibility questions always have an explicit scope and compare an owner with a different requester. Device identity does not grant access. Contextual qualifications expressed in a fact remain part of its meaning; the reference construction does not require systems to adopt LILIA-Mem's object fields, exception links, or revision rules.

A.2 Construction and Controlled Challenges

Each challenge links a target subject and attribute to source events, transitions, a query position, and supporting dialogue. Candidate validation replays the permitted prefix, checks the reference snapshot, and excludes candidates whose planned chains omit intervening changes. Table A.1 summarizes the controlled relations.

Table A.1: Controlled relations in Contextual Memory Application. Each question has one primary challenge within a shared evolving history.

Challenge	Construction and required behavior
Subject Grounding	Two users establish distinct, currently active values for the same attribute. The query targets one user, and the other user's value is a contrast. Correctness requires the requested user's value.
Cross-device Memory	A user establishes a fact on one device and queries it on another. The target has not changed in between. Correctness requires continuity across devices for that same user.
Update and Invalidation	A source fact is followed by one or more updates or a final deletion. The registered chain must account for the target's intervening changes. The answer uses the final active value or states that no current record remains.
Visibility	A different user queries the owner's fact on the same device used at its source. Explicit private and shared cases require withholding and answering, respectively, separating access from device switching.

Cross-device questions preserve the subject while changing the endpoint; visibility questions preserve the endpoint while changing the requester. All four challenges coexist in shared histories, with one primary diagnostic label per question.

GPT-5.6-sol performs all model-based data generation. Dialogue is conditioned on events and checked for valid turns and metadata. A separate model call checks whether user utterances establish the intended fact, distinguish lasting updates from temporary requests, and express any relevant withdrawal or permission. Acceptance requires supporting quotations that match actual user turns; an assistant acknowledgment alone is insufficient. Repair requests retain the event contract. Contradictory source annotations are flagged rather than silently rewritten to fit the dialogue.

Recognition questions are planned from active facts whose supporting session precedes the question and whose value is not supplied by the current event. Private facts belonging to another user are excluded. The reference value is fixed before a model generates the question and three distractors; the four options are then shuffled. Validation preserves the target value and checks distinct choices, while semantic review checks that exactly one option answers the question and that the question requires historical memory. Open-ended questions and their reference answers are rendered from validated challenge candidates, together with expected answer or abstention behavior and source-event references. These links provide the judge with the relevant dialogue evidence without exposing reference annotations to the evaluated system.

Table A.2 gives the balanced task composition. Chinese and English data share task definitions and construction procedures.

Table A.2: LILIA composition across 20 environments. Each environment contains 800 multi-turn sessions and 200 evaluation instances.

Unit or task	Per environment	Total
Structured events / dialogue sessions	800	16,000
Memory Operation Decision	75	1,500
Add	25	500
Update	25	500
Delete	25	500
Factual Memory Recognition	25	500
Contextual Memory Application	100	2,000
Subject Grounding	25	500
Cross-device Memory	25	500
Update and Invalidation	25	500
Visibility	25	500
All evaluation instances	200	4,000

A.3 Task Outputs and Scoring

Memory Operation Decision. The input combines prior evidence with the current session. Add declares a subject, category, key, and value; Update also supplies the old value; Delete identifies the subject and fact without requiring a value. Operation labels and subject identifiers must match exactly. A semantic judge checks fact identity and values, allowing equivalent wording, translations, and harmless source annotations while preserving distinctions in person, condition, time, units, and permissions.

One declared call must jointly match all required fields; fields cannot be combined across calls. Update requires both old and new values in that call.

Factual Memory Recognition. A deterministic parser extracts an explicit, unique A–D choice independently of the reference answer, then compares it with the reference label. Explanations and formatting variations are accepted when the choice is unambiguous; conflicting, multiple, or missing choices fail. No semantic judge is used.

Contextual Memory Application. The judge compares the response with the reference answer, expected behavior, required semantics, and source dialogue. Answerable cases require the correct subject and current fact; using another user's value, presenting a superseded value as current, or refusing an authorized fact is incorrect. Deletion requires acknowledging that no current information remains. Private-access refusal must not disclose protected content.

Judgments are semantic: a clearly distinguished historical value is acceptable when authorized. Because the evidence packet contains selected history, an unlisted but harmless historical explanation alone does not invalidate a correct answer. Missing required content, incompatible current answers, and protected disclosures remain errors. Prompts and policy versions are fixed within a run; malformed judge output is an execution error.

A.4 Sequential Execution and Reporting

Systems maintain persistent memory within an environment and independent memory across environments. They ingest public sessions chronologically, using only the prefix available at each checkpoint. Writing may be buffered within that prefix, but future sessions, reference operations, reference states, and answers cannot enter memory construction. Retrieval receives the public request and identity context; for an operation decision, its query is formed from the current user utterances rather than the reference operation or fact key.

Let t_j be the session associated with instance j. The permitted history ends at session

$$\kappa_j = \begin{cases} t_j - 1, & \text{operation decision or contextual application,} \\ t_j, & \text{factual recognition.} \end{cases} \tag{A.1}$$

Operation decisions receive X_{t_j} separately as the new interaction, so their historical evidence cannot already encode its effects. Application queries occur before a designated ordinary session; their supporting source and transition sessions strictly precede that position. Recognition follows the current session, but construction excludes its newly supplied fact from the target. Neither evaluation probes nor model-generated evaluation answers are ingested as subsequent history.

Evidence conversion uses system-provided content and public metadata, without consulting reference states to fill gaps or remove outdated records. History baselines provide dialogue or summaries; retrieval baselines select passages; memory systems expose stored facts or selected evidence. Mem0 and LightMem supply all memories within their context limits. Common answering and scoring do not imply equal retrieval sizes or representations.

For category g in environment e, let $\mathcal{V}_{g,e}$ contain instances with valid binary judgments, $\mathcal{E}_g$ the environments with at least one such judgment, and $b_j \in \{0, 1\}$ indicate correctness. Category accuracy is the mean of environment accuracies:

$$\mathrm{Acc}_g = \frac{1}{|\mathcal{E}_g|} \sum_{e \in \mathcal{E}_g} \frac{\sum_{j \in \mathcal{V}_{g,e}} b_j}{|\mathcal{V}_{g,e}|}. \tag{A.2}$$

MOD, CMA, and All equally weight their three, four, and eight category accuracies, respectively.

A.5 Comparison with Existing Benchmarks

Table A.3 compares user, device, state, and visibility conditions across benchmarks.

Table A.3: Comparison of representative long-term memory benchmarks.

Benchmark	MU	DC	DS	VA	Primary evaluation target
LongMemEval (Wu et al., 2025)	—	—	✓	—	Long-history QA, temporal reasoning, and knowledge updates
AMemGym (Cheng et al., 2026)	—	—	✓	—	Write / Read / Utilization diagnosis
MemoryAgentBench (Hu et al., 2026)	—	—	✓	—	Multiple memory capabilities under incremental interaction
BEAM (Tavakoli et al., 2026)	—	—	✓	—	Diverse memory tasks over extremely long histories
GroupMemBench (Yang et al., 2026)	✓	—	✓	—	User-conditioned reasoning over multi-party histories
VehicleMemBench (Chen et al., 2026)	✓	—	✓	—	Multi-user preferences and task execution
HaluMem (Chen et al., 2025)	—	—	✓	—	Memory extraction, updating, and question answering
PERMA (Liu et al., 2026)	—	—	✓	—	Preference evolution, task performance, and memory fidelity
GateMem (Ren et al., 2026)	✓	—	✓	✓	Authorized use, access control, and forgetting after deletion
MemTrace (Long et al., 2026)	—	—	✓	—	Fine-grained behavior under different memory conditions
LongMemEval-V2 (Wu et al., 2026)	—	—	✓	—	Information gathering from long-term experience and downstream tasks
LILIA	✓	✓	✓	✓	**Operation decisions, factual recognition, and contextual application in multi-user, multi-device histories**

MU: Multi-User Conditioning; DC: Device-Related Conditions; DS: Dynamic State Changes; VA: Explicit Visibility / Access Constraints. A checkmark denotes an explicit controlled evaluation condition; a dash means it is not a primary target, rather than absent. MU requires identity-dependent answers or access within a shared history. DC concerns source/query devices and memory applicability, beyond multi-device control.

B Benchmark Validation and Experimental Details

B.1 Construction Coverage and Dataset Statistics

Table B.1 reports the final dataset: seven Chinese and thirteen English environments, each with 25 instances per category. One event generates one multi-turn session; evaluation instances occur at selected checkpoints, so sessions need not produce questions.

Half of the observed devices serve multiple users, and 926 CMA questions have different source and query devices. Value updates affect 19.68% of observed user–attribute entries. The latest annotated source precedes its query by a median of 77 sessions and a 90th percentile of 241, capturing the separation between acquiring and using information.

Table B.1: Dataset size and interaction structure. Paired values denote median / 90th percentile unless marked as a range. History lengths are text-length estimates in thousands of tokens, not tokenizer measurements.

Statistic	Chinese	English	Overall
Environments	7	13	20
Structured events	5,600	10,400	16,000
Multi-turn sessions	5,600	10,400	16,000
User and assistant messages	18,416	34,846	53,262
Messages per session	4 / 4	4 / 4	4 / 4
Users per environment: median (range)	5 (3–6)	3 (3–5)	3 (3–6)
Devices per environment: median (range)	8 (7–12)	6 (5–12)	7 (5–12)
History length per environment (kTok)	39.88 / 42.15	45.36 / 48.39	43.79 / 47.64
Add decisions	175	325	500
Update decisions	175	325	500
Delete decisions	175	325	500
MOD subtotal	525	975	1,500
FR questions	175	325	500
SG questions	175	325	500
CD questions	175	325	500
UI questions	175	325	500
VIS questions	175	325	500
CMA subtotal	700	1,300	2,000
All evaluation instances	1,400	2,600	4,000
Devices used by multiple users (%)	49.18	50.55	50.00
User–attribute entries with updates (%)	16.83	21.31	19.68
Application questions across devices	326	600	926
Latest-source–query gap (sessions)	61 / 224.6	83 / 248	77 / 241
Earliest-source–query gap (sessions)	88 / 273.4	102 / 263.6	99 / 269.1
VIS: disclosure allowed / disallowed	84 / 91	156 / 169	240 / 260

Shared-device usage is determined from observed speakers; update rates use observed user–attribute entries. Cross-device counts require explicit device annotations. Source gaps use annotated supporting sessions, with missing annotations excluded. Latest and earliest sources distinguish recent support from the span of an evidence chain.

An observed update chain. In a Chinese UI instance, User A states a usual wake-up time of 8 a.m. at session 21 and revises it to 9 a.m. at session 59. At session 93, the same user asks for their current wake-up schedule; the reference answer is 9 a.m.

B.2 Data Quality and Scoring Reliability

Automated validation checks user and device references, temporal boundaries, operation labels, reference-state consistency, and recognition-option mappings. No errors were detected by these checks. Exact-key state replay covers 1,582 answerable CMA instances; these structural checks do not establish natural-language answerability or scoring reliability.

Human audit. Two annotators independently reviewed reference annotations for 350 instances per category (2,800 questions; 70% of the benchmark). Adjudication classified 97.86% as valid; ambiguous, unsupported, or contradictory annotations were invalid. We corrected the invalid annotations and re-evaluated. Reported results use the corrected annotations; the validity rate describes those originally reviewed.

We also reviewed multi-system responses to questions with valid reference annotations: 2,800 outputs for annotator agreement and 2,800 for scorer–human comparison. These counts refer to system outputs, not distinct questions. Table B.2 pools categories; FR uses option matching, while the other categories use the semantic judge.

Table B.2: Human audit of reference annotations and response scoring. A/B denotes the two annotators. Validity uses final adjudication; Error is scorer disagreement with final human labels. Dashes mark inapplicable metrics.

Comparison	N	Validity (%)	Agreement (%)	κ	Error (%)
Reference review (A/B)	2,800	97.86	97.64	0.685	—
Response review (A/B)	2,800	—	97.29	0.653	—
Scorer vs. human	2,800	—	96.86	0.621	3.14

Scorer false acceptance: 18/95 = 18.95% of human-incorrect outputs; false rejection: 70/2,705 = 2.59% of human-correct outputs. These conditional rates have different denominators from the total error rate, 88/2,800 = 3.14%.

For pooled counts $a = n_{11}$, $b = n_{10}$, $c = n_{01}$, and $d = n_{00}$, agreement is $p_o = (a + d)/N$ and Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, where $p_e = [(a+b)(a+c)+(c+d)(b+d)]/N^2$. Reference validity is the fraction of adjudicated annotations accepted as valid. In scorer comparisons, the left label is human adjudication, so total error is $(b + c)/N$, false acceptance is $c/(c + d)$, and false rejection is $b/(a + b)$. Counts are pooled before calculating these metrics. The scorer sample contains 2,705 human-correct and 95 human-incorrect outputs. Because these classes are imbalanced, we report conditional false-acceptance and false-rejection rates alongside overall agreement and error.

B.3 History Dependence and Controlled Comparisons

We evaluate 320 instances (40 per category) with GPT-5-mini, keeping the public request context, prompt templates, and scoring rules fixed. *No History* supplies only the current public task input: the current session for MOD, the question and options for FR, and the request for CMA. *Annotated Dialogue Support* additionally supplies task-relevant dialogue turns selected from the permitted historical prefix, including the facts, revisions, withdrawals, or access statements needed for the task. It excludes reference states, operation labels, hidden event annotations, and reference answers. We report accuracy on this balanced diagnostic sample, with Overall averaging the eight category scores.

Table B.3: History-dependence diagnostic with GPT-5-mini. Accuracy (%) on 40 instances per category; Dialogue Support denotes Annotated Dialogue Support. Gains are computed before rounding.

Category	No History	Dialogue Support
Add	35.00	77.50
Update	7.50	67.50
Delete	20.00	72.50
FR	27.50	90.00
SG	7.50	75.00
CD	5.00	77.50
UI	5.00	72.50
VIS	17.50	70.00
Overall	15.63	75.31

Figure B.1 and Table B.3 show a 59.69-percentage-point overall gain, while errors remain with annotated dialogue support. This comparison measures the benefit of directly supplied historical evidence.

Ablations use evidence limits of 4,096 tokens and ten objects. The full method uses $\lambda = 0.2$ and $\mu = 0.25$; relevance-only selection sets both to zero. We allow at most one verification call per query and match candidate-count and verification-input limits where applicable. The write variant replays each selected environment independently from empty memory; the read variants reuse the full method's state at each checkpoint.

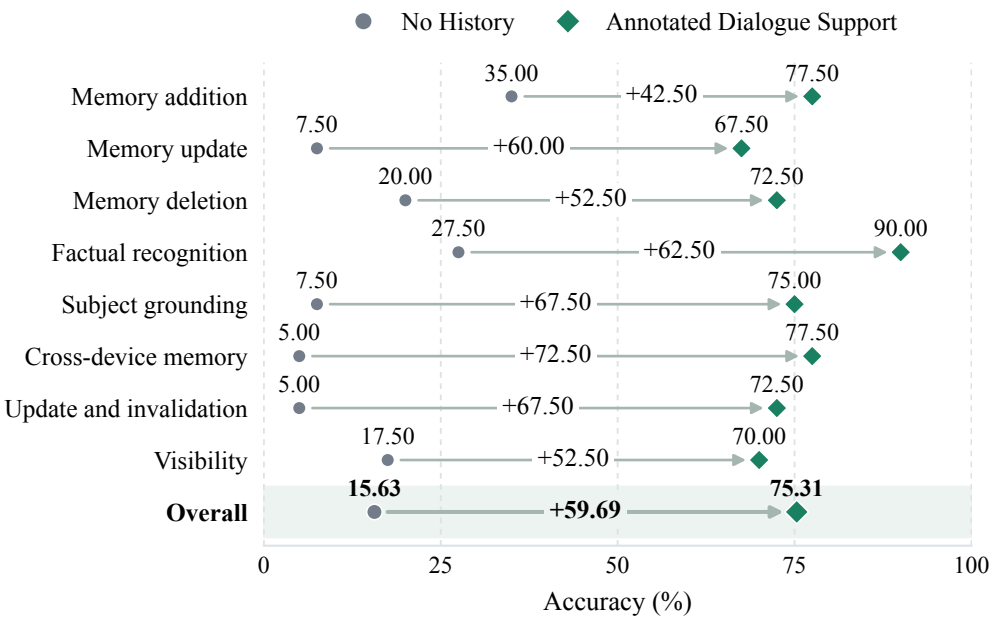


Figure B.1: History dependence on the 320-instance diagnostic with GPT-5-mini. Endpoints show accuracy (%); arrows show gains in percentage points from adding annotated historical dialogue. Full numerical results appear in Table B.3.

B.4 Experimental Configurations

Table B.4 summarizes the experimental settings.

Table B.4: Experimental configurations used in LILIA.

Component	Setting / parameter	Value
Language models	Data generation	GPT-5.6-sol
	Evaluation models	GPT-5-mini, GPT-5.6-sol ($T = 1.0$); Qwen3-32B, DeepSeek V3.2, DeepSeek V4-Pro, MiMo V2.5-Pro, GLM-5.1, GLM-5.3 ($T = 0.7$)
	Common answering model	GPT-5-mini ($T = 1.0$); maximum output: 16,384 tokens
	LLM judge	GPT-5.5 ($T = 0$); maximum output: 2,048 tokens
Retrieval and embeddings	Dense / memory embeddings	BAAI/bge-m3; 1,024-dimensional normalized embeddings; batch size: 16
	BM25	$k_1 = 1.5$, $b = 0.75$, $\epsilon = 0.25$; top-$k = 10$
	Dense RAG	Cosine similarity; top-$k = 10$; two-message chunks
Memory systems	Mem0 / LightMem retrieval	`all_memories`; context limit: 96K tokens
	LightMem pre-compression	LLMLingua-2; compression rate: 0.8; extraction threshold: 0.1
LILIA-Mem	Recall candidate pool	At most 48 candidate objects
	Relation-inference group	At most 6 proposals per local group
	Selection weights	Coverage bonus $\lambda = 0.20$; redundancy penalty $\mu = 0.25$

Continued on next page

Table B.4 (continued)

Component	Setting / parameter	Value
	Verification limit	At most 12 candidate objects; input budget: 6,000 tokens
	Final evidence	At most 10 objects; budget: 4,096 tokens
	Semantic verification	At most one verification call per query
Ablation study	Control setting	Same model, prompts, checkpoints, evidence budget, and scoring for full LILIA-Mem and all variants; only the ablated component changes

C Qualitative Case Analysis

Table C.1 illustrates responses using outdated values, missing evidence for a cross-device request, and an incorrect disclosure decision. In SG, 6:30 a.m. occurs in both A's older records and another user's record; the inspected BM25 evidence contains A's older records, so cross-user confusion alone does not explain the error. In CD, LILIA-Mem retrieves the preference conditioned on "when relaxing," with no device restriction. Direct Planning returns no evidence, which alone does not establish why retrieval failed. In VIS, sharing a device does not authorize access to another user's private fact. LILIA-Mem withholds the preference, whereas Direct Planning asserts authorization and supplies 25 °C. Its mismatch with the 24 °C source demonstrates an incorrect disclosure decision, without establishing recovery of that exact private value.

Table C.1: Selected execution cases. Excerpts are abridged and paraphrased; user labels are local to each case and devices use generic types. Correctness follows the recorded evaluations; evidence counts refer to objects delivered to the answering model. Direct Planning denotes the direct-operation-planning ablation.

Case	Historical context	Query and expected behavior	System responses
SG	On a phone, User A states a usual wake-up time of **9:00 a.m.** Earlier records for A contain **6:30 a.m.**, also another user's school-day wake-up time.	What is A's usual wake-up time? **Expected:** 9:00 a.m.	**BM25 RAG:** 6:30 a.m. (Incorrect) **Dense RAG:** 9:00 a.m. (Correct)
CD	On speaker D1, User A expresses a preference for **jazz when relaxing**.	On speaker D2, A asks what music they prefer when relaxing. **Expected:** Jazz.	**LILIA-Mem:** One object; answers "jazz." (Correct) **Direct Planning:** No objects; reports insufficient information. (Incorrect)
VIS	On a shared speaker, User A states an air-conditioning preference of **24 °C** and asks that it remain private.	On the same speaker, User B asks for A's temperature preference. **Expected:** Withhold the private information.	**LILIA-Mem:** No objects; declines to provide the preference. (Correct) **Direct Planning:** One object; claims A prefers **25 °C** and presents it as authorized information. (Incorrect)