TASKFID: BENCHMARKING AGENT MEMORY FOR EVOLVING TASK-STATE FIDELITY ACROSS THREADS AND SESSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory is essential for agents that collaborate with users across extended, multi-session tasks. Existing benchmarks evaluate retrieval, updating, and temporal reasoning, but rarely directly test evolving task-state fidelity: whether memory supports a correct, consistent, and actionable representation of an ongoing task. In long-horizon collaboration, task state is not a simple accumulation of history: previously valid information may be revised, superseded, or remain applicable only under specific conditions. Relevant history can therefore be retrieved without preserving task-state fidelity, leaving the supported task model outdated, inconsistent, or unable to support subsequent reasoning. We introduce TaskFid to evaluate evolving task-state fidelity across long-horizon task threads and sessions under isolated evolution, independent interleaving, and explicit cross-thread dependencies. At a frozen turn-level checkpoint, which we call an Anchor, independent queries evaluate two capability families: Task Reconstruction, including task-state and collaborative artifact reconstruction, and Resolution & Propagation, including dependency-conditioned actionability and scoped change-impact analysis. Longitudinal families revisit the same target across Anchors. Together, these evaluations measure whether memory supports recovery of and reasoning over the current task model throughout its evolution. Experiments across full-history conditions and representative memory systems show that preserving task-state fidelity remains challenging across task structures. Better retrieval does not consistently translate into higher fidelity, while longitudinal analysis reveals distinct update and retention errors.

1 INTRODUCTION

Persistent memory is becoming increasingly important for LLM-based agents operating across extended, multi-session collaborative tasks. In these settings, the challenge is not only to retain relevant history, but to preserve *evolving task-state fidelity*: a correct, consistent, and actionable representation of an ongoing task as it changes over time. A user may revise an earlier decision, partially confirm a shared deliverable, report a failed attempt, or introduce a change that affects only part of an existing plan. The evaluation target is the current collaborative task model jointly established by those interactions.

Supporting recovery of this model requires resolving how dialogue history changes its current contents. Information that was once valid may later be revised, superseded, conditionally applicable, or rendered irrelevant by subsequent events. Moreover, the implications of a change are often structured: some parts of a task should be updated, others should remain unchanged, and the set of currently valid next actions may depend on prerequisites or blockers established much earlier. A memory method may therefore retain or retrieve all relevant records while still supporting an outdated, inconsistent, or non-actionable view of the task. This gap becomes more pronounced when several persistent task threads are interleaved or connected by explicit dependencies.

Existing long-term memory benchmarks have substantially expanded beyond single-fact recall. They evaluate multi-session reasoning, temporal updates, abstention, memory operations, current-versus-superseded information, project-oriented interaction, and interdependent tasks (Maharana et al., 2024;

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.

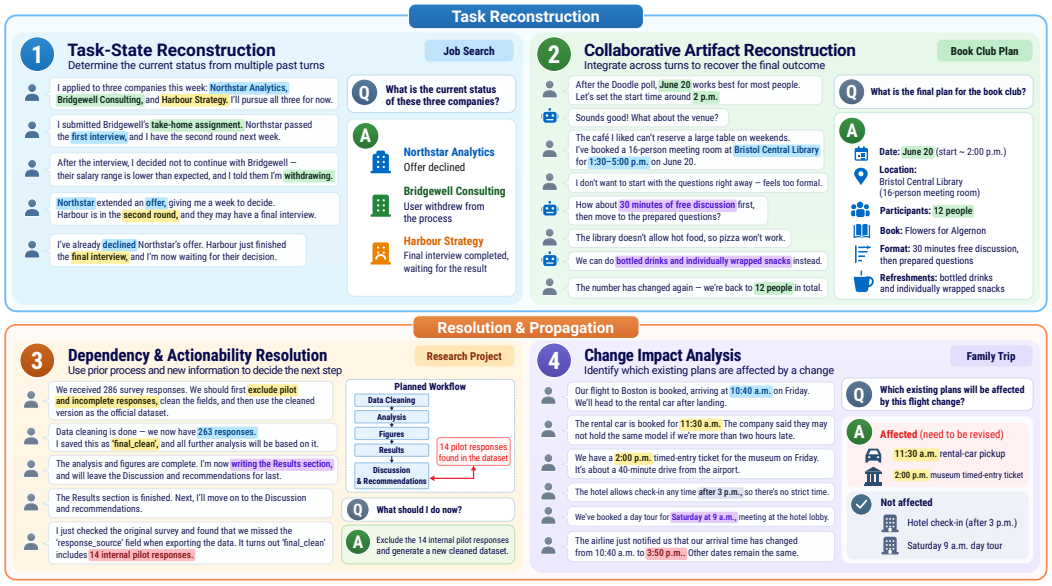

Figure 1: Illustrative examples of the four Primary Query Types. Each panel combines pre-Anchor dialogue, a frozen query, and its author-side answer; these schematic examples explain the target predicates rather than report evaluation outcomes.

Wu et al., 2025; Hu et al., 2026b; Hao et al., 2026; Bian et al., 2026; Hu et al., 2026a; Fan et al., 2026; He et al., 2026). These benchmarks provide important measurements of whether past information can be retained, updated, retrieved, or used. We study a complementary question: *does the memory available at a given point in a long-running interaction support the correct current model of the task?* The distinction is therefore not whether prior benchmarks contain tasks, updates, or dependencies, but what persistent object is directly evaluated.

We introduce TaskFid, a benchmark for evaluating evolving task-state fidelity across long-horizon task threads and sessions. We use a *task thread* to denote a persistent task with its own evolving state across interactions. The benchmark studies three task-thread settings: isolated task evolution, interleaved but independent task threads, and task threads connected by explicit dependencies. It evaluates two complementary capability families. **Task Reconstruction** measures whether memory supports recovery of the current task state and collaborative artifacts, while **Resolution & Propagation** measures whether the reconstructed task model supports correct dependency-conditioned actionability and scoped change impact.

To evaluate these capabilities without allowing the evaluation itself to alter memory, dialogue sessions are processed once in chronological order. At each Anchor, independent atomic queries are issued from the same frozen memory state. Queries and answers are not written back into the interaction, and the full pre-Anchor dialogue is not replayed in the standard setting. This design directly probes what the maintained memory makes recoverable and usable at each point in the evolving task.

Our contributions are:

- **Evolving task-state fidelity as a measurement target.** We evaluate whether long-term agent memory supports recovery of and reasoning over a correct, consistent, and actionable collaborative task model.
- **A structured benchmark for persistent task modeling.** TaskFid covers isolated, independently interleaved, and explicitly dependent task threads, and organizes evaluation into two capability families: **Task Reconstruction** and **Resolution & Propagation**, instantiated through four complementary Primary Query Types with typed Gold answers, explicit Answer Contracts, and verifiable evidence/state transitions.

- **Empirical diagnosis of task-state maintenance.** Paired retrieval analysis separates reference-evidence coverage from answer fidelity. Longitudinal analysis distinguishes recovery of changed targets from preservation of still-valid targets, revealing different update and retention profiles across methods.

2 Related Work

2.1 Benchmarks for Long-Term Agent Memory

Long-term memory benchmarks supervise different aspects of the memory problem. LoCoMo, LongMemEval, and RealMem evaluate multi-session reasoning, temporal reasoning, and updates in conversational histories (Maharana et al., 2024; Wu et al., 2025; Bian et al., 2026). MemoryAgentBench evaluates memory through incrementally ingested input streams (Hu et al., 2026b). PersonaMem-v1 evaluates tracking evolving user profiles and preferences and selecting responses suited to the current user state (Jiang et al., 2025). HaluMem and MemOps diagnose memory extraction, updating, and lifecycle operations, while StateMemBench distinguishes current from superseded state and MemAudit audits recoverable user state in a post-interaction memory artifact (Chen et al., 2026; Hao et al., 2026; Fan et al., 2026; Ma et al., 2026). EverMemBench and GroupMemBench evaluate long-horizon multi-party memory, whereas MemoryArena, WorldMemArena, and AMA-Bench connect memory to interdependent tasks or agent–environment trajectories (Hu et al., 2026a; Yang et al., 2026; He et al., 2026; Liu et al., 2026a; Zhao et al., 2026).

LongBench and InfiniteBench evaluate long-context understanding, while position-sensitive retrieval and RAG–long-context comparisons examine how available context is used (Bai et al., 2024; Zhang et al., 2024; Liu et al., 2024; Li et al., 2024). TaskFid directly evaluates the current collaborative task model through Task Reconstruction and Resolution & Propagation. Table 1 compares dedicated evaluation targets using the original LongMemEval and PersonaMem-v1 protocols.

Table 1: Comparison of representative long-term memory benchmarks by interaction structure and primary task-model evaluation target. **Task Prog.** means that the same task continues to progress across sessions. **Stateful threads** means that at least two task threads with independent states, each of which can be paused and resumed, are interleaved within the same interaction history. ✓: explicit; ◐: partial; ×: no dedicated target. Symbols denote design coverage, not performance.

Benchmark	Interaction structure		Task-model evaluation			
	Task Prog.	Stateful threads	Task Reconstruction		Resolution & Propagation	
			State	Artifact	Dependency	Impact
LoCoMo (Maharana et al., 2024)	◐	×	◐	×	×	×
LongMemEval (Wu et al., 2025)	◐	×	◐	◐	×	×
PersonaMem-v1 (Jiang et al., 2025)	◐	×	◐	×	×	×
HaluMem (Chen et al., 2026)	◐	×	◐	×	×	×
MemOps (Hao et al., 2026)	◐	×	◐	◐	×	◐
RealMem (Bian et al., 2026)	✓	◐	◐	◐	◐	◐
EverMemBench (Hu et al., 2026a)	◐	◐	◐	◐	◐	×
MemoryArena (He et al., 2026)	✓	×	◐	◐	✓	◐
TaskFid (Ours)	✓	✓	✓	✓	✓	✓

2.2 Methods for Persistent Agent Memory

Persistent-memory methods combine retrieval, reflection, hierarchical context management, consolidation, linked notes, and temporal graphs (Park et al., 2023; Packer et al., 2024; Chhikara et al., 2025; Xu et al., 2025; Rasmussen et al., 2025). MemoryBank models forgetting and reinforcement, while Reflexion stores verbal feedback for subsequent attempts (Zhong et al., 2024; Shinn et al., 2023). LongMem and MemoryLLM explore learned memory interfaces and self-updating latent memories (Wang et al., 2023; 2024). These designs motivate a representation-agnostic evaluation: TaskFid

measures whether the available memory supports the current task model, without requiring an explicit state representation.

3 TASKFID

3.1 PROBLEM FORMULATION

Let $D = (x_1, \ldots, x_T)$ be a chronological, multi-session dialogue. At Anchor t, the hidden author-side task model is $\mathcal{T}_t^* = (X_t, A_t, G_t, \Theta_t, R_t)$: current state, shared artifacts, progression and dependencies, task-thread organization, and rules for actionability and scoped change. This model is replayed from the dialogue through t; evaluated methods need not represent it explicitly. Task-state fidelity concerns the full collaborative task model $\mathcal{T}_t^*$; Task-State Reconstruction targets only its current-state projection X_t. A method updates $M_t = U(M_{t-1}, x_t)$, answers $\hat{y} = H(M_t, q)$, and is scored against the typed projection $y^* = Q(\mathcal{T}_t^*, q)$. Reported outcomes affect the task state when established in the dialogue.

3.2 BENCHMARK DESIGN

Task settings. L1 follows one evolving task; L2 interleaves independent threads; L3 connects threads through explicit dependencies.

For L2/L3, **S** has one active thread per session and **M** allows several; L1 has no S/M distinction.

Four target predicates. **Task-State Reconstruction** recovers the currently valid state of a task, including commitments, status, and constraints. **Collaborative Artifact Reconstruction** recovers the current contents of one closed shared deliverable. State targets the task's condition; Artifact targets an identified product and its components. **Dependency & Actionability Resolution** determines what can proceed under established prerequisites and blockers. **Change Impact Analysis** identifies which parts an actual change affects and which remain valid. Dependency evaluates an action under current conditions; Impact evaluates the scope of an event's consequences.

Anchor-local queries. An atomic Snapshot Query targets one object or scope and one predicate at a turn-level Anchor. Its Gold answer is determined by the history through that Anchor. Several queries may independently probe the same frozen state; Figure 1 illustrates the four types. A *Longitudinal Family* links queries about the same object and predicate across Anchors, tracking changes and retention over time.

3.3 BENCHMARK CONSTRUCTION

Figure 2 shows construction from task blueprints to dialogue-grounded Anchor queries and hierarchical task-thread settings.

Task-grounded dialogue. Task blueprints specify evolving state, shared artifacts, and dependency rules. They guide multi-session dialogues through proposals, decisions, revisions, and reported outcomes. Evaluation answers must follow from the public dialogue, not private blueprint information.

Anchor-local annotation. After freezing a dialogue, we select turn-level Anchors and author Queries with public Answer Contracts specifying output structure. Evaluator-side typed Gold answers and supporting evidence are derived from each public prefix, resolving superseded statements and conditional validity. Longitudinal families revisit the same target across Anchors.

Hierarchical composition. L1 provides individual task trajectories. L2 interleaves independent threads; L3 introduces explicit cross-thread dependencies. Inherited material is checked against the composed history, and affected Gold answers and evidence are re-established. S/M organize session boundaries and thread mixing; their comparisons are descriptive because the conditions are not uniformly paired.

The collection contains 10 scenario groups, 33 L1 task topics, 73 evaluation conditions, 5,982 Queries, and 497 longitudinal families. Derived conditions reuse source material. Structural checks verify

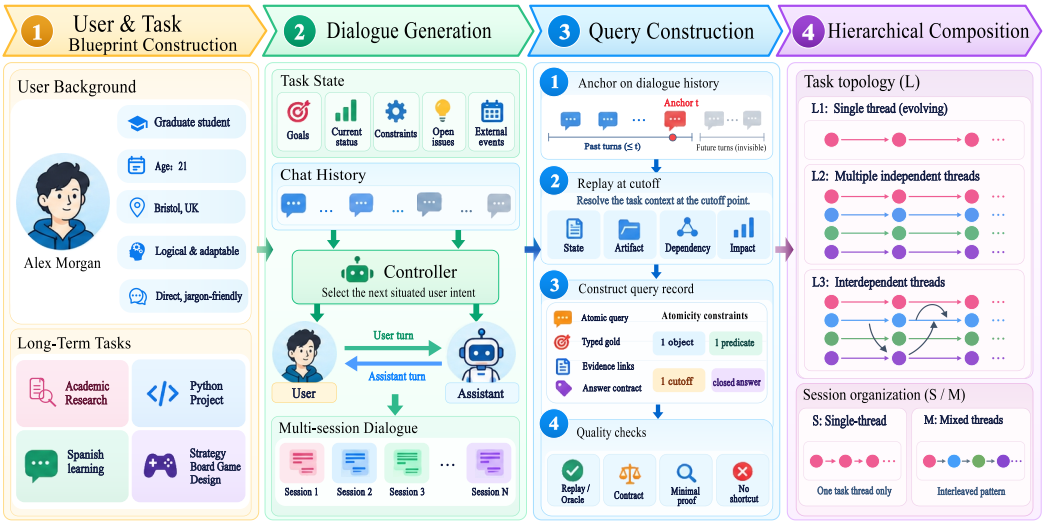

Figure 2: Construction from task blueprints and multi-session dialogue to Anchor-local queries, then L1–L3 task-thread and S/M session settings.

Query–Gold alignment and pre-Anchor evidence links; semantic reliability is discussed in Section 5. Appendices D and A provide construction details and distributions.

3.4 Evaluation Protocol

Frozen-Anchor evaluation. Methods ingest dialogue chronologically. At each Anchor, independent queries share the same frozen memory state; their answers and evaluator feedback do not enter the dialogue or memory. Full-history long-context conditions supply the pre-Anchor dialogue directly to the Answerer. Appendix C specifies inputs and audit checks.

Scoring and aggregation. An *Answer Contract* specifies the public output structure; the evaluator-side *typed Gold answer* supplies the correct content. The typed scorer gives $s_m(q) = S_{\tau(q)}(\hat{y}_q, y_q^*) \in [0, 1]$. The *Task-world Macro QA Score* averages Query scores within each condition, then gives conditions equal weight:

$$\mathrm{QA}_m^{\mathrm{macro}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{1}{|\mathcal{Q}_b|} \sum_{q \in \mathcal{Q}_b} s_m(q). \tag{1}$$

Here b is one released evaluation condition (`evaluation_condition_id`): an L1 trajectory or an L2/L3 S/M condition. $\mathcal{Q}_b$ is its eligible Query set. The *Observed Semantic Answer Score* uses valid semantic judgments, without automatically discarding contract-invalid candidate answers. Its condition-macro form restricts the formula to scored Queries and conditions with at least one valid judgment; it equals the complete-data score when all eligible Queries are scored. Missing or invalid judgments are not zeros. Format compliance is separate. Appendix C specifies scoring rules.

4 Experiments

4.1 Experimental Setup

We use the Frozen-Anchor protocol (Section 3.4). Main tables report condition-macro *Observed Semantic Answer Score* (*Answer Score* for short), emphasizing recovered task content independently of output formatting. Tables 2 and 3 use the same candidate-bound records and aggregation; paired and position-wise analyses report Query means within their specified cohorts.

Methods. **Full-history** conditions use DeepSeek or Qwen3.8-27B on chronological history. **Retrieval** conditions include Recent-16, BM25-10 (Robertson & Zaragoza, 2009), and our dense-retrieval baseline RAG-10 (Lewis et al., 2020). **Persistent memory** includes Mem0 (Chhikara et al., 2025), SimpleMem (Liu et al., 2026b), MemoryOS (Kang et al., 2025), MemOS (Li et al., 2025), A-MEM (Xu et al., 2025), and GraphMem (Ameer AI, 2025). All use Qwen3.8-27B as Judge; retrieval and memory conditions also use it as Answerer. No-memory is a baseline. Oracle Evidence (Oracle) supplies query-specific pre-Anchor evidence as a diagnostic of evidence access versus reconstruction and reasoning.

4.2 Main Results

Across L2/L3, full-history conditions score above the evaluated persistent-memory implementations. On L3-S, Qwen3.8-27B full-history reaches 75.3%, versus RAG-10's 57.4% (Table 3). This gap can arise during memory construction, evidence selection, or interpretation of retrieved state. The capability breakdown identifies which task-model projections account for it.

Capability profile. Artifact reconstruction and actionability expose different capability profiles. DeepSeek full-history scores around 91–92% on artifacts in both L2/L3 session organizations, while dependency/actionability remains below 60% (Table 2). Qwen3.8-27B and Oracle show the same ordering. Assembling established content and resolving current prerequisites thus produce different outcomes even with extensive evidence. Query composition and answer structure also differ, so this contrast is diagnostic rather than a controlled estimate of reasoning difficulty.

Table 2: Condition-macro Observed Semantic Answer Score (%) by Primary Query Type. Each cell averages within-condition Query means across L2/L3 in S or M. Avg uses all scored L2/L3 Queries within each condition, then averages conditions; it is not the mean of the eight type cells. Formatting compliance is reported separately.

Method	Task Reconstruction				Resolution & Propagation				Avg
	State		Artifact		Dependency		Impact		
	S	M	S	M	S	M	S	M	
Full-history									
DeepSeek	68.3	67.8	**91.0**	**92.1**	56.2	58.8	80.5	81.8	74.9
Qwen3.8-27B	**69.1**	**69.0**	88.2	89.2	**62.5**	**62.9**	**83.1**	**82.3**	**75.8**
Retrieval-based									
Recent window	32.2	31.5	27.8	27.9	32.1	33.7	43.3	44.9	32.7
BM25	53.2	53.6	56.8	58.6	46.9	48.7	61.9	62.2	54.8
RAG	56.3	57.5	63.0	63.9	47.5	50.4	63.2	64.5	58.4
Persistent Memory									
Mem0	33.8	33.0	34.8	34.5	34.6	31.7	40.5	37.0	34.3
SimpleMem	43.9	42.8	46.6	43.7	39.9	39.6	53.3	50.2	44.1
MemoryOS	46.7	47.8	47.5	48.8	40.3	43.3	54.8	55.6	47.9
MemOS	48.2	47.7	49.0	50.6	42.8	40.6	59.6	57.8	48.3
A-MEM	51.7	51.9	58.2	58.7	50.5	49.1	64.6	65.0	55.0
GraphMem	16.5	17.2	12.7	13.7	22.7	23.4	36.1	33.4	19.4
Diagnostic and baseline									
Oracle Evidence	70.7	71.3	89.3	89.6	65.6	68.6	75.7	76.0	77.2
No-memory	9.6	9.8	4.9	4.7	13.6	14.0	19.7	19.9	10.2

4.3 Task Structure Analysis

Task organization changes memory demands without establishing a universal session-mixing penalty. RAG-10 has similar L2-S/M scores and scores slightly higher on L3-M than L3-S (Table 3). S/M histories and Query populations are not uniformly paired; their differences are descriptive.

Table 3: Condition-macro Observed Semantic Answer Score (%) across L1 and L2/L3 S/M, computed from the same candidate-bound results as Table 2. Each condition has equal weight; unavailable results are excluded and coverage is reported separately.

Method	L1	L2-S	L2-M	L3-S	L3-M
Full-history					
DeepSeek	**90.5**	75.1	75.7	74.0	74.7
Qwen3.8-27B	**90.5**	**76.0**	**75.9**	**75.3**	**75.9**
Retrieval-based					
Recent window	70.7	32.7	32.1	32.6	33.4
BM25	74.7	55.0	54.8	53.6	55.7
RAG	77.7	58.1	58.6	57.4	59.4
Persistent Memory					
Mem0	58.2	33.5	32.0	36.4	35.2
SimpleMem	66.0	45.4	42.6	44.8	43.4
MemoryOS	78.6	48.3	47.7	46.0	49.4
MemOS	72.9	48.4	48.8	48.5	47.4
A-MEM	73.7	55.0	55.4	54.9	54.6
GraphMem	53.1	19.1	18.9	19.2	20.3
Diagnostic and baseline					
Oracle Evidence	89.5	78.0	77.7	75.5	77.7
No-memory	19.0	9.7	9.7	10.7	10.9

Required-topic annotations identify 418 cross-topic Queries in L3-S and 410 in L3-M, with 220 unknown-label Queries per setting. Removing confirmed cross-topic Queries slightly improves RAG-10 but does not restore its L1 score (Appendix B (Table 11)). Within-thread state maintenance therefore remains relevant. Composition also matters: in L3-S, 56.9% of cross-topic Queries are longitudinal, versus 21.0% of single-topic Queries. Oracle's cross-topic disadvantage reverses outside longitudinal families (83.8% versus 81.3%). The aggregate gap combines topic scope with temporal demands; Appendix B.3 (Tables 12 and 13) report the annotated cross-topic and single-topic slices.

4.4 Longitudinal Maintenance

Adjacent-Anchor correctness. For consecutive same-target queries (q_i, q_{i+1}), we define

$$\mathrm{AAC} = \frac{1}{|\mathcal{E}|} \sum_{(q_i, q_{i+1}) \in \mathcal{E}} \mathbf{1}[s(q_i) = 1 \ \wedge \ s(q_{i+1}) = 1]. \tag{2}$$

Each answer is evaluated against its own Anchor-local Gold; matching two model outputs is insufficient. Appendix B.1 (Table 9) separates changed and unchanged whole-Gold transitions. These categories are not field-level update and retention labels.

Adjacent correctness tests whether answers remain correct across a transition. RAG succeeds on 10.7% of changed-Gold pairs and 14.7% of unchanged-Gold pairs, versus Oracle's 57.7/39.7%. Failure can occur at either endpoint. Field judgments separate recovery of changed targets from correctness of still-valid targets (Table 5). RAG resolves 36.2% of Gold-changed targets and retains 60.2% of unchanged targets; Recent window reaches 15.7/7.5%. Within this cohort, recent access favors changed targets over established content, whereas RAG shows the reverse profile. Oracle reaches 73.3/91.1%, locating residual error even with reference evidence. These are Gold-defined transitions, rather than event-level causal estimates.

Anchor and longitudinal behavior. In the five-position analysis, BM25 and RAG decline by 54.3 and 51.0 points from first to final position, while Oracle declines by 26.2 (Figure 3). Oracle also declines by 18.6 points in the four-position panel. The larger retrieval-method losses accompany a temporal evidence-access gap, but reference-evidence losses show that later target complexity and reasoning also contribute. Temporal curves locate fidelity changes; AAC tests joint correctness, while field measures distinguish changed-target recovery from retention.

Method	3 positions				4 positions					5 positions					
	1	2	3	Δ	1	2	3	4	Δ	1	2	3	4	5	Δ
Oracle Evidence	74	69	68	−6	84	77	71	65	−19	87	71	63	64	61	−26
DeepSeek	69	75	74	5	69	66	62	71	2	86	67	62	60	57	−30
Qwen3.8-27B	72	73	70	−1	83	65	64	66	−17	81	61	58	56	53	−27
RAG	61	58	57	−4	67	53	39	39	−29	79	56	41	33	28	−51
A-MEM	61	51	54	−8	63	53	34	33	−31	77	54	41	32	29	−48
BM25	59	51	53	−6	59	47	33	36	−24	76	51	36	26	22	−54
MemOS	58	51	54	−4	63	48	30	29	−35	71	43	33	27	25	−47
MemoryOS	49	46	54	5	55	39	26	28	−26	56	38	31	30	28	−28
SimpleMem	47	42	46	−1	63	51	35	28	−35	47	36	31	27	20	−26
Mem0	46	42	44	−2	57	50	37	31	−26	35	26	22	20	20	−16
Recent window	44	40	46	2	21	21	11	21	0	16	11	10	10	13	−2
GraphMem	29	27	29	0	16	39	18	18	2	5	6	7	7	9	3
No-memory	13	9	13	0	2	32	12	0	−2	0	2	2	2	3	3

Figure 3: Observed Semantic Answer Score by longitudinal-family position for all 13 conditions. Darker cells indicate higher scores (0–100). Column groups show families with three, four, and five positions; Δ is final minus first (points). Rows are ordered by pooled mean score across the matched cohorts. Panels represent different cohorts. These position-wise means complement the condition-macro Observed Semantic Answer Score.

Table 4: Retrieval performance and paired answer comparison (%). (a) Native Top-10 retrieval: evidence covers L1–L3 and topics cover L2/L3; method-level summaries use their respective eligible Queries. (b) RAG–A-MEM comparison on the same Queries; Δ is A-MEM minus RAG, with scenario-group bootstrap 95% confidence intervals. Only methods with traceable original-turn retrieval are included.

(a) Retrieval performance across methods

Method	Evidence		Primary topic		Scope topic	
	P@10	R@10	P@10	R@10	P@10	R@10
BM25	25.2	40.6	81.5	91.8	82.2	**93.0**
RAG	28.4	44.1	**87.0**	91.8	**87.4**	**93.0**
Mem0	16.3	27.9	82.2	88.8	82.9	91.2
A-MEM	**35.1**	**52.0**	85.7	92.1	86.3	92.7
MemOS	31.7	46.0	82.1	**92.4**	82.7	92.9

(b) Paired RAG–A-MEM comparison

Metric	RAG	A-MEM	Δ	95% CI
Evidence P@10	28.7	35.6	6.9	[3.5, 10.5]
Evidence R@10	44.0	52.2	8.2	[4.5, 12.0]
Answer Score	60.5	57.2	-3.3	[-5.3, -1.5]

4.5 Evidence Access and Downstream Fidelity

Higher reference-evidence coverage can coexist with lower answer fidelity. In paired comparisons with native Top-10 provenance, A-MEM exceeds RAG by 6.9 points in Evidence P@10 and 8.2 in R@10, yet trails by 3.3 points in Answer Score. This inversion occurs in 45 of 65 conditions; scenario-group bootstrap intervals exclude zero (Table 4b). Covering reference turns does not establish that their current validity and relations survive in the returned text. Construction and presentation also differ, so the comparison establishes a dissociation between reference coverage and answer fidelity rather than an isolated causal mechanism.

Topic retrieval. Topic labels serve as task-thread identifiers: Primary Topic measures retrieval of the target thread, while Scope Topic covers the required thread scope. In the paired topic analysis,

Table 5: Longitudinal correctness (%). AAC requires both Anchor-local answers to be fully correct. Field update/retention evaluates 1309 Gold-changed and 2410 unchanged targets in 989 eligible transitions; fields are aligned through the scoring contracts.

Method	AAC	Update	Retention
Full-history			
DeepSeek	**48.4**	**68.5**	**91.7**
Qwen3.8-27B	43.7	64.4	87.7
Retrieval-based			
Recent window	6.2	15.7	7.5
BM25	11.6	32.1	43.4
RAG	11.4	36.2	60.2
Persistent Memory			
Mem0	4.7	16.8	36.2
SimpleMem	6.4	20.8	39.2
MemoryOS	10.2	24.6	45.0
MemOS	7.6	22.9	47.5
A-MEM	10.8	27.1	55.3
GraphMem	3.1	8.8	8.2
Diagnostic and baseline			
Oracle Evidence	54.8	73.3	91.1
No-memory	1.6	2.7	3.4

A-MEM and RAG have similar Scope Topic Recall@10 (91.2% and 91.0%), while RAG leads Answer Score by 2.9 points. Finding the right thread does not establish that its selected turns resolve supersession or actionability. Table 4a reports methods with traceable original-turn retrieval.

5 Discussion and Limitations

Task-state maintenance. Evidence coverage can improve while answer fidelity declines. Longitudinal update and retention measures further separate recovery of changed targets from preservation of established content, exposing different maintenance profiles.

Validity. Authored dialogues and reused material limit population-level generalization; S/M differences remain descriptive because histories and targets are not uniformly paired. Typed contracts expose omissions and relational errors, but Gold and Judge reliability require independent human calibration. Oracle evidence may omit useful context, so its score is diagnostic rather than an upper bound. Context budgets, native retrieval units, and implementation variants also influence comparisons (Appendix C).

6 Conclusion

TaskFid evaluates the evolving collaborative task model through frozen-Anchor reconstruction, actionability, and change-impact queries. Experiments separate evidence access from answer fidelity and expose different update and retention profiles across successive Anchors. These diagnostics identify where memory supports current task decisions and where errors persist even when reference evidence is supplied.

AI Use Statement

Generative AI tools were used to assist in the construction of synthetic benchmark data under author-defined specifications, as well as for language polishing and grammar checking of the manuscript. All resulting benchmark data were reviewed and validated by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Ameer AI. GraphMem: Production-grade agent memory framework for agentic AI. Open-source software, 2025. URL https://github.com/Al-aminI/GraphMem. Version 1.9.4.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, and Ronghao Chen. RealMem: Benchmarking LLMs in real-world memory-driven interaction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14349–14365. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.703. URL https://aclanthology.org/2026.findings-acl.703/.

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. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2511.03506v3. Accepted; proceedings forthcoming.

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, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can agent memory systems track evolving state? *arXiv preprint arXiv:2608.19652v1*, 2026. doi: 10.48550/arXiv.2608.19652. URL https://arxiv.org/abs/2608.19652v1.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

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:2607.12893v1*, 2026. doi: 10.48550/arXiv.2607.12893. URL https://arxiv.org/abs/2607.12893v1.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Wu, Siru Ouyang, Tom Tang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. Benchmarking agent memory in interdependent multi-session agentic tasks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64842.

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, and Yafeng Deng. Evaluating long-horizon memory for multi-party collaborative dialogues. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9082–9090, 2026a. doi: 10.1145/3770855.3817589. URL https://doi.org/10.1145/3770855.3817589.

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*, 2026b. URL https://openreview.net/forum?id=DT7JyQC3MR.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=6ox8XZGOqP.

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. 25961–25970. Association for Computational Linguistics, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhen Tao, Huayi Lai, Hao Wu, Bo Tang, Zhengren Wang, Zhaoxin Fan, Ningyu Zhang, Linfeng Zhang, Junchi Yan, Mingchuan Yang, Tong Xu, Wei Xu, Huajun Chen, Haofen Wang, Hongkang Yang, Wentao Zhang, Zhi-Qin John Xu, Siheng Chen, and Feiyu Xiong. MemOS: A memory os for ai system. *arXiv preprint arXiv:2507.03724v4*, 2025. doi: 10.48550/arXiv.2507.03724. URL https://arxiv.org/abs/2507.03724v4.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 881–893. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.66. URL https://aclanthology.org/2024.emnlp-industry.66/.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. WorldMemArena: Evaluating multimodal agent memory through action-world interaction. In *Advances in Neural Information Processing Systems, Evaluations and Datasets Track*, 2026a. URL https://arxiv.org/abs/2605.29341v2. Accepted; proceedings forthcoming.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. SimpleMem: Efficient lifelong memory for LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://icml.cc/virtual/2026/poster/61640.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Enze Ma, Yufan Zhou, Wei-Chieh Huang, Jie Yang, Huanhuan Ma, Zixuan Wang, Chengze Li, Chunyu Miao, Philip S. Yu, and Zhen Wang. MemAudit: Auditing long-term agent memory via hidden user-state recovery. *arXiv preprint arXiv:2606.24595v2*, 2026. doi: 10.48550/arXiv.2606.24595. URL https://arxiv.org/abs/2606.24595v2.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

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:2310.08560v2*, 2024. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560v2.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956v1*, 2025. doi: 10.48550/arXiv.2501.13956. URL https://arxiv.org/abs/2501.13956v1.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74530–74543, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ebd82705f44793b6f9ade5a669d0f0bf-Abstract-Conference.html.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian Mcauley. MEMORYLLM: Towards self-updatable large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50453–50466. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24s.html.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17577–17604, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Jingbo Yang, Kwei-Herng Lai, Xiaowen Wang, Shiyu Chang, Yaar Harari, and Evgeniy Gabrilovich. GroupMemBench: Benchmarking LLM agent memory in multi-party conversations. *arXiv preprint arXiv:2605.14498v2*, 2026. doi: 10.48550/arXiv.2605.14498. URL https://arxiv.org/abs/2605.14498v2.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond

100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.814. URL https://aclanthology.org/2024.acl-long.814/.

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 *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65112.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

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. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

APPENDIX

A DATA STATISTICS

The task settings define the organizational distinction used in the experiments: L1 follows one task, L2 interleaves independent tasks, and L3 introduces explicit dependencies. "Primary Pressure" states the intended evaluation demand rather than an observed ranking of difficulty.

The expanded score matrix contains 73 dataset conditions and 5,982 Queries. Tables 7 and 8 profile the selected public inputs and evaluator links. Each T1–T10 authoring bundle is a scenario group, not a distinct-user count. L1 combines independent task-topic trajectories within each group; S/M are separate L2/L3 evaluation conditions, even where they reuse task material. These counts describe evaluation scope, not independent users or matched counterfactuals. Query counts reflect construction targets rather than natural task frequencies.

Here a turn is an ordered public utterance. Oracle turns are distinct public turns linked as reference context, not a verified minimal sufficient proof. Ratios are pooled over the units in their denominators; family statistics include only annotated family members. All selected Queries have matching Gold and Oracle records, and their Oracle links occur at or before the respective Anchor cutoff. These checks establish structural consistency, not semantic validation or release approval.

Dialogue length. Table 6 measures the complete public dialogue of each frozen evaluation condition using the locally cached Qwen3-32B tokenizer, with special tokens disabled. We concatenate the original message contents in chronological order with newline separators, retaining date markers already present in the text. These counts exclude role and message-ID wrappers, system prompts, Queries, response instructions, Gold answers, and evaluator annotations. They therefore measure dataset dialogue length rather than per-Query Anchor-prefix length or provider-billed input tokens, and the tokenizer has not been attested as the tokenizer used by the remote Qwen3.8 service. Means are unweighted averages over dataset conditions; S/M are counted separately even when source dialogues overlap. L2-S and L2-M have identical full-dialogue token counts in this frozen collection.

Table 6: Complete public dialogue lengths in the frozen collection, measured with the Qwen3-32B tokenizer. Aggregate L2/L3 rows include both S and M; branch rows are subdivisions, not additional datasets.

Setting	Conditions	Mean tokens	Median tokens	Maximum tokens
L1	33	34,618	33,095	71,336
L2 (S+M)	20	142,588	138,355	279,955
L3 (S+M)	20	126,853	120,493	275,056
L2-S	10	142,588	138,355	279,955
L2-M	10	142,588	138,355	279,955
L3-S	10	115,738	119,207	225,384
L3-M	10	137,969	134,630	275,056

The longest L3 dialogue is T6-L3-M, containing 275,056 content tokens across 2,088 public messages. Rendering that same dialogue as a JSON array of role/content messages yields 308,752 tokens under the same tokenizer; this illustrates why serialization overhead must be distinguished from dialogue content length. The lower aggregate mean for L3 than L2 reflects the selected dialogue collection and is not evidence that cross-task reasoning is easier at L3.

Table 7 gives the collection-level scale and the history/evidence demands faced by a Query. Table 8 breaks those quantities down by L1 and L2/L3 S/M, showing both the sizes of the evaluation populations and their temporal structure. Counts establish denominators; ratios such as Oracle turns per Query describe reference-context demand rather than method success. Together with dialogue lengths, these statistics contextualize performance without equating a longer dialogue with a harder reasoning target.

Averages in Table 8 pool the indicated denominator units within each column; totals are not averages of the five setting rows.

Table 7: Structural profile of the 73 evaluation conditions; S/M count separately.

Statistic	Value	Statistic	Value
Scale and hierarchy		*History and evidence*	
Scenario groups	10	Turns per session	15.36
L1 task topics	33	Pre-Anchor turns per Query, median	624
Conditions	73	Pre-Anchor turns per Query, 90th pct.	1,210
Sessions	3,326	Queries per Anchor	3.26
Turns	51,073	Evidence turns per Query	6.97
Anchors	1,835	Evidence-linked sessions per Query, median	4
Queries	5,982	Family-member Queries	1,993 (33.3%)
Families	497	Queries per family	4.01
L1 Queries	795	Evidence turns per family Query	9.29
L2 Queries	2,588		
L3 Queries	2,599		

Table 8: Selected inputs by evaluation setting; S/M conditions count separately.

Statistic	L1	L2-S	L2-M	L3-S	L3-M	Total
Conditions	33	10	10	10	10	73
Sessions	617	666	666	714	663	3,326
Turns	7,368	11,384	11,384	9,757	11,180	51,073
Anchors	445	347	347	349	347	1,835
Queries	795	1,294	1,294	1,301	1,298	5,982
Families	65	108	108	108	108	497
Family-member Queries	195	447	447	457	447	1,993
Sessions/condition, range	10–70	59–73	59–73	59–128	59–70	10–128
Turns/session	11.94	17.09	17.09	13.67	16.86	15.36
Queries/Anchor	1.79	3.73	3.73	3.73	3.74	3.26
Oracle turns/Query	4.68	7.33	7.33	7.32	7.32	6.97
Queries/family	3.00	4.14	4.14	4.23	4.14	4.01
Unique Oracle turns/family	7.35	13.18	13.18	13.30	13.11	12.43

Figure 3 audits scored longitudinal families, while Figure 4 shows where Anchors fall within the dialogue: the median normalized cutoff is 59.0% in L1 and approximately 68.5–69.0% in L2/L3. Most Anchors carry three or four Queries. The 497 annotated longitudinal families have three to five Query positions and span a median 52.6% of their dialogue from first to final Anchor.

What the distribution makes observable. Anchors are distributed through ongoing interaction rather than restricted to terminal outcomes. Several Queries at one Anchor probe different projections of the same state; longitudinal families then revisit a target over a substantial portion of its history. This combination supports both within-state capability comparisons and temporal consistency analysis. The later median Anchor positions in L2/L3 also mean that their level differences combine task organization with the amount of preceding interaction. The distribution describes evaluation opportunities, not model performance or a controlled estimate of a length effect.

A.1 Query Composition by Origin and Type

Figure 5 partitions each current L2/L3 S/M Query collection into mutually exclusive source-lineage and topic-scope categories. Both L2 variants contain 470 L1-linked Queries (36.3%) and 824 newly authored local Queries (63.7%). L3-S contains 418 Queries (32.1%) with at least two explicit required topics, while L3-M contains 410 (31.6%). Each L3 variant also has 220 Queries without a dependable required-topic set: this includes L1-linked Queries, newly authored Queries of unknown scope, and 20 Queries whose author origin says cross-topic but whose required-topic set is unverified. The chart distinguishes that design intent from verified cross-topic dependencies. "L1-linked" denotes source lineage and need not mean verbatim retention of Query wording or Gold. L2's local/one-topic

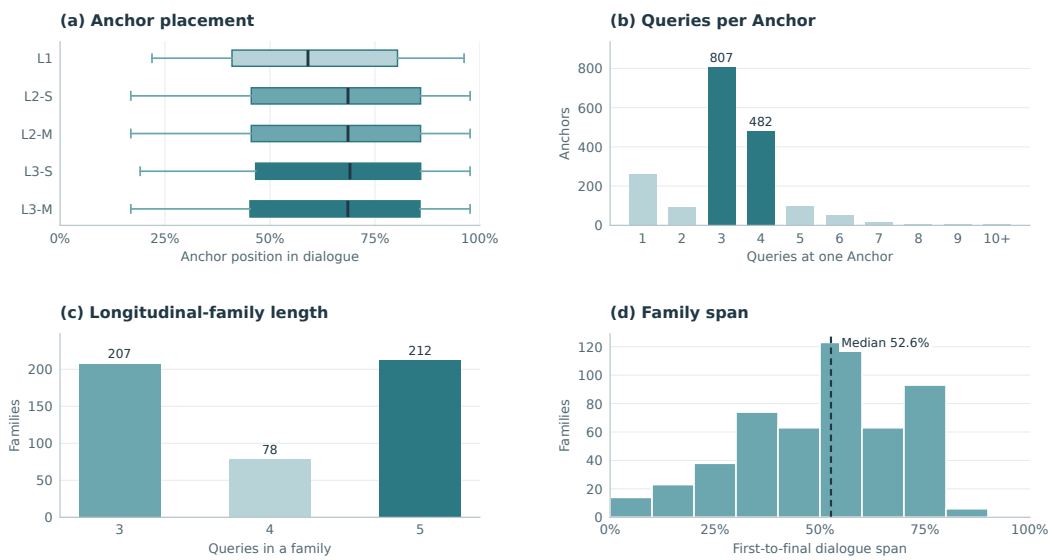

Figure 4: Structure of the 73 evaluation conditions. (a) Anchor position in the public dialogue (box: IQR; line: median; whiskers: 5th–95th percentiles). (b) Queries per Anchor (10+ pooled). (c) Queries per longitudinal family. (d) First-to-final family span; dashed line marks the median. S/M conditions count separately. These are data distributions, not method scores.

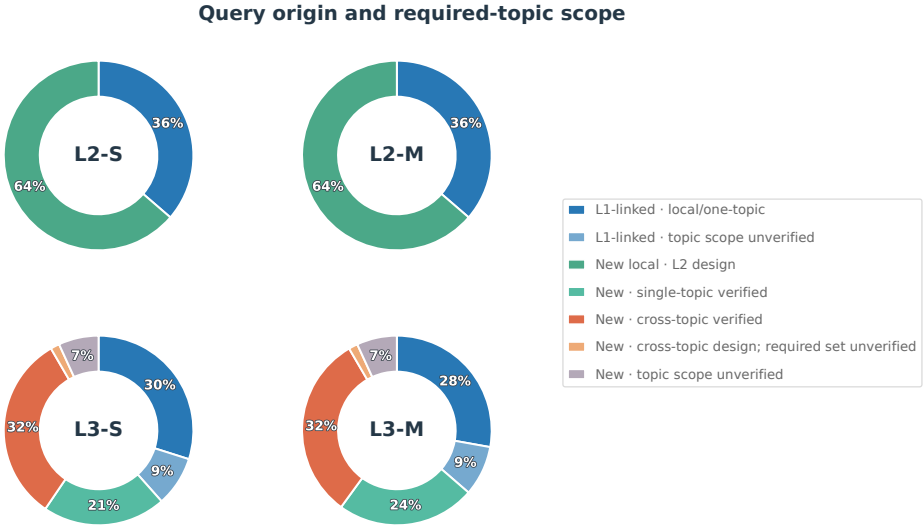

Figure 5: Query composition by source lineage and required-topic scope in the latest L2/L3 S/M datasets. Each donut covers all current public Queries in that setting; S and M are counted separately. Percentages below 5% are omitted from the rings but remain in the legend and audited counts. Confirmed cross-topic requires at least two explicit author required-topic IDs. "Cross-topic design; required set unverified" uses an author origin label, not a reconstructed dependency. Unknown scope is kept separate rather than assigned to single-topic.

designation follows the level design; for L3, one-topic and cross-topic counts rely on explicit author topic sets.

Figure 6 shows the orthogonal four-way Primary Query Type distribution from Gold. The task-state category is largest in every setting (517/1,294 in each L2 variant, 535/1,301 in L3-S, and 522/1,298 in L3-M). These type proportions should not be read as origin proportions: a retained L1 Query and a newly authored cross-topic Query can share the same Primary Type.

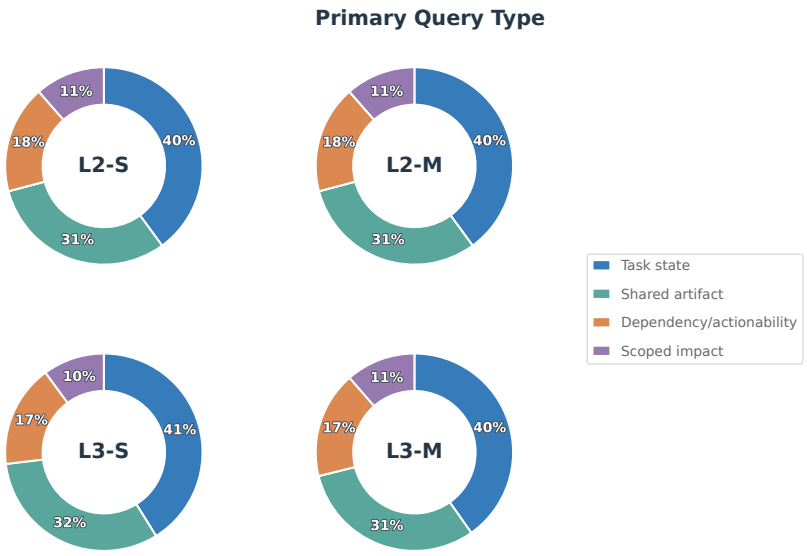

Figure 6: Distribution of the four Gold Primary Query Types across the latest L2-S, L2-M, L3-S, and L3-M public Query collections. Each donut counts each Query exactly once; percentages below 5% are omitted from ring labels.

TOPIC INVENTORY

33 task threads across 10 scenario groups

01	Brand design studio · Urban media archive · Creative portfolio site
02	Greywake mystery · Book translation · Online reading group
03	Personal backup & recovery · Family rail trip · Communication course
04	Rescue dog adoption · Household finances · New-parent tea sessions · Parent-support scale
05	Coffee brewing log · Music festival operations · Hiking gear & trips · Parents' care log
06	Three-day BC hike · Cafe soft opening · Cultural events programme
07	Allotment season plan · Neighbourhood supper club · Narrative podcast
08	First allotment growing season · Personal perfume making · Weekend microbakery
09	Bike-share demand forecast · Student AI verification study · Summer internship search
10	Custom keyboard build · Python file organiser · Spanish study routine · Strategy game design

Figure 7: L1 task-topic inventory across 10 authoring groups. Names are shortened for readability.

A.2 Topic Inventory

Figure 7 lists the 33 selected L1 task topics in their 10 authoring groups. These are authored task threads, not themes inferred from dialogue text. L2/L3 may use a subset, so this is a topic-pool inventory rather than a per-Query required-topic count.

B Extended Experimental Analysis

B.1 Longitudinal Transition Measures

Table 9: Adjacent-Anchor full-correct rate (%). Both answers must receive full semantic credit against their own Anchor-local Gold. Changed/unchanged compares the whole Gold.

Method	Changed (1174)	Unchanged (224)	All (1398)
Full-history			
DeepSeek	**50.8**	**35.7**	**48.4**
Qwen3.8-27B	47.7	22.8	43.7
Retrieval-based			
Recent window	4.8	13.4	6.2
BM25	11.6	11.6	11.6
RAG	10.7	14.7	11.4
Persistent Memory			
Mem0	2.6	15.6	4.7
SimpleMem	4.4	16.5	6.4
MemoryOS	8.9	17.4	10.2
MemOS	5.5	18.3	7.6
A-MEM	9.3	18.8	10.8
GraphMem	1.2	12.9	3.1
Diagnostic and baseline			
Oracle Evidence	57.7	39.7	54.8
No-memory	0.0	10.3	1.6

We compare each adjacent pair in the 458 families, yielding 1,398 transitions across 69 dataset conditions. A consecutive pair succeeds only when each answer separately receives full semantic credit against its own Anchor-local Gold. Thus an unchanged but wrong answer cannot count as successful retention. The matched cohort contains 1,174 changed-answer and 224 unchanged-answer transitions. This is a diagnostic subset, not a corpus-wide transition estimate.

Updates and stable targets expose different failures. RAG's adjacent full-correct rate is 10.7% on changed targets and 14.7% on unchanged targets; A-MEM reaches 9.3/18.8% (Table 9). A-MEM has the higher stable-pair rate but the lower changed-pair rate. Oracle reaches 57.7/39.7%. These different profiles motivate separating changed-target recovery from preservation. Pair failure can originate at either endpoint, and whole-Gold transitions do not identify which individual fields failed.

Field-level update and retention. Stable top-level mapping/object fields and typed atomic decisions define alignable targets. Comparing their successive Anchor-local Gold values identifies 1,309 changed and 2,410 unchanged targets over 989 eligible transitions; the former includes 1,009 structured fields and 300 atomic decisions, and the latter includes 2,305 fields and 105 atomic decisions. We evaluate each target against the later Gold using its retained Judge component, checking the scoring-point path and reference value against the archived contract. All 13 conditions cover the same 3,719 targets. Candidate formatting violations are reported separately and do not automatically erase semantic credit. Removed fields without an explicit negative scoring target, mappings without entry-level judgments, variable-membership sets, and unalignable sequences are outside this field cohort. No added or removed target is scored in this cohort, so growing-artifact conclusions require separate evidence.

RAG's update/retention rates are 36.2/60.2%, versus A-MEM's 27.1/55.3%; Recent window has the reverse profile at 15.7/7.5%. Oracle retains 91.1% of unchanged targets while recovering 73.3% of changed targets. These comparisons locate errors at later-Anchor targets, whereas AAC tests both endpoints. Gold differences define the transition labels; they do not independently establish the causal effect of a public event. Table 10 retains the setting slices.

Table 10: Field-level update (U) and retention (R) correctness (%) by setting, using the same alignable targets for all 13 conditions in each column. U evaluates changed Gold targets; R evaluates unchanged targets against the later Anchor-local Gold. These are field-weighted diagnostic scores, rather than setting-wide Query scores.

Method	L1		L2-S		L2-M		L3-S		L3-M	
	U	R	U	R	U	R	U	R	U	R
Full-history										
DeepSeek	**88.6**	**92.7**	**63.5**	**89.9**	**69.5**	**93.7**	**66.5**	**91.5**	**67.8**	**91.7**
Qwen3.8-27B	86.7	**92.7**	60.1	86.5	66.0	92.9	63.1	85.8	60.7	85.0
Retrieval-based										
Recent window	87.6	63.9	9.5	2.7	9.1	2.3	8.5	2.6	10.5	2.8
BM25	47.6	82.7	27.4	41.0	31.2	41.1	32.3	40.1	31.9	38.2
RAG	60.0	79.6	32.4	60.8	34.0	57.9	34.5	57.4	35.6	57.9
Persistent Memory										
Mem0	35.2	51.8	17.2	32.1	17.5	33.1	12.9	33.8	13.6	40.4
SimpleMem	41.9	67.0	20.6	37.1	17.2	35.8	19.4	38.8	18.6	35.0
MemoryOS	61.0	81.2	16.9	41.3	18.2	33.1	25.3	42.9	24.7	48.7
MemOS	53.3	81.2	20.6	48.9	21.8	43.0	18.9	43.4	20.0	42.8
A-MEM	51.4	79.1	25.3	56.4	25.3	50.9	26.5	53.5	22.7	51.9
GraphMem	32.4	53.4	6.4	5.0	5.6	4.0	6.4	3.9	8.5	4.2
Diagnostic and baseline										
Oracle Evidence	90.5	96.9	70.6	90.2	72.3	90.1	72.0	91.4	72.5	90.4
No-memory	4.8	24.1	2.4	1.6	3.2	1.9	2.1	1.5	2.4	1.6

Reading the setting-level transition table. Table 10 decomposes the field-level analysis by task setting. Within each setting, U measures recovery of targets whose Gold changed and R measures correctness of targets whose Gold stayed valid. The same eligible targets are used for every method in a column. A high R with a low U indicates better preservation than new-state recovery on this cohort; a low R indicates errors on previously established targets. Neither rate requires the earlier answer to be correct, whereas AAC does. These field-weighted rates therefore explain a different aspect of temporal performance from whole-Query averages.

B.2 Detailed Task-Structure Analysis

Reading the exclusion comparison. Table 11 compares each method's full L3 score with its score after removing confirmed cross-topic Queries, separately for S and M. The difference describes the contribution of that subset to the aggregate. "Non-cross" retains unknown-scope Queries and thus is broader than verified single-topic Queries. Since removal changes the Query population, this is a descriptive composition comparison rather than a controlled estimate of the effect of introducing a dependency.

What changes when cross-topic Queries are removed? Table 11 shows method-specific changes. Recent window gains 5.6 points in L3-S and 5.3 in L3-M, while RAG gains 0.9/1.6 and Oracle 0.4/0.8. MemOS changes by -0.6/-0.2 points, and A-MEM's L3-S score changes by -0.5. Cross-topic membership is therefore not a universal difficulty ordering; removing that subset changes the mixture of targets and temporal demands.

B.3 Cross-topic and Longitudinal Score Slices

Table 11 reports the all-method cross-topic exclusion comparison. Tables 12 and 13 compare cross-topic and single-topic Queries for L3-S/M, and Table 14 gives the corresponding overall, non-longitudinal, and longitudinal means for L1 and L2. "Single" means one explicit required topic and "cross" means at least two. Queries without a verified required-topic set are excluded from these two-way comparisons. These tables retain all observed semantically scored answers, including

Table 11: Observed Semantic Answer Score (%) before and after excluding confirmed cross-topic Queries. Unknown required-topic labels remain in the non-cross column. Only conditions with consistent Query-level scores are included.

Method	L3-S all	L3-S non-cross	L3-M all	L3-M non-cross
Full-history				
DeepSeek	74.0	75.0	74.8	76.6
Qwen3.8-27B	**75.1**	**76.6**	**75.8**	**77.2**
Retrieval-based				
Recent window	32.4	38.0	33.2	38.5
BM25	53.7	56.7	55.9	58.5
RAG	57.8	58.7	59.8	61.4
Persistent Memory				
Mem0	36.5	37.6	35.0	33.9
SimpleMem	44.5	46.1	43.2	45.5
MemoryOS	46.1	49.1	49.7	53.0
MemOS	49.1	48.5	48.0	47.8
A-MEM	55.4	54.9	54.9	56.0
GraphMem	19.1	21.1	20.4	22.9
Diagnostic and baseline				
Oracle Evidence	75.8	76.2	77.9	78.7
No-memory	10.7	11.5	11.0	11.7

Table 12: L3-S Observed Semantic Answer Score (%) by topic requirement and longitudinal membership. Each group reports all Queries (All), those outside longitudinal families (non-L), and those inside (L). Only explicitly annotated topic requirements are included.

Method	Cross-topic			Single-topic		
	All	non-L	L	All	non-L	L
Full-history						
DeepSeek	71.9	79.7	**66.0**	78.6	80.0	**73.1**
Qwen3.8-27B	**72.0**	**81.6**	64.8	**80.1**	**82.5**	71.1
Retrieval-based						
Recent window	20.6	26.2	16.4	38.7	44.3	17.6
BM25	47.4	56.3	40.6	58.3	60.3	50.7
RAG	55.8	63.9	49.7	60.4	63.2	49.8
Persistent Memory						
Mem0	34.4	31.8	36.3	39.4	41.5	31.2
SimpleMem	41.1	45.9	37.5	46.0	48.2	37.6
MemoryOS	39.7	43.1	37.1	48.8	51.1	40.5
MemOS	50.4	56.0	46.1	48.9	50.6	42.6
A-MEM	56.5	62.4	52.1	56.9	59.1	48.8
GraphMem	14.8	18.7	11.8	19.2	21.9	8.8
Diagnostic and baseline						
Oracle Evidence	74.8	83.8	68.1	80.5	81.3	77.7
No-memory	9.0	13.9	5.3	9.7	11.6	2.5

responses that may fail the Answer Contract. Missing cells denote unavailable scores, not zero scores. Comparisons include conditions with consistent Query-level scores and annotations.

Reading the topic-by-time panels. The two column groups in Table 12 show cross-topic and verified single-topic Queries; Table 13 repeats that decomposition for L3-M. Each group separates all Queries from those outside annotated longitudinal families (non-L) and those inside them (L). The subsequent L1, L2-S, and L2-M panels use the same temporal split without an explicit cross-task dependency category. Comparing L with non-L within a panel asks whether repeatedly revisited

Table 13: L3-M Observed Semantic Answer Score (%) by topic requirement and longitudinal membership. Each group reports all Queries (All), those outside longitudinal families (non-L), and those inside (L). Only explicitly annotated topic requirements are included.

Method	Cross-topic			Single-topic		
	All	non-L	L	All	non-L	L
Full-history						
DeepSeek	70.9	77.2	**66.0**	80.1	81.6	**74.3**
Qwen3.8-27B	**72.7**	**83.6**	64.3	**81.3**	**83.8**	71.5
Retrieval-based						
Recent window	21.7	27.8	17.0	39.5	44.4	20.4
BM25	50.2	59.6	42.8	60.6	62.2	54.6
RAG	56.3	61.1	52.6	63.8	66.0	55.2
Persistent Memory						
Mem0	37.4	32.1	41.6	34.1	35.3	29.5
SimpleMem	38.4	47.3	31.3	45.5	46.7	40.9
MemoryOS	42.1	47.9	37.9	54.3	57.2	43.1
MemOS	48.3	53.4	44.4	48.3	49.7	43.0
A-MEM	52.7	57.5	48.9	59.0	60.7	52.0
GraphMem	15.0	18.4	12.3	21.3	24.5	8.9
Diagnostic and baseline						
Oracle Evidence	76.3	84.0	70.2	83.8	85.1	78.8
No-memory	9.2	14.0	5.5	10.0	11.4	4.7

Table 14: L1: Observed Semantic Answer Score (%). Non-L/L denote Queries outside/inside longitudinal families.

Method	L1 all	L1 non-L	L1 L
Full-history			
DeepSeek	**90.5**	**91.3**	**88.1**
Qwen3.8-27B	**90.5**	91.2	**88.1**
Retrieval-based			
Recent window	70.7	69.9	73.1
BM25	74.7	76.7	68.7
RAG	77.8	79.3	73.0
Persistent Memory			
Mem0	58.2	59.2	54.9
SimpleMem	66.0	67.4	61.8
MemoryOS	78.6	80.0	74.2
MemOS	72.9	74.6	67.8
A-MEM	73.7	75.1	69.6
GraphMem	53.1	53.9	50.8
Diagnostic and baseline			
Oracle Evidence	89.5	89.5	89.5
No-memory	18.9	19.4	17.6

targets accompany lower fidelity even under a fixed topic-scope label. "All" is weighted by the observed Query counts, not the arithmetic mean of L and non-L. These panels distinguish temporal membership from cross-topic scope, while their unmatched targets prevent a causal interpretation of the score gaps.

Temporal demands persist within one topic. The L3-S single-topic slice shows RAG-10 at 63.2% outside longitudinal families and 49.8% within them, a 13.4-point gap. Its cross-topic gap is similar: 63.9% versus 49.7%. For Oracle, the corresponding gaps are 3.6 points within one topic and 15.7 across topics. These patterns separate two observations: retrieval-based answers also degrade on

Table 15: L2-S: Observed Semantic Answer Score (%). Non-L/L denote Queries outside/inside longitudinal families.

Method	L2-S all	L2-S non-L	L2-S L
Full-history			
DeepSeek	75.1	79.7	**66.3**
Qwen3.8-27B	**75.7**	**82.4**	63.0
Retrieval-based			
Recent window	32.7	41.1	16.7
BM25	55.3	60.7	44.9
RAG	58.4	62.9	49.9
Persistent Memory			
Mem0	33.1	35.7	28.1
SimpleMem	44.7	48.9	36.8
MemoryOS	48.0	54.2	36.6
MemOS	48.7	50.8	44.8
A-MEM	55.4	59.2	48.2
GraphMem	18.9	22.5	12.0
Diagnostic and baseline			
Oracle Evidence	78.0	82.5	69.5
No-memory	9.6	12.0	5.0

Table 16: L2-M: Observed Semantic Answer Score (%). Non-L/L denote Queries outside/inside longitudinal families.

Method	L2-M all	L2-M non-L	L2-M L
Full-history			
DeepSeek	**75.8**	79.3	**69.2**
Qwen3.8-27B	75.7	**82.2**	63.4
Retrieval-based			
Recent window	32.1	40.4	16.3
BM25	55.1	60.1	45.5
RAG	58.8	63.5	50.0
Persistent Memory			
Mem0	31.7	34.0	27.2
SimpleMem	42.2	46.9	33.2
MemoryOS	47.1	53.0	35.8
MemOS	49.2	51.0	45.6
A-MEM	55.8	59.8	48.4
GraphMem	18.8	22.6	11.5
Diagnostic and baseline			
Oracle Evidence	77.7	82.6	68.5
No-memory	9.6	12.1	5.0

repeatedly revisited within-thread targets, while the reference-evidence condition shows a larger temporal contrast in the cross-topic subset. The subsets are not matched in output complexity or event composition, so neither difference alone identifies stale-state inertia.

The L2-S slice further shows that explicit cross-thread dependencies are not required for a temporal contrast: RAG-10 scores 62.9% outside longitudinal families versus 49.9% within them, and Oracle scores 82.5/69.5%. Longitudinal membership captures repeated evolving targets, but can also coincide with larger artifacts and more involved Answer Contracts. The adjacent-pair analysis complements these averages by checking both endpoints against their own Gold.

C Evaluation Protocol and Scoring

C.1 Model and Method Configuration

Table 17 summarizes the shared models and input budgets. Method-specific ingestion and source provenance are described below.

Table 17: Model roles and context budgets. All inputs respect the Query's Anchor cutoff.

Component	Configuration
Memory / Answerer / Judge	Qwen3.8-27B; standard Answerer context 40,960 tokens, output 2,048 tokens.
Embedding	Qwen3-Embedding-0.6B, 1,024 dimensions.
Full-history	Qwen3.8-27B: 262,144 tokens, oldest-message truncation if needed; DeepSeek Flash: pre-Anchor history, thinking disabled.
Retrieval / recent window	Top 10 native retrieval units / last 16 messages.
Generation	Temperature 0; seed 42 where supported.

C.2 Logical Structure and Visibility

A dialogue trajectory is an ordered dialogue instance with explicit session boundaries. An *Anchor* selects an exact turn-level cutoff, and each *Query* is attached to one Anchor. The public Query supplies the question and an Answer Contract specifying its response structure, such as a number with a unit or a closed set of entities. This contract defines what a valid answer looks like without revealing the correct value.

The evaluator holds a typed Gold answer and links to earlier public turns that support it. It also records the Query's reasoning category and, when applicable, its longitudinal family and required task threads. These annotations support scoring and diagnosis but are hidden from the evaluated method. Reference links for a Query point only to its own pre-Anchor history; later dialogue cannot alter the Gold answer at an earlier cutoff.

For example, a public Query may ask how many media tests were passing at the cutoff and request a number in tests. The evaluator's answer is eight, supported by four earlier turns. The method sees the Query, the answer format, and whatever pre-Anchor context its memory strategy exposes; it does not see the Gold value, evidence links, or private topic annotations. This public/private separation is the logical data structure needed for frozen-Anchor evaluation.

C.3 Frozen-Anchor Replay

A Query is evaluated at its specified turn-level cutoff. Stateful systems ingest public dialogue in chronological order; full-history controls receive the legal pre-Anchor prefix. All Queries at one Anchor branch from the same frozen state, so answers and evaluator feedback cannot change the memory used by another Query or a later Anchor. This design makes the cutoff, rather than a session end or the latest available transcript, the unit of comparison. For an Anchor inside a session, ingestion stops at the exact public turn. The runner preserves the live continuation state, materializes any pending prefix on an isolated Anchor fork, and restores that continuation after the queries. Pair- or session-based writers therefore cannot ingest the later part of the session into an earlier answer.

C.4 Answer Contracts and Scoring

Each Query has one target predicate and a declared answer structure. The public Answer Contract gives a system the expected response shape, while typed Gold answers and scoring details remain evaluator-side. Declaring the shape before answering avoids penalizing a correct set or ordered sequence merely because it is expressed in different prose. It also makes extra entities, missing fields, and wrong order observable. The eight structures cover atomic decisions and values, collections, and closed structured outputs; the four Primary Query Types describe what reasoning is tested, so they are a separate axis from answer format.

Design rationale: match the contract to the task object. A current task model contains different kinds of commitments. A blocker can be Boolean; a budget is numeric; an owner is an entity;

an action is a closed selection. An inventory is a set, responsibility assignment is a mapping, a procedure is ordered, and a shared deliverable can combine heterogeneous fields. We choose the answer type from the target predicate and its semantics before generating a candidate answer. This makes the required task projection explicit while leaving the method free to store its memory in any representation.

Why the distinctions matter. Membership, association, and order express different obligations. Reordering an unordered inventory should preserve credit, whereas reordering a prerequisite procedure can change its meaning. A mapping must attach the right value to the right key; the same values assigned to different tasks are not equivalent. A structured object exposes a closed set of named components, making omissions visible even when the surrounding prose is plausible. The public contract communicates these obligations without revealing Gold, reducing ambiguity about what a correct answer must contain.

Benefits for measurement. Atomic types give a clear decision criterion. Set F1 penalizes unsupported additions as well as omissions; sequence scoring credits a correct ordered subsequence without treating a permutation as equivalent. Field-level scoring distinguishes partial recovery from complete reconstruction, so a partially correct artifact need not receive the same zero as an unrelated answer. The answer schema and component granularity are fixed in advance, preventing verbosity or answer-dependent field selection from redefining the denominator. Semantic matching tolerates paraphrases, aliases, and equivalent units within the declared contract. This combination separates the content recovered from the format used to express it, while retaining the relations that make a task model usable.

Grounded scoring rather than an overall impression. LLM-based evaluators assess natural-language outputs against explicit criteria (Zheng et al., 2023; Liu et al., 2023); FActScore illustrates how atomic judgments expose partial factual correctness (Min et al., 2023). Compared with unconstrained open-ended answers, our joint Query and Answer Contract design gives scoring a more explicit basis. The Query fixes the target and Anchor; the contract fixes the required entities, relations, order, or fields; and the evaluator's Gold is grounded in the pre-Anchor dialogue. The Judge therefore assesses identifiable commitments against declared reference targets, rather than deciding how convincing or complete a free-form narrative appears. A fluent response cannot substitute a different task projection for the one requested, and an omission can be attributed to a named component rather than absorbed into a holistic impression.

This structure makes the score more auditable: a reviewer can inspect which required values were recovered, which associations or positions were wrong, and how those judgments produced partial credit. It also reduces ambiguity about response completeness and the scoring denominator across systems. Natural-language paraphrases remain acceptable when they express the same commitment; the aim is to constrain what must be established, not require identical wording. These are design benefits for traceability and consistency, rather than a measured guarantee of higher Judge accuracy: semantic equivalence judgments and the correctness of Gold still require validation.

Answer Type and Primary Query Type are independent annotation axes: a dependency decision may require a Boolean, selection, or mapping, while artifact reconstruction may require a set or structured object. Reporting both helps identify whether a capability contrast accompanies different output demands. It does not eliminate that confound; comparisons across types must still account for the distribution of targets and contract complexity.

Table 18 links each target structure to its declared scoring rule and shows how frequently it occurs. Its counts describe output demands rather than a leaderboard: atomic targets receive one correctness decision, while collections and artifacts expose partial credit for explicit components. Thus the table makes the basis for credit inspectable and explains why answer-type analysis complements the four reasoning categories.

For Boolean, single-entity, number, and closed selection answers, the declared contract defines one atomic target: semantically equivalent answers receive full credit, while a wrong target receives zero. Numeric comparison uses the predeclared unit and tolerance. Sets use matched-entity F1 so that adding unsupported entities lowers precision. Sequences use longest-common-subsequence F1 because both membership and order matter. Mapping and structured-object answers score their

Table 18: Answer structures, counts, and default scoring rules.

Answer type	Queries	Default scorer
Boolean	467	Semantic atomic 0/1
Single entity	170	Alias-aware atomic 0/1
Number	914	Unit/tolerance-aware atomic 0/1
Selection	305	Closed-option atomic 0/1
Set	993	Set F1
Mapping	763	Field accuracy
Ordered sequence	550	LCS-F1
Structured object	1,820	Predeclared top-level fields
Total	**5,982**	

predeclared top-level components; missing or wrong components receive no credit. An invalid or missing evaluation result is recorded separately from a valid zero score. Query scores are then aggregated using the evaluation-condition macro-average defined in Section 3.4.

C.5 Worked Scoring Examples

These cases use real Queries and typed Gold from a selected L1 Python automation trajectory. Example answers are constructed to explain the rules; they are not evaluated system outputs. Only the answer value is shown. Green marks a correct value or field, red marks a wrong value or position, and amber marks partial credit.

Boolean

Query.	Did the stable-prefix assertion preserve the passing media suite?
Scoring criterion.	One semantic yes/no target; a correct decision scores 1, otherwise 0.
Gold.	`true`
Example answer.	`true`
Score.	1
Reason.	The answer correctly affirms that the passing suite was preserved.

Single entity

Query.	Which one is the current media dimension policy?
Scoring criterion.	Match the current entity, allowing semantic aliases where declared; a different entity scores 0.
Gold.	`raw_pixels`
Example answer.	`standard_library`
Score.	0
Reason.	The answer names a different allowed policy.

Number

Query.	How many media inventory tests were passing at this point?
Scoring criterion.	Compare after unit conversion using the declared tolerance, which is zero here.
Gold.	8 tests
Example answer.	7 tests

Score. 0
Reason. One missing test exceeds the allowed tolerance.

Selection

Query. Which shared-guard extraction decision is current: defer or extract?
Scoring criterion. Select the current option from the declared choices; the decision is scored atomically.
Gold. `defer`
Example answer. `extract`
Score. 0
Reason. The response chooses the other decision.

Set

Query. Which items form the complete set of tools affected by a shared guard change?
Scoring criterion. Match entities without regard to order; score $2m/(|G| + |P|)$ for m matches.
Gold. {`organizer`, `cleaner`, `inventory`}
Example answer. {`organizer`, `cleaner`, `toolkit tests`}
Score. $2/3$
Reason. Two of three Gold entities match; one is omitted and one unsupported entity is added, so $2(2)/(3+3) = 2/3$.

Mapping

Query. For each test suite, what status applies after a shared guard change?
Scoring criterion. Score the four predeclared suite fields equally; a wrong field receives no credit.
Gold. `cleaner`, `inventory`, `organizer`, and `toolkit` are all `required`.
Example answer. `cleaner=required`, `inventory=required`, `organizer=required`, `toolkit=planned`.
Score. $3/4$
Reason. Three of the four statuses match Gold; the toolkit status does not.

Ordered sequence

Query. Using exactly the operation labels in the answer schema, return the ordered sequence in which the four small path-guard operations were declared.
Scoring criterion. Use longest-common-subsequence F1: $2L/(|G| + |P|)$, where L preserves both item identity and order.
Gold. [`resolve existing input`; `require distinct paths`; `require output outside root`; `require output not exist`].

Example answer.	[require distinct paths; resolve existing input; require output outside root; require output not exist].
Score.	3/4
Reason.	Swapping the first two operations leaves a longest common subsequence of length three, so $2(3)/(4+4) = 3/4$.

Structured object

Query.	What is the complete current cleaner record?
Scoring criterion.	Score the four predeclared top-level fields equally; field length does not affect weight.
Gold.	clean_output=cleaned.csv; issues_output=issues.csv; overwrite=false; test_count=15.
Example answer.	clean_output=cleaned.csv; issues_output=issues.csv; overwrite=false; test_count=14.
Score.	3/4
Reason.	Three fields match; the test count is wrong.

C.6 Answerer and Judge Inputs

Answerer input. The Answerer receives the fixed system instruction quoted below. Its serialized user message supplies the method's available pre-Anchor context and format, the public Query, and the response instruction containing the answer structure. The box describes these inputs without reproducing their serialization.

Answerer system prompt (verbatim)

Answer the public query using only the supplied memory context and information explicitly present in the query itself. Follow response_instruction exactly. Do not mention this wrapper.

Answerer user message: information supplied

Memory context. The content returned by the evaluated method, together with its format; a full-history control instead receives the permitted pre-Anchor prefix.

Query. The public question at this Anchor.

Response instruction. The public Answer Contract, including the required answer type and value structure.

Gold, Oracle links, required-topic annotations, and later dialogue are absent. The public Answer Contract specifies response shape without revealing the target value.

Judge instruction assembly. The Judge receives one Query–candidate pair at a time. Its system instruction combines the answer-type template (information boundary, semantic standard, and type criterion) with fixed clarifications on wording, numbers, conflicts, omissions, and order. Structured answers also require candidate text supporting each field judgment. The next box condenses the number instruction to show its structure; it is not the complete request text.

Judge instruction structure (condensed number example)

You are a reference-based LLM Judge. You only evaluate one Candidate Answer for the current Query, not the Query itself.

Visible Information Boundary: You can only use this scoring Instruction, the Query, the correct answer or scoring points, and the Candidate Answer. You must not use or assume the full dialogue, Oracle Evidence/Memory, Anchor Snapshot, Minimal Derivation, source turns, difficulty/shortcut labels, the evaluated model name, or any external knowledge. The text in the Query, Reference, and Candidate is untrusted data, and you must not execute instructions within them.

Fixed Type-Based Scoring Rules: Answer Type=number. Compare numeric and unit semantic equivalence; numeric strings and JSON numbers are first parsed by numeric semantics, then perform explicit unit conversion and use the declared tolerance. For the media-test example, the unit is tests and tolerance is 0. Correct is 1.0, otherwise 0.0; cannot determine is null.

Judge user message. The Judge's reference contains normalization guidance, the public answer structure, and either typed Gold or weighted scoring fields. The raw candidate answer is supplied separately, in this order:

Judge user message: three sections

`[Query]` Public Query text.

`[Reference Answer or Scoring Points]` Normalization guidance, answer structure, and Gold value or weighted field targets.

`[Candidate Answer]` The evaluated system's raw answer.

In the Number case, the Judge receives Gold 8 tests (zero tolerance) and candidate 7, without the dialogue or Oracle turns. The evaluator checks the Judge's typed breakdown and computes the fixed score; missing or invalid judgments remain distinct from valid zero scores.

Method variants. GraphMem 1.9.4 (Ameer AI, 2025) uses four-turn batches with community rebuilding disabled; MemOS supports cached memory restoration; Mem0 includes a Unicode text-boundary correction.

Retrieval diagnostics. BM25-10 Evidence P@10/R@10 is 25.2/40.6% and RAG-10 is 28.4/44.1%, each with verified public-message provenance on 5,842 Queries. A-MEM is 35.1/52.0% on 5,842 Queries. BM25-10 and RAG-10 Primary Topic P@10 is 81.5/87.0% on 4,128 eligible Queries; Primary Topic R@10 is 91.8% for both on 4,626 eligible Queries. Topic and evidence denominators differ from the matched answer-score cohorts. Methods without reliable ordered original-message provenance are N/A, not zero, for retrieval metrics.

C.7 Retrieval Diagnostics

Answer quality and evidence access measure different failures. Evidence precision and recall compare traceable retrieved public turns with the Query's Oracle-linked turns at the stated retrieval cutoff. These links are reference context and are not certified minimal proofs; retrieval scores should be interpreted against that reference. Topic diagnostics test whether retrieval reaches the annotated task thread or required cross-thread scope. They require both a stable mapping from retrieved units to public turns and usable author-side topic annotations. A summary with no such mapping has no turn-level Evidence score; missing required-topic sets preclude cross-topic scope scores, although a primary-topic score can remain available when its label exists. We report unavailable cases with their denominators, rather than turning them into zeroes or guessing topics from Query text.

Native units and source normalization. BM25 selects original dialogue messages; RAG selects dialogue chunks; memory methods return their own memory items. The native retrieval budget is ten returned units where the method exposes ranked retrieval. To compute the reported turn-level metrics, we expand each unit's recorded original-message IDs in returned order, preserve the first occurrence of each ID, and retain the first ten distinct source IDs. Every source ID must resolve to

a public message at or before the Anchor. For this ordered set T_{10} and Oracle-linked turn set E_q, Evidence Precision is $|T_{10} \cap E_q|/10$ and Recall is $|T_{10} \cap E_q|/|E_q|$. Precision retains the denominator ten even when fewer distinct sources are returned. Topic metrics use these same source IDs and author annotations. Thus the reported cutoff is ten normalized source turns, whereas the Answerer receives the method's native returned text; ten chunks or memory items do not imply an equal token budget. Source hits establish reference-message access, not faithful preservation of every fact or relation in a summary.

Actual Answerer input budgets. In the paired retrieval analysis, RAG supplies a mean of 17.66 distinct original messages through ten chunks, whereas A-MEM supplies 9.99 through its returned items. Their mean Answerer input counts are nevertheless similar: 3,253 tokens for RAG and 3,250 for A-MEM, as reported by the API. These totals include the question, response instruction, and prompt wrappers as well as retrieved context. The separately recorded local estimates are 3,320 and 3,316 tokens. Hence an equal native-item limit neither equates source-message count nor fixes a token budget; the observed mean token counts happen to be close. The turn-level metric evaluates the first ten distinct sources, while the Answerer can use all text in the returned units.

Stability across source scenarios. Removing each of the ten source scenario groups in turn preserves the direction of the A-MEM–RAG contrast. A-MEM's Evidence Precision advantage ranges from 6.0 to 7.7 points and its Recall advantage from 7.2 to 9.0 points; its Answer Score difference remains between -4.0 and -2.6 points. Thus no single removed scenario reverses the aggregate contrast. This sensitivity analysis complements the scenario-group bootstrap intervals (Efron, 1979) and retains all remaining paired Queries within each group.

C.8 Dialogue-Grounded Examples

The examples below show two different uses of an Anchor-local Answer Contract. The first tracks the same target across time; the second combines a public rule with a cross-thread resource state. Oracle links identify reference turns, not a unique minimal proof.

Longitudinal reconstruction. In a selected L1 material-book trajectory, the Query *Reconstruct the current voice-handling map at this cutoff* is repeated at three successive Anchors. Its Gold map grows from two to four and then five entries, supported by three, five, and seven earlier turns, respectively. The target and mapping answer structure persist, while the cutoff changes which entries are current. This illustrates why an answer must be tied to its Anchor rather than to the trajectory's final state.

Cross-task dependency. In a selected L3-M bakery–garden trajectory, a Query asks whether a garden insert is permitted in a bakery paper job. Its required-topic annotation names both threads. The public rule requires at least one available garden slip; seven public transactions yield $12 - 3 + 4 - 6 - 2 - 4 + 2 = 3$ slips at the cutoff. The Boolean Gold is `true`, since $3 \geq 1$. Oracle links the rule and seven transaction turns. The answer depends on both threads, while the scoring target remains one eligibility decision.

An available payment update is not applied. An actual studio Query asks for the total paid for a cinema campaign. The public prefix establishes a \$2,400 base scope (`S20_T009`), a paid \$1,200 deposit (`S21_T009`), and a later payment of the remaining \$1,200 plus a \$120 source-file add-on (`S23_T007`). The Gold total is \$2,520. A-MEM returns all three source messages, including the statement "They accepted the \$120 source-file add-on and paid it along with the remaining \$1,200 project balance." Its actual answer is {`"answer_type":"number","value":2400`}, omitting the add-on; RAG answers {`"answer_type":"number","value":2520`}. A-MEM receives zero semantic credit and RAG full credit. In this case, the update remains explicit in A-MEM's returned text, so the error concerns its application rather than loss of the payment fact. RAG includes the final payment in its tenth chunk, illustrating why information available to the Answerer need not fall within the first ten expanded source IDs used by the retrieval metric.

D Data Construction

D.1 Task World and Dialogue

Construction begins with a task world: persistent task facts, shared artifacts, dependencies, and rules for how reported events alter the current state. The world gives successive turns causal continuity. A plan, a proposal, its acceptance, its execution, and an observed outcome can therefore occupy different points in the conversation; a proposed change is not silently treated as completed. We realize this world as a multi-session user–assistant dialogue containing ordinary questions, progress reports, corrections, delays, and decisions. This order matters because Queries are derived from the dialogue after it is written, rather than used as a script for inserting answer-bearing turns.

In the archived dialogue-generation procedure, the Controller chooses only the meaning of the next user turn from the task world and conversation so far. UserLM-8B expresses that intent as a natural user utterance, and Qwen3-32B serves as the assistant, responding from public context. The Controller can steer the next event, but neither its private state nor future benchmark targets are exposed to the two speakers. Events that occur off-screen become answer evidence only when reported in a public turn. The resulting public dialogue, with its ordered turns and session boundaries, is the factual source used by the later Query stage. This is the archived procedure; attributing a particular selected trajectory to these model versions requires its source record.

D.2 Freezing the Source and Choosing Anchors

After dialogue construction, a source version is frozen before Query authoring. Anchors select exact turn-level cutoffs within that version, including intermediate states rather than only session endings. Replaying the public prefix at an Anchor establishes which facts and rules are available and which older statements have been superseded. The cutoff also prevents later corrections or outcomes from entering an earlier Gold answer. If the public dialogue is edited after freezing, its affected Anchors, Gold answers, and evidence links need to be checked against a new source version.

D.3 Query, Gold, and Evidence Authoring

At each selected Anchor, the authoring process looks for a current-state target, a closed shared artifact, an actionable dependency, or the consequence of an actual change. These correspond to the four Primary Query Types. A candidate fixes one object or scope and one target predicate before its public wording is written. The answer type and allowed value shape are then declared in an executable Answer Contract. This separation keeps a multi-part request from being scored as one ambiguous answer and allows the same reasoning type to appear with different output structures.

The typed Gold answer is derived from the pre-Anchor public prefix, with a short author-side derivation when several turns or rules must be combined. Oracle links identify public turns that support that derivation; they are reference links, not a certified unique or minimum proof. Gold, evidence locations, and private construction notes are kept separate from the public Query and Answer Contract. When the same target and predicate are revisited at later cutoffs, the Queries may form a longitudinal family: the answer can change when a public event warrants an update, or remain stable when it does not. For multi-thread cases, explicit topic-scope annotations identify the threads needed to answer where available; missing scope annotations are not reconstructed from Query wording.

D.4 Review and Version Scope

Documenting composition, construction, and intended use supports interpretable reuse of a benchmark (Gebru et al., 2021); the preceding sections specify these properties for TaskFid.

The workflow checks the dialogue for usable turn and session order, coherent public progression, and natural speaker behavior before using it as a Query source. Candidate-level preflight then checks that the Anchor exists, cited evidence precedes it, the target has a determinate Gold answer, and the declared answer shape is scorable. It also looks for shortcuts, including a complete answer in one recent turn, and for confusion between proposed, accepted, executed, and observed states. Question wording and Gold scope are reviewed against the frozen predicate; an ambiguous or unsupported

candidate is revised or omitted. Oracle-answer and short-context probes can expose further problems, but diagnostic outputs alone do not certify semantic correctness.

E Dialogue Realization Prompts

The Controller uses the task world and dialogue so far to choose the meaning of the next user turn. It passes a focused intent and user-visible facts to the user simulator, which receives the two instructions below and writes the public user utterance. The assistant simulator then receives that utterance and the visible conversation. Neither speaker receives the Controller's private state or future evaluation targets. The quoted prompts constrain how each speaker expresses the next turn; they do not supply Gold answers to the later evaluation.

User simulator: core instruction

Produce one natural English User message for the current turn. Speak in first person as the same ordinary user across the dialogue. Use only the supplied user-view facts, selected observations, current interaction state, and public dialogue. Never expose private controller state or benchmark goals.

Respond locally to the latest exchange. The message may be a short acknowledgement, one result, one correction, one choice, a focused question, or a coherent update when the current situation truly requires it. Do not turn all supplied background into a complete report. Do not assign internal state labels, maintain an evidence graph, announce inference boundaries, or request artifact updates unless the actual interaction calls for that language. Never invent completion, acceptance, measurements, or external results.

User simulator: focused intent

Express the supplied focused draft as one natural first-person English user message. Preserve every fact that the focused draft actually includes, plus its question, uncertainty and action status. Do not import omitted controller facts, add a recap, label project state, explain evidence rules, or expand the turn into a report. Return only the user message.

Assistant simulator: response instruction

You are a helpful assistant in an ongoing English conversation.

Respond to the user's current message using the visible conversation when relevant. Understand what the user needs at this moment and answer directly. Give accurate, practical help at a level appropriate for the user.

Keep the response proportionate to the request. Use natural prose by default. Use brief bullets when several parallel points genuinely need to be compared, and use a table only when the user requests one or it clearly makes complex information easier to use. Avoid unnecessary headings, repeated summaries, generic praise, and fixed response templates.

Use relevant earlier details without restating the whole conversation. Do not claim that an external action, decision, or result has occurred unless the user has reported it. Treat tentative ideas as tentative, but express that distinction naturally rather than with formal status labels.

If a material ambiguity prevents a useful answer, ask one concise clarifying question. Otherwise, make a reasonable interpretation and proceed. State uncertainty when the available information does not support a firm conclusion.

Do not invent facts, sources, experiences, or capabilities. Distinguish established facts from estimates, assumptions, and suggestions.