ANSWER ERROR ≠ FORGETTING: DISENTANGLING MEMORY FROM TEMPORAL REASONING OVER CONTINUAL KNOWLEDGE STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory is essential for language models operating over continual knowledge streams, where information is presented sequentially as knowledge accumulates, is updated, or is superseded over time. Recent long-term memory benchmarks extend beyond factual recall to evaluate capabilities such as knowledge updating and temporal reasoning over long histories. However, these capabilities are typically assessed through end-to-end question answering, making it difficult to determine whether an incorrect answer reflects a memory-related failure or insufficient temporal reasoning. We introduce **DiMeR**, a diagnostic benchmark that presents knowledge units sequentially and evaluates the same temporal questions under controlled evidence conditions. DiMeR contains 1,956 temporal question–answer pairs constructed from 674 Wikipedia articles, each associated with source-linked gold evidence in the corresponding knowledge stream. We further divide the questions into Easy and Hard groups according to whether the temporal cues expressed in the question are explicitly present in the knowledge stream. For each question, we construct five core evaluation conditions that vary when and how the supporting evidence is available. To more directly separate memory-dependent degradation from reasoning limitations, we additionally evaluate memory conditions on questions that the same model answers correctly when the relevant evidence is provided directly. Beyond answer accuracy, we measure evidence completeness and introduce Refresh-4K, which reintroduces the original gold evidence after continued knowledge ingestion. Experiments show that Hard questions are more sensitive to continued ingestion and that similar answer accuracy can correspond to different levels of evidence completeness. Evidence reintroduction also does not always recover the correct answer, indicating that answer correctness alone is insufficient to characterize long-term memory performance.

1 INTRODUCTION

Long-term memory is essential for large language models (LLMs) operating over continual knowledge streams, where information is presented sequentially as knowledge accumulates, is updated, or is superseded over time. Recent long-term memory benchmarks have therefore moved beyond factual recall to evaluate capabilities such as multi-session reasoning, knowledge updating, and temporal reasoning over long histories (Maharana et al., 2024; Wu et al., 2025; Li et al., 2026). These tasks require models not only to retain previously observed information, but also to reason over it when answering later questions.

However, long-term memory is still commonly evaluated through end-to-end question answering. An incorrect answer may arise because the model fails to recover the required historical evidence, or because it fails to reason correctly even when that evidence is available. For temporal questions, this distinction is particularly important because the model must identify which fact is valid under the queried temporal constraint. Consequently, answer correctness alone does not reveal whether a failure should be attributed to long-term memory or temporal reasoning.

As shown in Figure 1, Llama-3.1-8B-Instruct gives the correct answer when queried after three knowledge units, whereas it produces an incorrect answer after five knowledge units, despite complete

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.

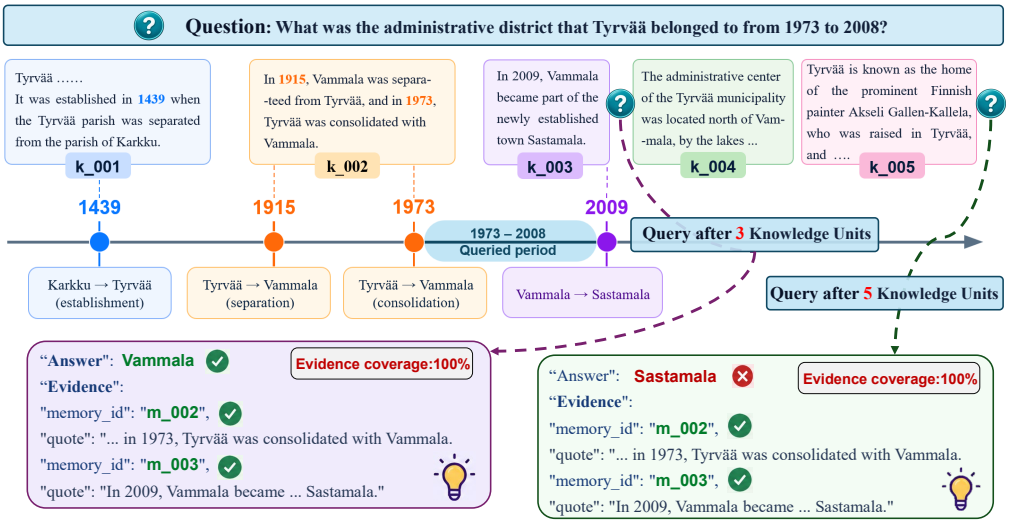

Figure 1: An example from DiMeR, in which Llama-3.1-8B-Instruct gives different answers to the same temporal question after three and five knowledge units despite complete evidence coverage.

evidence coverage in both settings. This example suggests that changes in answer correctness during continued reading are not always accompanied by corresponding changes in evidence availability. This motivates our central question: *when direct evidence is sufficient for correct temporal reasoning, how does performance change once the same evidence must be retained and recovered from a continual knowledge stream?*

To study this question, we introduce **DiMeR**, a diagnostic benchmark for temporal question answering over continual knowledge streams. DiMeR contains 1,956 temporal question–answer pairs constructed from 674 Wikipedia articles, each linked to source-grounded gold evidence. The questions are divided into Easy and Hard groups according to whether their temporal cues are explicitly present in the corresponding knowledge stream. For each question, we construct five controlled conditions—NO-EVIDENCE, Direct-Evidence, Immediate-Memory, Delay-2K, and Delay-4K—while keeping the question, reference answer, and gold evidence fixed. These conditions compare direct-evidence reasoning with memory-dependent answering before and after continued knowledge ingestion.

Crucially, we additionally evaluate memory-dependent performance on the subset of questions that the same model answers correctly under Direct-Evidence. This conditional evaluation reduces the confounding effect of questions that remain unsolved even when the required evidence is explicitly available. Beyond answer accuracy, we measure evidence completeness after sequential reading and introduce Refresh-4K, which reintroduces the original gold evidence after Delay-4K. Together, these diagnostics allow us to examine whether degradation is associated with incomplete evidence recovery and whether restoring the evidence is sufficient to recover the correct answer.

Our experiments reveal three main findings. First, continued knowledge ingestion affects Hard questions more strongly than Easy questions, including when evaluation is restricted to questions solved under Direct-Evidence. Second, similar answer accuracy can correspond to different levels of evidence completeness, showing that answer correctness and evidence recovery capture different aspects of model behavior. Third, reintroducing the gold evidence does not always recover the correct answer, indicating that residual errors cannot be explained by evidence incompleteness alone.

Our contributions are threefold:

- We introduce **DiMeR**, a diagnostic benchmark containing 1,956 temporal question–answer pairs from 674 Wikipedia articles, with source-linked gold evidence for evaluating temporal question answering over continual knowledge streams.
- We design five controlled evidence conditions and a Direct Evidence-conditioned evaluation protocol that compares memory-dependent performance only on questions that the same model can solve when the required evidence is explicitly available.

- We jointly analyze answer correctness and evidence completeness, and introduce evidence reintroduction through Refresh-4K to examine the relationship between evidence recovery and answer recovery after continued knowledge ingestion.

2 Related Work

2.1 Long-Term Memory for LLM Agents

Recent long-term memory benchmarks extend evaluation beyond factual recall to reasoning over extended histories. LoCoMo evaluates question answering over long conversations, including single-hop, multi-hop, temporal, and commonsense reasoning (Maharana et al., 2024), while LongMemEval further considers capabilities such as temporal reasoning and knowledge updating (Wu et al., 2025). More recent benchmarks examine implicit conversational constraints and experience accumulated from agent trajectories (Li et al., 2026; Wu et al., 2026).

Existing methods improve long-term information access through long-context modeling (Kitaev et al., 2020; Fu et al., 2024; An et al., 2024), dedicated memory modules (Zhong et al., 2022; Wang et al., 2023), and context compression or external memory (Mu et al., 2023; Chevalier et al., 2023; Jiang et al., 2023; Shi et al., 2024; Xu et al., 2025; Gutiérrez et al., 2025). However, long-term memory is still commonly evaluated through end-to-end answer correctness, making it difficult to determine whether an incorrect answer results from incomplete recovery of the required evidence or unsuccessful reasoning over recovered evidence. DiMeR targets this distinction by jointly examining answer correctness and evidence completeness under controlled conditions.

2.2 Continual Knowledge Streams

Continual knowledge streams present information incrementally, requiring models to retain and use knowledge observed earlier as new information arrives (Kim et al., 2026). HippoRAG 2 studies non-parametric accumulation and retrieval of continually observed information (Gutiérrez et al., 2025), while long-context studies show that effective information use can depend on context length and evidence position (Bai et al., 2024; Hsieh et al., 2024).

Unlike these settings, DiMeR preserves the original order of the knowledge stream and evaluates the same temporal question at controlled points during continued reading, enabling matched comparisons as additional knowledge is ingested after the supporting evidence.

2.3 Temporal Reasoning

Temporal reasoning requires models to understand temporal constraints and relations among events (Chen et al., 2021). Existing benchmarks evaluate increasingly complex forms of temporal reasoning, including temporal comparison, aggregation, multi-hop reasoning (Gruber et al., 2025), reasoning over unseen event scenarios (Uddin et al., 2025), and exhaustive multi-answer reasoning over long contexts (Wu, 2026).

Temporal reasoning has also been incorporated into long-term memory systems. TSM organizes historical information along semantic timelines (Su et al., 2026), while Chronos structures conversational histories as temporal events for time-sensitive retrieval (Sen et al., 2026). DiMeR instead focuses on diagnosing temporal question answering over continual knowledge streams, distinguishing failures associated with historical evidence access from those that remain when the required evidence is directly available.

DiMeR differs from prior evidence-controlled memory evaluations in three respects. First, it evaluates memory over sequentially ingested knowledge streams rather than static evidence contexts. Second, it introduces controlled post-evidence ingestion checkpoints, allowing the same question to be evaluated immediately after its supporting evidence and again after additional knowledge has been consumed. Third, Refresh-4K explicitly reintroduces the original evidence after delayed ingestion, enabling a matched diagnostic comparison between evidence recovery and answer recovery. These controls are further analyzed under temporal-cue locatability, separating questions whose queried timestamps are explicitly recoverable from those requiring less direct temporal localization.

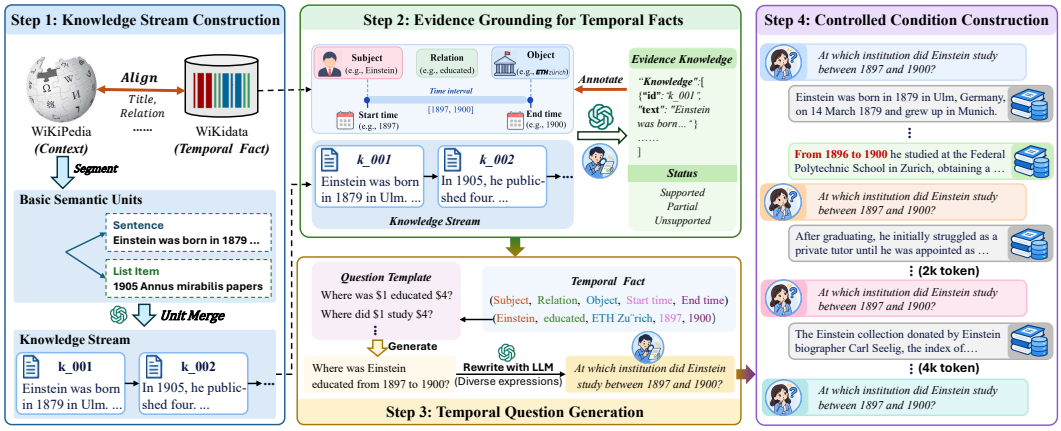

Figure 2: The overall framework for DiMeR. It consists of four steps: knowledge stream construction, evidence grounding, temporal question generation, and controlled condition construction.

3 METHODS

This section details the problem formulation, construction pipeline and statistical analysis of DiMeR. As illustrated in Figure 2, DiMeR is constructed by four phases: **Knowledge Stream Construction**, **Evidence Grounding**, **Temporal Question Generation**, and **Controlled Condition Construction**.

3.1 PROBLEM FORMULATION

In DiMeR, each evaluation instance is represented as a 4-tuple $(\mathcal{K}, q, a, E)$. Here, $\mathcal{K} = (K_1, K_2, \ldots, K_T)$ is a continual knowledge stream consisting of T knowledge units presented sequentially in their original source order, where K_t denotes the knowledge unit at step t. q is a temporal question associated with evidence contained in $\mathcal{K}$, and a denotes its reference answer. $E = \{K_{e_1}, \ldots, K_{e_m}\}$ is the annotated gold evidence set supporting the question-answer pair (q, a), where $1 \leq e_1 < \cdots < e_m \leq T$. Based on this formulation, we next construct the knowledge stream, annotate the corresponding gold evidence set, generate the temporal question and its reference answer.

3.2 KNOWLEDGE STREAM CONSTRUCTION

Following MulTR (Wu, 2026), we align structured temporal facts with their corresponding Wikipedia articles. Each temporal fact is represented as

$$\tau = (\text{subject}, \text{relation}, \text{object}, t_{\text{start}}, t_{\text{end}}), \tag{1}$$

where the temporal boundaries retain their original precision. The aligned articles provide the textual source for constructing continual knowledge streams, while the temporal facts are used for subsequent evidence grounding and question generation.

We first segment each article into minimal semantic units, including sentences and individual list items, while preserving their original source order. The units are then merged into knowledge units using Qwen3-30B-A3B (Yang et al., 2025). Consecutive list items are grouped into the same knowledge unit, whereas adjacent sentences are merged when they are semantically related and jointly describe a complete event. The merging process only changes unit boundaries and does not rewrite the original text. We retain articles containing at least five knowledge units after merging, producing an ordered knowledge stream $\mathcal{K} = (K_1, \ldots, K_T)$. Additional details on source filtering, knowledge-unit construction, and evidence annotation are provided in Appendix A.

3.3 EVIDENCE GROUNDING

For each temporal fact τ, we first identify the supporting knowledge units from the complete knowledge stream by Qwen3-30B-A3B, producing an evidence set $E = \{K_{e_1}, \ldots, K_{e_m}\}$. Then, we

Knowledge Units	**[K_1]** Tyrvää … was a municipality in the Satakunta region, Turku and Pori Province, Finland. It was established in 1439 when the Tyrvää parish was separated from the parish of Karkku. **[K_2]** In 1915, the market town of Vammala was separated from Tyrvää, and in **1973**, Tyrvää was consolidated with **Vammala**. **[K_3]** In **2009**, Vammala became a part of the newly established town **Sastamala**. …
Temporal Question	Which administrative district did Tyrvää belong to from 1973 to 2008?
Reference Answer	Vammala
Gold Evidence	$\{K_2, K_3\}$
Condition	**Pre-Question Input**
No-Evidence	$\varnothing$
Direct-Evidence	(K_2, K_3)
Immediate-Memory	(K_1, K_2, K_3)
Delay-2K	$(K_1, \ldots, K_{69})$; **2,060** additional text tokens relative to Immediate-Memory
Delay-4K	$(K_1, \ldots, K_{126})$; **4,102** additional text tokens relative to Immediate-Memory

Table 1: An example from DiMeR illustrating the five evaluation conditions. K_j denotes the jth knowledge unit. All five conditions share the same question, reference answer, and gold evidence.

use GPT-5 to verify the temporal fact together with the annotated evidence and assign a binary label indicating whether the evidence sufficiently supports the fact. Annotations rejected by the verifier are re-annotated and subsequently inspected manually to correct missing, redundant, or incorrectly matched evidence units.

3.4 Temporal Question Generation

For each verified temporal fact, we generate a base question using a relation-specific template, where the subject and temporal constraint form the query and the object serves as the reference answer. We use 23 templates covering 8 temporal relation types. The generated questions are then paraphrased using GPT-5 while preserving the subject, queried relation, temporal constraint, reference answer, and evidence mapping. More details are provided in AppendixB.

We further divide the questions into Easy and Hard groups according to whether the temporal cues mentioned in the question are explicitly present in the corresponding knowledge stream. For questions with a time point, the question is categorized as Easy if its mentioned time point appears in the knowledge stream; otherwise, it is categorized as Hard. For questions with a time interval, both the start and end times must appear in the knowledge stream for the question to be categorized as Easy; otherwise, it is categorized as Hard.

3.5 Controlled Condition Construction

For each instance $(\mathcal{K}, q, a, E)$, we construct five core evaluation conditions that keep the question, reference answer, and gold evidence fixed while varying when and how the supporting evidence is available (Table 1).

No-Evidence provides a baseline without evidence from the current stream. **Direct-Evidence** supplies only the gold evidence and establishes whether the model can answer the temporal question when the required information is explicitly available. **Immediate-Memory** instead presents the evidence naturally within the ordered knowledge stream and queries the model immediately after the final gold evidence unit.

To examine the effect of continued reading, Delay-2K and Delay-4K extend the stream by at least 2,048 and 4,096 text tokens, respectively, after the final evidence unit. The original knowledge-unit order is preserved, and extensions terminate at complete unit boundaries. When the source article is insufficient, additional articles are appended in source order. These continuations preserve the sequential-reading setting but are not independently certified to be irrelevant to the target question.

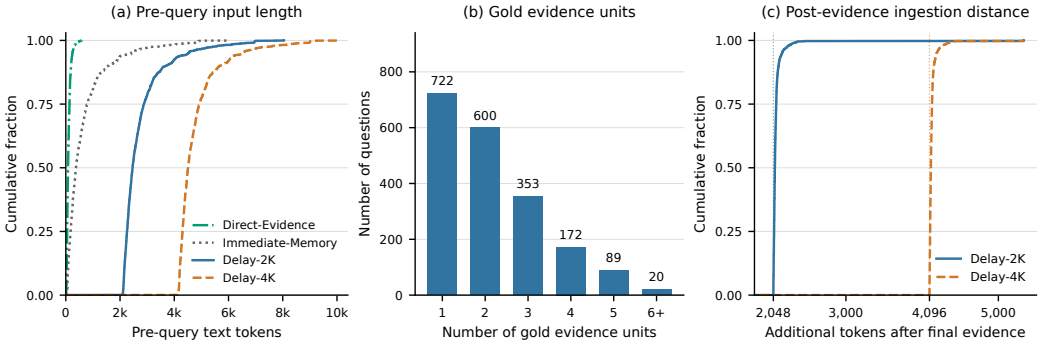

Figure 3: Statistics of DiMeR. (a) Pre-query input length under Direct-Evidence, Immediate-Memory, Delay-2K, and Delay-4K. (b) The number of gold evidence per question. (c) Additional text tokens ingested after the final gold evidence under Delay-2K and Delay-4K.

Together, these conditions enable matched comparisons across direct-evidence, immediate-memory, and delayed-memory settings.

3.6 Dataset Statistics

DiMeR contains 1,956 temporal questions from 674 Wikipedia articles, including 1,219 easy and 737 hard questions. The questions cover 8 temporal relation types and are generated from 23 relation-specific templates. Figure 3 summarizes the number of gold evidence units, pre-query context lengths, and post-evidence ingestion distances.

4 Experiments

4.1 Experimental Setup

Baselines We evaluate a diverse set of baselines to provide a comprehensive assessment on DiMeR. These baselines are categorized into two groups according to whether they explicitly construct external memory: *base models* and *memory-augmented models*.

We evaluate four representative language models: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), GLM-4-9B-Chat (GLM Team et al., 2024), and GPT-5.4 mini (OpenAI, 2026). These models directly process the knowledge provided under each evaluation condition without constructing an explicit external memory.

We also evaluate three long-term memory systems: A-MEM (Xu et al., 2025), Mem0 (Chhikara et al., 2025), and HippoRAG 2 (Gutiérrez et al., 2025). These systems sequentially ingest the knowledge stream, construct persistent memory, and retrieve relevant information for question answering. All three memory systems use Llama-3.1-8B-Instruct as the backbone model for answer generation.

Evaluation Metrics. We evaluate model performance from two complementary perspectives: *answer correctness* and *evidence completeness*. For answer correctness, we report **Answer Accuracy**, which measures whether the normalized model prediction matches the reference answer. For evidence completeness, we use three metrics: **Evidence Hit (EH)**, which measures whether at least one gold evidence unit is identified; **Evidence Recall (ER)**, which measures the proportion of gold evidence units identified; and **All-Evidence Recall (AR)**, which measures whether the complete gold evidence set is covered. Together, EH, ER, and AR characterize evidence completeness at progressively stricter levels, from identifying any gold evidence to covering the complete gold evidence set. Detailed definitions are provided in Appendix C.

Table 2: Answer accuracy (%) across five conditions, with Easy and Hard modes.

Method	No-Evidence		Direct-Evidence		Immediate-Memory		Delay-2K		Delay-4K	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Llama-3.1-8B	17.31	9.09	57.59	47.76	55.54	46.68	54.72	35.82	53.57	35.41
Qwen2.5-7B	7.63	6.38	56.52	37.18	56.44	34.60	51.76	25.64	49.71	25.51
GLM-4-9B	10.34	5.43	**63.90**	**51.02**	**62.10**	**49.53**	**59.23**	**39.35**	**59.15**	**39.21**
GPT-5.4-mini	**21.97**	**11.89**	62.43	44.76	61.27	44.06	58.38	39.16	56.07	38.46
A-MEM	–	–	56.36	47.08	51.44	44.50	46.24	36.36	45.66	36.36
Mem0	–	–	37.65	21.85	47.91	35.96	49.13	25.17	43.93	23.08
HippoRAG 2	–	–	57.18	48.03	52.09	44.50	53.18	34.97	52.60	34.27

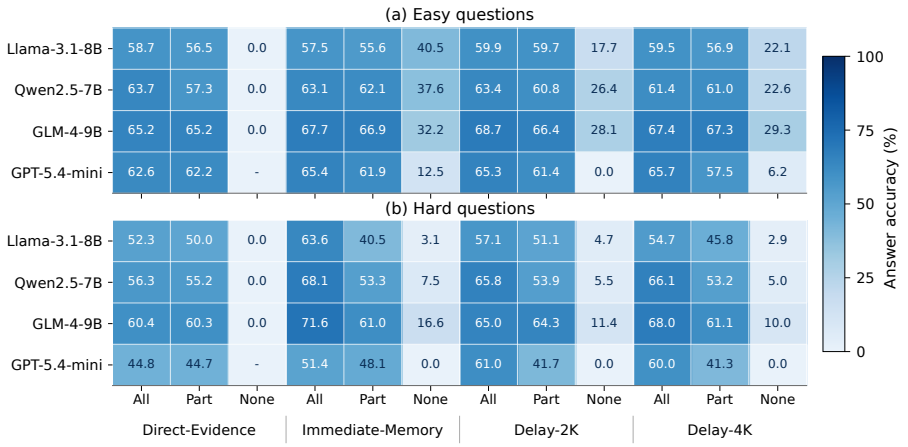

Figure 4: Answer accuracy conditioned on evidence coverage across four evaluation conditions.

4.2 Overall Performance

The main experimental results are summarized in Table 2. Our analysis reveals how answer performance changes as models move from direct evidence access to memory-dependent reasoning over continued knowledge streams.

Direct evidence substantially improves temporal answering. Providing the gold evidence directly yields large gains over No-Evidence for all base models. For example, Llama-3.1-8B improves from 17.31% to 57.59% on Easy questions, while GLM-4-9B improves from 10.34% to 63.90%; similar gains appear on Hard questions. No-Evidence accuracy indicates that some questions can also be answered from parametric knowledge or other model priors. Therefore, We interpret the controlled-condition comparisons as measuring the effect of evidence availability and continued ingestion, rather than assuming that all correct answers arise from the provided knowledge stream.

Explicit memory augmentation does not consistently improve answer accuracy. The three memory-augmented systems do not consistently outperform their shared Llama-3.1-8B backbone. Under Delay-4K, for example, HippoRAG 2 achieves 52.60%/34.27% on Easy/Hard questions, close to Llama-3.1-8B at 53.57%/35.41%, while A-MEM and Mem0 show mixed or lower performance.

Continued ingestion affects Hard questions more strongly. For Llama-3.1-8B, accuracy decreases from 46.68% to 35.41% on Hard questions from Immediate-Memory to Delay-4K, compared with 55.54% to 53.57% on Easy questions. The additional change from Delay-2K to Delay-4K is relatively small, suggesting that degradation does not increase proportionally with additional post-evidence context. Because aggregate accuracy still mixes questions that models can and cannot solve under Direct-Evidence, we report Direct-Evidence-conditioned results in Appendix E.1 and next examine evidence completeness.

Table 3: Evidence recall metrics (%) across four conditions, with Easy and Hard questions.

Method	Direct-Evidence			Immediate-Memory			Delay-2K			Delay-4K		
	EH	ER	AR	EH	ER	AR	EH	ER	AR	EH	ER	AR
DiMeR-Easy												
Llama-3.1-8B	99.18	86.54	69.73	93.11	73.44	51.68	87.94	66.32	43.31	87.04	66.06	44.38
Qwen2.5-7B	94.50	61.81	36.42	75.80	47.75	27.15	71.78	45.01	25.35	70.30	43.09	23.79
GLM-4-9B	98.03	65.29	38.23	85.48	55.42	31.99	79.57	50.18	27.81	78.42	50.34	28.47
GPT-5.4-mini	**100.00**	77.23	52.60	95.38	70.01	46.82	**92.49**	65.64	41.62	90.75	64.78	40.46
DiMeR-Hard												
Llama-3.1-8B	92.13	83.96	73.95	82.36	**68.91**	55.22	62.28	49.67	37.31	67.44	53.56	40.16
Qwen2.5-7B	66.76	46.86	31.34	51.15	35.62	24.69	37.86	24.62	15.47	37.99	25.64	16.82
GLM-4-9B	84.53	61.78	44.91	66.49	46.46	32.02	52.51	35.87	24.02	53.87	36.77	24.56
GPT-5.4-mini	**100.00**	83.23	67.13	**88.11**	68.74	51.75	**74.83**	**57.02**	**41.26**	**74.13**	**56.90**	**41.96**

4.3 Evidence Recall

The evidence completeness results are summarized in Table 3, while Figure 4 reports answer accuracy conditioned on different levels of evidence coverage.

The Effect of Continued Ingestion on Evidence Completeness. Across all base models and conditions reported in Table 3, EH is consistently higher than ER and AR, indicating that recovering at least one gold evidence unit does not imply complete evidence coverage. Evidence completeness decreases from Immediate-Memory to the delayed conditions, with larger reductions on Hard questions. For Llama-3.1-8B, AR on Hard questions decreases from 55.22% under Immediate-Memory to 37.31% under Delay-2K, compared with 51.68% to 43.31% on Easy questions. From Delay-2K to Delay-4K, the corresponding changes are relatively small, reaching 40.16% and 44.38%, respectively. Similar trends are observed across the other models. The effect of continued ingestion is therefore more pronounced on Hard questions, while further increasing the post-evidence context does not produce a proportional decline.

The Relation between Evidence Completeness and Answer Correctness. As shown in Figure 4, Answer accuracy is substantially lower when no gold evidence is identified, whereas the difference between complete and partial coverage is often modest. Complete evidence coverage nevertheless does not guarantee a correct temporal answer. Under Delay-4K, Llama-3.1-8B achieves 54.7% answer accuracy on Hard questions even when all gold evidence is identified, while GLM-4-9B shows nearly identical accuracy under complete and partial evidence coverage for easy questions. These results indicate that evidence completeness and answer correctness are related but distinct aspects of model behavior. Overall, continued knowledge ingestion reduces evidence completeness, particularly for Hard questions, while complete evidence coverage remains insufficient to ensure correct temporal reasoning. Additional analyses of evidence cardinality and temporal-relation effects are provided in Appendices E.3 and E.4, respectively.

5 Analysis

5.1 Answer Recovery under Evidence Reintroduction

To examine whether the performance degradation after continued knowledge ingestion can be recovered by restoring the gold evidence, we introduce Refresh-4K as a diagnostic intervention. Starting from the same pre-query stream as Delay-4K, Refresh-4K appends an exact copy of the gold evidence immediately before the question. Thus, the two conditions differ only in whether the gold evidence is reintroduced before answering.

Evidence reintroduction provides only partial answer recovery. As shown in Figure 5(a)–(b), Refresh-4K improves answer accuracy over Delay-4K for all evaluated models, but the improvements are generally limited. On Hard questions, GLM-4-9B exhibits the largest increase, from 39.21% to 43.42%, whereas Llama-3.1-8B changes only from 35.41% to 35.55%. Moreover, Refresh-4K remains below Immediate-Memory across all models, with the gap being particularly pronounced on

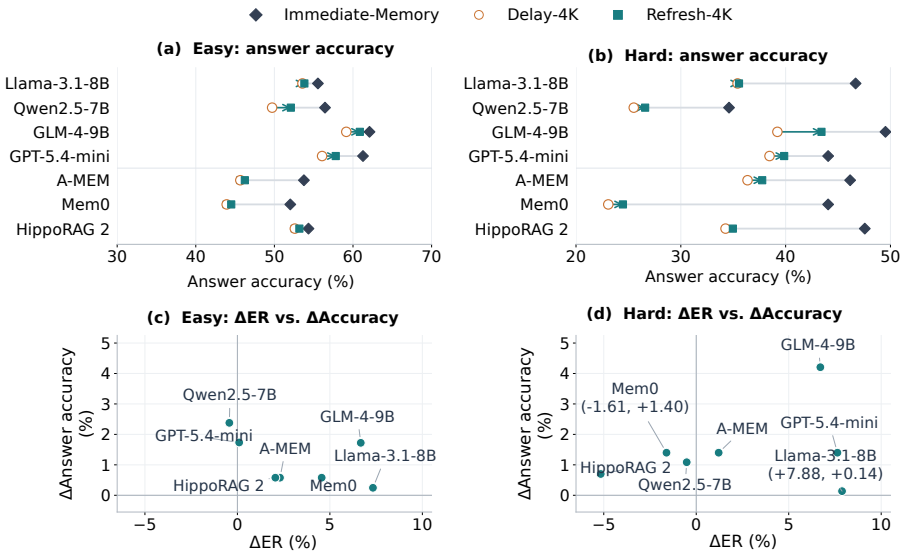

Figure 5: Answer and evidence recovery under Refresh-4K. Panels (a) and (b) compare answer accuracy under Immediate-Memory, Delay-4K, and Refresh-4K. Panels (c) and (d) show the relation between changes in evidence recall and answer accuracy from Delay-4K to Refresh-4K.

Hard questions. These results indicate that reintroducing the gold evidence recovers only part of the answer performance lost after continued knowledge ingestion.

5.2 Relation between Evidence Recovery and Answer Recovery

Evidence recovery and answer recovery are not tightly coupled. Figure 5(c)-(d) compares the change in Evidence Recall with the corresponding change in answer accuracy from Delay-4K to Refresh-4K. The two quantities exhibit substantial variation across models. For example, on Hard questions, the ER of Llama-3.1-8B increases by 7.88 percentage points, while answer accuracy increases by only 0.14 points. In contrast, GLM-4-9B shows increases of 6.71 points in ER and 4.21 points in answer accuracy. Some models even exhibit improved answer accuracy despite reduced ER, such as Mem0 and HippoRAG 2 on Hard questions. These results show that changes in evidence recovery do not translate proportionally into changes in answer accuracy.

Evidence recovery alone does not explain residual answer errors. The limited correspondence between evidence recovery and answer recovery suggests that incomplete evidence recovery is not sufficient to explain all errors after continued knowledge ingestion. In particular, substantial improvements in ER can coexist with negligible improvements in answer accuracy. Therefore, answer errors that persist after evidence reintroduction cannot be attributed to evidence incompleteness alone, motivating a distinction between recovering the supporting evidence and successfully using it for temporal reasoning. Paired bootstrap confidence intervals and exact McNemar tests for the three open base models are reported in Appendix E.2.

6 Conclusion

We introduce DiMeR, a diagnostic benchmark for temporal question answering over continual knowledge streams under controlled evidence conditions. By jointly evaluating answer correctness and evidence completeness, DiMeR helps distinguish incomplete evidence recovery from errors that persist despite available evidence. Our results show that these signals capture different aspects of model behavior, enabling a more precise analysis of long-term memory and temporal reasoning. One limitation is that GPT-5 is used during evidence verification and question paraphrasing, while GPT-5.4-mini is evaluated as a baseline, which may introduce mild model-family bias. Future work will validate the annotations using independent model families and broader human review.

AI USE STATEMENT

Large language models were used as supporting tools during dataset construction and manuscript preparation. Specifically, during **knowledge-unit construction**, we used Qwen3-30B-A3B to merge sentences into knowledge units while preserving the original text and source order. During **evidence grounding**, Qwen3-30B-A3B was used to identify supporting knowledge units for each temporal fact, and GPT-5 was subsequently used to verify whether the annotated evidence sufficiently supported the corresponding subject, relation, object, and temporal constraints. Annotations rejected during verification were re-annotated and manually inspected to correct missing, redundant, or incorrectly matched evidence units. During **question generation**, GPT-5 was used to paraphrase relation-template-generated questions while preserving the subject, queried relation, temporal constraint, reference answer, and evidence mapping. Manual quality checks were additionally performed by random sampling during dataset construction. Generative AI tools were also used for language polishing during manuscript preparation.

REFERENCES

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your LLM fully utilize the context. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems*, 2024.

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, 2024.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*, 2021.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, 2023.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *Proceedings of the 28th European Conference on Artificial Intelligence*, volume 413, pp. 2993–3000, 2025.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. In *Proceedings of the Forty-first International Conference on Machine Learning*, volume 235, pp. 14125–14134, 2024.

GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv:2406.12793*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. Complextempqa: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9100–9112, 2025.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the Forty-second International Conference on Machine Learning*, volume 267, 2025.

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

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023.

Jiyeon Kim, Hyunji Lee, Dylan Zhou, Sue Hyun Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Sungmin Cha, and Minjoon Seo. Can large language models keep up? benchmarking online adaptation to continual knowledge streams. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42240–42272, 2026.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 2026.

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, 2024.

Jesse Mu, Xiang Li, and Noah D. Goodman. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, 2023.

OpenAI. GPT-5.4 Mini Model. OpenAI API Documentation, 2026.

Sahil Sen, Elias Lumer, Anmol Gulati, and Vamse Kumar Subbiah. Chronos: Temporal-aware conversational agents with structured event retrieval for long-term memory. *CoRR*, abs/2603.16862, 2026.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: retrieval-augmented black-box language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, 2024.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. In *Findings of the Association for Computational Linguistics*, pp. 29935–29951, 2026.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven Corman, and Chitta Baral. UnSeenTimeQA: Time-sensitive question-answering beyond LLMs' memorization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 1873–1913, 2025.

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, 2023.

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.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv:2605.12493*, 2026.

Shaojuan Wu. From existence to exhaustiveness: Unveiling the compounding failures of llms in multi-answer event temporal reasoning. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2015–2026, 2026.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv:2505.09388*, 2025.

Zexuan Zhong, Tao Lei, and Danqi Chen. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5657–5673, 2022.

A Dataset Construction Details

This appendix supplements the construction procedure in Section 3. It describes the source collection, knowledge-unit construction, and evidence quality control.

A.1 Source Data and Filtering

Each source record associates a Wikipedia article with structured temporal facts of the form (s, r, o, t_s, t_e). The article supplies the knowledge stream, while the facts specify the subject, relation, reference object, and temporal boundaries used for evidence annotation and question generation. Article and fact identifiers preserve these correspondences throughout construction.

The source collection contains 882 articles. Knowledge-unit construction succeeds for 751 articles, and 696 remain after the post-merge five-unit filter. Evidence grounding, binary verification, re-annotation, and human review produce validated annotations for 1,956 temporal facts from a corpus of 674 articles. One question is generated for each validated fact, yielding 1,219 Easy and 737 Hard questions in the current release.

A.2 Knowledge Unit Construction

Articles are first segmented into minimal semantic units using rule-based processing. Sentences and individual list items retain their original types and source order. A locally hosted Qwen3-30B model then merges these units into knowledge units. Consecutive list items are grouped together, while adjacent narrative sentences are merged when they are semantically related and jointly describe a complete event. Unrelated events remain separate.

Merging changes unit boundaries without rewriting the source. Original expressions, including pronouns, are preserved, and the resulting units retain the article's order. The article-size criterion is applied after merging: an article must contain at least five knowledge units, and all of its valid units are retained. It is neither a five-sentence threshold nor a restriction to the first five units.

Each frozen knowledge unit contains `memory_id`, `unit_type`, and `text`. The legacy field `memory_id` identifies a source knowledge unit rather than a model's internal memory. Article–fact–unit correspondences are maintained separately. Evidence annotation and question generation operate on these frozen units without changing their text or order.

A.3 Evidence Annotation and Quality Control

Qwen3-30B receives the complete knowledge stream and a temporal quintuple, then identifies its supporting knowledge units. This stage precedes natural-language question generation. Annotation checks the subject, relation, object, start time, and end time against the source, preserving the original precision of the temporal boundaries. Evidence may include multiple units when identity, relation, and temporal information are distributed across the article. The annotation record contains the fact identifier, evidence-unit IDs, a support status, and unsupported fields. An illustrative record is:

```
{
  "triple_id": "t_001",
  "evidence_memory_ids": ["m_003", "m_005"],
  "evidence_status": "supported",
  "unsupported_fields": []
}
```

Table 4: Relation inventory and question distribution in DiMeR.

Relation	Meaning and object role	n_T	Easy	Hard	Total
P39	**Position held**: o is the position held by person s.	4	614	403	1,017
P54	**Member of sports team**: o is the team for which player s played.	3	257	132	389
P937	**Work location**: o is the place where s worked.	3	97	89	186
P108	**Employer**: o is the employer for whom s worked.	4	126	59	185
P488	**Chairperson**: o is the person serving as chair or head of s.	2	73	5	78
P131	**Administrative territorial entity**: o is the administrative entity containing s.	2	27	42	69
P1037	**Director / manager**: o is the person directing or managing s.	3	19	7	26
P4791	**Commanded by**: o is the person commanding s.	2	6	–	6
Total		23	1,219	737	1,956

The annotation statuses `supported`, `partial`, and `unsupported` describe support for all, some, or none of the five fact fields, respectively. Referenced IDs must resolve to units in the corresponding frozen stream. Operational failures, such as invalid output formatting, are distinct from a semantic judgment of insufficient support.

GPT-5 receives the full knowledge stream $\mathcal{K}$, temporal quintuple τ, and annotated evidence set E. It returns label 1 when the annotation correctly supports the fact and label 0 when revision is required. For samples labeled 0, the evidence set is first re-annotated using the stream and temporal fact. The revised annotation then undergoes human review to correct missing, redundant, or incorrectly matched evidence units. Frozen stream content and order are preserved throughout. The binary labels are separate from the initial annotation statuses. This is the quality-control sequence described in Section 3.3.

Verification evaluates the temporal fact at its recorded precision. Question generation subsequently expresses its temporal constraint at year level, as specified in Appendix B. Neither the number of annotated evidence units nor the binary verification label defines the minimum number of units needed to answer a question.

B Question Generation

B.1 Relation Inventory and Templates

For each verified temporal fact $\tau = (s, r, o, t_s, t_e)$, a relation-specific template combines the subject s with the temporal constraint; the object o is the reference answer. The dataset uses eight relation types and 23 templates. Table 4 defines each relation and reports its distribution. For example, P108 concerns an employer, P937 concerns a work location, and P39 concerns a position held; these answer categories are preserved during question rewriting.

The placeholders `$1` and `$4` denote the subject and temporal expression, respectively. Let $y_s = \mathrm{year}(t_s)$ and $y_e = \mathrm{year}(t_e)$. The temporal expression is

$$\phi(t_s, t_e) = \begin{cases} \text{in } y_s, & y_s = y_e, \\ \text{from } y_s \text{ to } y_e, & y_s \neq y_e. \end{cases}$$

Month and day are omitted from the question, while the original timestamps remain associated with the source fact. The final dataset contains 139 same-year questions and 1,817 questions with distinct endpoint years.

Templates are rotated within each relation: the jth occurrence of relation r, indexed from zero, uses template $j \bmod n_T(r)$. Rotation is performed within generation batches; stored template indices are retained during consolidation rather than recomputed over the final release. Each generated question preserves its article, fact, answer, and evidence mappings.

Table 5: Templates of Temporal Questions for the Eight Retained Relations

Relation	Index	Base template
P39	0	Which position did `$1` hold `$4`?
	1	What was the position of `$1` `$4`?
	2	What position did `$1` take `$4`?
	3	`$1` took which position `$4`?
P54	0	Which team did `$1` play for `$4`?
	1	`$1` played for which team `$4`?
	2	Which team did the player `$1` belong to `$4`?
P937	0	Where did `$1` work `$4`?
	1	What was the working location for `$1` `$4`?
	2	`$1` worked in which location `$4`?
P108	0	Which employer did `$1` work for `$4`?
	1	Who did `$1` work for `$4`?
	2	What was the name of the employer `$1` work for `$4`?
	3	`$1` was an employee for whom `$4`?
P488	0	Who was the chair of `$1` `$4`?
	1	Who was the head of `$1` `$4`?
P131	0	Which adminstrative district was `$1` located `$4`?
	1	Which adminstrative district did `$1` belong to `$4`?
P1037	0	Who was the director or manager of `$1` `$4`?
	1	Who directed or managed `$1` `$4`?
	2	`$1` was managed or directed by whom `$4`?
P4791	0	Who commanded `$1` `$4`?
	1	Who was the commander of `$1` `$4`?

B.2 Question Paraphrasing

We use GPT-5 to paraphrase template-generated questions to vary their wording while preserving the subject, queried relation, temporal constraint, reference answer, and evidence mapping. Rewriting must not introduce additional temporal restrictions, change the requested answer category, or replace an interval with an unspecified time within it. The original and rewritten questions remain linked by question ID.

Consistency checks compare subject strings and numeric tokens and preserve the non-question fields. Such checks can identify structural changes but cannot replace semantic review. Therefore, we require the model to rewrite and correct wording without altering semantic information.

C Evaluation Metrics

We evaluate model performance from two complementary perspectives: answer correctness and evidence completeness. Answer correctness measures whether the model generated the target answer, while evidence completeness evaluates how completely the supporting evidence is recovered when answering the question. Accordingly, we employ four metrics: Answer Accuracy for answer correctness, and Evidence Hit (EH), Evidence Recall (ER), and All-Evidence Recall (AR) for evidence completeness.

Answer Accuracy (AnswerAcc). AnswerAcc measures whether the predicted answer matches the reference answer after normalization. Let $\hat{a}_i$ and a_i^* denote the predicted and reference answers for question q_i, respectively. The metric is defined as

$$\text{AnswerAcc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{a}_i = a_i^*]\,,$$

where $\mathbb{I}[\cdot]$ is the indicator function and N is the number of evaluated questions.

To evaluate evidence completeness, let G_i denote the set of gold evidence knowledge units associated with question q_i, and R_i denote the set of knowledge units identified by the model in its final evidence citations. We evaluate evidence completeness at three levels: whether any gold evidence is identified, how much of the gold evidence is identified, and whether the complete gold evidence set is covered.

Evidence Hit (EH). EH measures whether the model identifies at least one gold evidence unit:

$$\mathrm{EH}_i = \mathbb{I}[G_i \cap R_i \neq \varnothing] .$$

EH captures the minimum level of evidence coverage. A question is counted as a hit as long as at least one gold evidence unit is identified, regardless of whether the remaining required evidence is recovered.

Evidence Recall (ER). ER measures the proportion of gold evidence units identified by the model:

$$\mathrm{ER}_i = \frac{|G_i \cap R_i|}{|G_i|}.$$

Compared with the binary EH metric, ER provides a finer-grained measure of evidence completeness by distinguishing different degrees of partial coverage.

All-Evidence Recall (AR). AR is the strictest of the three evidence-completeness metrics. It measures whether all gold evidence units associated with a question are identified:

$$\mathrm{AR}_i = \mathbb{I}[G_i \subseteq R_i] = \mathbb{I}[|G_i \cap R_i| = |G_i|] .$$

A question receives an AR score of 1 only when the complete gold evidence set is covered. Unlike strict set exact match, AR does not penalize additional non-gold evidence as long as all gold evidence units are identified. This design is intentional because our objective is to measure the completeness of the evidence required for answering, rather than whether the predicted evidence set exactly matches the reference set.

Importantly, the maximum gold-evidence set size is 8 for Easy questions and 7 for Hard questions, which does not exceed the maximum of eight evidence citations allowed by the output schema. Therefore, the citation limit does not structurally constrain All-Evidence Recall. We report EH, ER, and AR as macro-averages over all evaluated questions with non-empty gold evidence sets:

$$\mathrm{EH} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{EH}_i, \qquad \mathrm{ER} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{ER}_i, \qquad \mathrm{AR} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AR}_i.$$

The three metrics provide progressively stricter views of evidence completeness:

$$\mathrm{AR} \leq \mathrm{ER} \leq \mathrm{EH}.$$

Specifically, EH reflects whether any relevant evidence is identified, ER quantifies the degree of partial evidence coverage, and AR evaluates whether the complete evidence set required by the question is recovered. Together with AnswerAcc, these metrics allow us to distinguish final answer correctness from the completeness of the evidence identified during answering.

D Experimental Details

D.1 Model Configurations

We use the instruction-tuned checkpoints `Meta-Llama-3.1-8B-Instruct`, `Qwen2.5-7B-Instruct`, and `glm-4-9b-chat`. Llama and Qwen are loaded from frozen repository snapshots; GLM is loaded from a frozen local checkpoint with its configuration recorded in the run manifest. All three use their native chat templates and BF16 inference. Local runs use NVIDIA RTX 5090 (32 GB) and RTX 5880 Ada (48 GB) hosts. The archived serving environment uses vLLM 0.23.0; launch commands, environment inventories, checkpoint paths, and configuration hashes are retained with each run.

Table 6 lists answer-generation settings. Local serving permits two concurrent sequences per service. The maximum output length is 2,048 tokens, distinct from the model context window and any

Table 6: Answer-generation configurations. Context denotes the configured local serving window, not the model's advertised maximum. "Unspecified" indicates a setting not controlled in the API request. The memory-system reader uses the Llama configuration.

Model	Precision	Context	Temp.	Top-p	Seed	Max. output
Llama-3.1-8B	BF16	40,960	0	1	42	2,048
Qwen2.5-7B	BF16	32,768	0	1	42	2,048
GLM-4-9B	BF16	40,960	0	1	42	2,048
GPT-5.4-mini	Unspecified	Unspecified	0	1	Unspecified	2,048

retrieved-input budget. Local answer generation uses repetition penalty 1.0. The rendered prompt is checked against the configured context window with output tokens reserved; overflowing inputs are not silently truncated.

GPT-5.4-mini is accessed through a third-party OpenAI-compatible endpoint using the requested model identifier `gpt-5.4-mini`. Requests set `reasoning_effort=none` and `response_format=json_object`; they are issued serially without automatic request retries. No inference seed or repetition penalty is specified for the API model, and its server-side precision and chat template are not locally controlled. The archived response identifier records the requested service rather than independently verifying the backend checkpoint.

All readers return an answer string or `null`, together with evidence citations containing a knowledge-unit ID and a quotation. The output protocol permits at most eight evidence entries and 160 characters per quotation. Local readers use schema-constrained JSON generation; the API response is checked against the same answer/evidence structure after generation. Transport errors, output truncation, and format failures are logged separately from ordinary incorrect or null answers. Technical recovery runs retain their source manifests; an incorrect answer is not itself a retry criterion.

D.2 Memory-System Configurations

A-MEM, Mem0, and HippoRAG 2 share Llama-3.1-8B-Instruct for answer generation. Their memory construction and retrieval are implemented through the frozen adapters used in the experiment archive. The evaluation follows a common source-reading interface: each method writes and retrieves through its own backend, retrieved records are mapped to original knowledge-unit IDs, and a shared reader answers from the corresponding original text. Thus, the evaluated systems are adaptations of the memory backends, rather than reproductions of every component of their original question-answering pipelines.

Table 7: Retrieval configurations of memory systems.

Method	Embedding model	Top-k	Retrieval budget	Reader budget
A-MEM	all-MiniLM-L6-v2	20	8,192	8,192
Mem0	BAAI/bge-m3	20	8,192	8,192
HippoRAG 2	nvidia/NV-Embed-v2	20	8,192	8,192

The embedding checkpoints are frozen per method. The archived Mem0 environment uses `mem0ai` 2.0.20. HippoRAG 2 uses FP16 embedding inference with batch size one and a configured embedding sequence limit of 2,048 tokens; its answer reader remains BF16. The adapters and runtime patches are identified by snapshot hashes rather than an assumed common package version. Memory-construction output limits are separate from the answer limit: A-MEM permits up to 4,096 tokens per construction call, while Mem0 and HippoRAG 2 permit up to 8,192, with task-specific extraction limits recorded in their configurations.

All adapters are called with $k = 20$ and an 8,192-token retrieval budget. Because their native outputs differ, equal k does not imply an equal number of original knowledge units. Source IDs are deduplicated in retrieval order. The reader-material selector then counts each rendered unit, including its ID marker, using the Llama tokenizer and includes a whole unit only if it fits within the separate 8,192-token reader-material budget. Units that do not fit are logged and skipped without truncating

their text. Title, question, and system instructions are accounted for in the subsequent model-window check. Full-context base models do not use this retrieval-material cap.

For Mem0, a provenance sidecar tracks which input knowledge units contributed to memory insertions and content updates. These mappings establish operational source lineage, not independent semantic sufficiency. Direct-Evidence also passes through memory-system writing and retrieval before the common reader; it therefore differs from direct presentation of the evidence text to a base model. Reference answers and gold reference-ID lists are excluded from model-facing query objects.

D.3 Condition Execution

State Initialization and Isolation. Each memory-system case has a namespace determined by its condition, stratum, question ID, and input fingerprint. Before ingestion, the runner resets the namespace and checks that retrieval returns no previous records; source-lineage mappings are reset as well. Embedding models and service clients may be reused across cases, while question-specific memory contents are isolated. The exported case records contain only the selected condition's input units and current question. Predictions and scoring feedback are stored outside the ingestion stream.

The different conditions are executed as independent cases using their frozen input views. Base models receive the relevant view in a fresh request. Memory systems reconstruct the corresponding view by ingesting its knowledge units sequentially from an empty namespace. Thus, the Immediate-Memory $\rightarrow$ Delay-2K $\rightarrow$ Delay-4K comparison aligns nested input prefixes; it does not carry a previous checkpoint's question or answer into a later condition. No answer-generation call is inserted while reconstructing a prefix.

Condition-Specific Execution. No-Evidence supplies only the question, with no article materials or external retrieval, and is evaluated for base models. Direct-Evidence supplies the annotated source units in their original order. For ordered reference positions $e_1 < \cdots < e_m$, Immediate-Memory uses the original prefix $(K_1, \ldots, K_{e_m})$, with no protocol step clearing that prefix before the question. Base models read the frozen prefix as context; memory systems ingest it through their own backends before retrieval. Delay-2K and Delay-4K extend this same prefix using the frozen continuation sequence.

Refresh-4K appends an exact copy of the Direct-Evidence packet to the Delay-4K input. For memory systems, the runner independently reconstructs the unqueried Delay-4K prefix, records its boundary, and then appends the evidence without resetting between prefix and suffix. It does not restore an internal state left by a previous Delay-4K answer. Prefix hashes and suffix counts check input consistency. The base-model implementation presents the corresponding combined text in a fresh request. Question text and reference answers remain fixed across conditions.

Delayed Input Construction and Token Counting. Continuation first consumes the remaining units of the source article. If these are insufficient, additional articles are ingested in their original within-article order. The candidate pool contains 75 articles outside the 674 annotated source articles. Additional articles are sampled without replacement using a fixed random seed for each question. The sampled article IDs and their ingestion order are recorded to support reproducibility. Appended units receive consecutive IDs; external-source annotations are omitted from model-facing inputs but retained in evaluation-side mappings. The constructed streams contain no earlier checkpoint questions, predictions, or scoring feedback.

The token distance is measured from the end of the Immediate-Memory prefix, using the frozen Qwen2.5-7B-Instruct tokenizer. For a delayed prefix ending at unit j, the additional distance is

$$d(j) = \sum_{i=e_m+1}^{j} \left|\mathrm{Tok}_{\mathrm{Qwen}}(\mathrm{text}(K_i))\right|.$$

Each unit text is tokenized separately without special tokens; titles, unit IDs, questions, and chat formatting are excluded. Delay-2K stops at the first complete-unit boundary with $d(j) \geq 2{,}048$, and Delay-4K stops at the first boundary with $d(j) \geq 4{,}096$. Both distances are measured relative to Immediate-Memory, so Delay-4K does not add another 4K tokens after Delay-2K. The conditions share the same continuation sequence, and the shorter input is a prefix of the longer one. These construction distances are distinct from the Llama-tokenized retrieval and reader budgets in Table 7.

In the final release, the median additional distances are 2,068 and 4,115 tokens. External articles are needed for 1,794 Delay-2K inputs and 1,924 Delay-4K inputs. Four questions share the two delayed checkpoints because a single complete unit crosses both thresholds. Actual distances and source mappings are retained for every question. The appended content is not independently certified irrelevant to the target question.

E Additional Results

E.1 Performance Conditioned on Direct-Evidence Success

The main results in Table 2 report performance over the full evaluation set. To more directly separate memory-dependent degradation from questions that a model cannot solve even when the required evidence is explicitly provided, we additionally condition the analysis on Direct-Evidence success.

For each model M, we define

$$\mathcal{Q}_M^{DE+} = \{q \mid M \text{ answers } q \text{ correctly under Direct-Evidence}\}.$$

We then evaluate Immediate-Memory, Delay-2K, and Delay-4K on the same subset $\mathcal{Q}_M^{DE+}$.

Table 8: Answer accuracy (%) conditioned on Direct-Evidence success. Each row uses the same $n_{\mathrm{DE+}}$ questions in all three conditions.

Model	Split	$n_{\mathrm{DE+}}$	Immediate-Memory	Delay-2K	Delay-4K
Llama-3.1-8B	Easy	697	84.94	79.48	76.90
	Hard	352	76.42	56.25	52.27
Qwen2.5-7B	Easy	686	86.44	76.82	72.01
	Hard	274	66.79	47.45	45.26
GLM-4-9B	Easy	776	87.63	81.44	81.57
	Hard	376	75.27	57.98	58.24

The degradation persists after controlling for Direct-Evidence success. Conditioning on Direct-Evidence success substantially raises the absolute accuracy of the memory-dependent conditions, as expected, but the overall degradation pattern remains. From Immediate-Memory to Delay-4K, Llama-3.1-8B decreases from 84.94% to 76.90% on Easy questions and from 76.42% to 52.27% on Hard questions. Qwen2.5-7B shows a similar pattern, decreasing from 86.44% to 72.01% on Easy and from 66.79% to 45.26% on Hard, while GLM-4-9B drops from 87.63% to 81.57% and from 75.27% to 58.24%, respectively.

Hard questions remain more sensitive to continued ingestion. The substantially larger decline on Hard questions persists even when the analysis is restricted to questions that the same model answers correctly under Direct-Evidence. This indicates that the stronger degradation observed on Hard questions cannot be explained solely by their being intrinsically unsolved under direct access to the required evidence. Instead, continued knowledge ingestion introduces an additional challenge for maintaining and using previously available temporal information.

E.2 Effects of Evidence Reintroduction

This analysis examines whether presenting the original evidence again after continued reading changes answer correctness and final citation coverage in the same way. Table 9 reports answer accuracy; Tables 10 and 11 report EH, ER, and AR. Immediate-Memory provides the reference point before delayed input, Delay-4K measures performance after continued reading, and Refresh-4K appends the Direct-Evidence packet to that delayed input.

Matched comparison and reporting scope. Within every model and stratum, Immediate-Memory, Delay-4K, and Refresh-4K are evaluated on identical question IDs. All models use the full final release of 1,219 Easy and 737 Hard questions, enabling matched cross-condition comparisons. The

execution details, including independent reconstruction of the unqueried memory prefix, are specified in Appendix D.

For Refresh-4K coverage, reference occurrences with exactly identical text are treated as equivalent sources, so citing either the original or reinjected copy covers that reference. This exact-text convention does not require both copies to be cited and does not merge merely similar passages. The archived occurrence-level scores remain separate. EH, ER, and AR measure final citation coverage, not independently verified quotation fidelity or successful use of the cited information. We further examine the magnitude of these changes, the remaining gap to Immediate-Memory, and their relationship to evidence coverage.

Answer improvements leave a gap to immediate performance. Across the displayed configurations, the net answer-accuracy increase ranges from 0.25 to 2.38 percentage points on Easy questions and from 0.14 to 4.21 points on Hard questions. Every Refresh-4K accuracy remains below its matched Immediate-Memory value. For example, GLM-4-9B on Hard questions increases from 39.21% to 43.42%, but remains 6.11 points below its Immediate-Memory accuracy of 49.53%. Llama-3.1-8B changes from 35.41% to 35.55% on the same stratum, corresponding to a net increase of one correct answer among 737 questions, with an 11.13-point gap to Immediate-Memory. These are net changes: improvements and regressions on individual questions can offset one another, so the accuracy difference is not an item-level recovery rate.

Table 9: Comparison of Immediate-Memory, Delay-4K, and Refresh-4K performance, with Δ denoting the gain from Delay-4K to Refresh-4K.

Method	Immediate-Memory		Delay-4K		Refresh-4K		Δ	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Llama-3.1-8B	55.54	46.68	53.57	35.41	53.81	35.55	+0.25	+0.14
Qwen2.5-7B	56.44	34.60	49.71	25.51	52.09	26.59	**+2.38**	+1.09
GLM-4-9B	**62.10**	**49.53**	**59.15**	**39.21**	**60.87**	**43.42**	+1.72	**+4.21**
GPT-5.4-mini	61.27	44.06	56.07	38.46	57.80	39.86	+1.73	+1.40
A-MEM	53.76	46.15	45.66	36.36	46.24	37.76	+0.58	+1.40
Mem0	52.02	44.06	43.93	23.08	44.51	24.48	+0.58	+1.40
HippoRAG 2	54.34	47.55	52.60	34.27	53.18	34.97	+0.58	+0.70

Evidence-coverage changes depend on the model and stratum. The evidence metrics do not improve uniformly after reinjection. On Hard questions, Llama-3.1-8B's ER rises from 53.56% to 61.44% and AR from 40.16% to 50.34%. GPT-5.4-mini also increases both metrics, with ER moving from 56.90% to 64.53% and AR from 41.96% to 46.85%. In contrast, Mem0's Hard ER decreases from 30.58% to 28.97% and AR from 20.98% to 16.78%. A-MEM's Hard AR remains at 34.97% despite increases in EH and ER. Reporting all three metrics therefore distinguishes broader evidence hits or higher average coverage from complete reference coverage.

Table 10: Evidence recall metrics (%) on the **Easy** questions.

Method	Immediate-Memory			Delay-4K			Refresh-4K		
	EH	ER	AR	EH	ER	AR	EH	ER	AR
Llama-3.1-8B	93.11	73.44	51.68	87.04	66.06	44.38	89.50	73.38	53.40
Qwen2.5-7B	75.80	47.75	27.15	70.30	43.09	23.79	63.66	42.65	25.84
GLM-4-9B	85.48	55.42	31.99	78.42	50.34	28.47	82.03	57.01	35.93
GPT-5.4-mini	95.38	70.01	46.82	90.75	64.78	40.46	91.91	64.87	39.88
A-MEM	**96.53**	79.98	59.54	83.82	62.79	42.77	85.55	65.10	47.40
Mem0	81.50	59.85	35.84	73.41	49.67	27.75	76.88	54.23	33.53
HippoRAG 2	94.80	**80.55**	**62.43**	**92.49**	**74.64**	**56.65**	**94.80**	**76.70**	**57.80**

Table 11: Evidence recall metrics (%) on the **Hard** questions

Method	Immediate-Memory			Delay-4K			Refresh-4K		
	EH	ER	AR	EH	ER	AR	EH	ER	AR
Llama-3.1-8B	82.36	**68.91**	55.22	67.44	53.56	40.16	71.51	61.44	**50.34**
Qwen2.5-7B	51.15	35.62	24.69	37.99	25.64	16.82	34.74	25.12	17.64
GLM-4-9B	66.49	46.46	32.02	53.87	36.77	24.56	58.48	43.48	31.07
GPT-5.4-mini	**88.11**	68.74	51.75	**74.13**	**56.90**	**41.96**	**85.31**	**64.53**	46.85
A-MEM	74.83	62.10	49.65	62.94	48.76	34.97	66.43	49.98	34.97
Mem0	72.03	54.00	37.76	43.36	30.58	20.98	44.06	28.97	16.78
HippoRAG 2	79.72	68.36	**55.94**	72.03	56.63	41.26	64.34	51.47	40.56

Answer changes and citation-coverage changes are not interchangeable. Llama-3.1-8B on Hard questions provides a clear aggregate contrast: reinjection produces a 10.18-point increase in AR but only a 0.14-point increase in answer accuracy. On Easy questions, its Refresh-4K AR reaches 53.40%, exceeding the Immediate-Memory value of 51.68%, while answer accuracy remains lower (53.81% versus 55.54%). Conversely, Hard answer accuracy of Mem0 increases from 23.08% to 24.48% while both ER and AR decrease. These comparisons show that a change in answer accuracy cannot serve as a direct substitute for a change in final evidence coverage. They do not establish that the individual questions gaining complete coverage are the same questions whose answers remain incorrect.

The results motivate examining answer outcomes jointly with evidence coverage. Reintroducing the source text changes its recency and repetition and, for memory systems, triggers additional memory processing. Residual errors cannot therefore be assigned exclusively to temporal reasoning, nor can improved citation coverage alone certify successful retention or comprehension. The supported conclusion is narrower: answer accuracy and reference-evidence coverage capture distinct aspects of performance, and both are needed to characterize the effects of evidence reintroduction.

Table 12: Paired changes from Delay-4K to Refresh-4K on the same questions as Table 9. ΔAcc and nominal 95% percentile bootstrap intervals are in percentage points. McNemar p values are exact, two-sided, and unadjusted.

Model	Split	N	ΔAcc	95% CI	McNemar p
Llama-3.1-8B	Easy	1219	+0.25	[-1.73, +2.14]	0.8663
	Hard	737	+0.14	[-2.44, +2.71]	1.0000
Qwen2.5-7B	Easy	1219	+2.47	[+0.25, +4.70]	0.0351
	Hard	737	+1.09	[-1.63, +3.80]	0.4884
GLM-4-9B	Easy	1219	+1.81	[+0.08, +3.54]	0.0448
	Hard	737	+4.21	[+1.76, +6.78]	0.0015

Paired statistical analysis. We assess the uncertainty of the Delay-4K$\rightarrow$ Refresh-4K accuracy changes for the three open base models using paired percentile bootstrap intervals and exact two-sided McNemar tests on matched question IDs. Most changes are small relative to their sampling variability. The nominal 95% bootstrap intervals exclude zero for Qwen2.5-7B on Easy questions and GLM-4-9B on both Easy and Hard questions, while the remaining intervals include zero. The largest and most stable change is observed for GLM-4-9B on Hard questions (ΔAcc $= +4.21$ points, 95% CI $[+1.76, +6.78]$, exact McNemar $p = 0.0015$). These results support interpreting the smaller Refresh-4K gains descriptively rather than as robust improvements.

E.3 Effects of Evidence Cardinality

All-Evidence Recall is strongly sensitive to evidence cardinality. Table 13 stratifies questions by the number of gold evidence units, $k = |G_q|$, and reports Answer Accuracy and All-Evidence Recall (AR) under Immediate-Memory and Delay-4K. Easy and Hard questions are analyzed separately, with identical question IDs used across models and conditions within each cardinality group.

Table 13: Performance by reference-evidence cardinality $k = |G_q|$ on the full final release. Answer Accuracy and AR are reported as percentages within each stratum–cardinality group. Each group uses identical question IDs across models and conditions; n denotes the number of questions.

Method	k	n	Answer accuracy		AR	
			Immediate-Memory	Delay-4K	Immediate-Memory	Delay-4K
Easy						
Llama-3.1-8B	1	365	57.53	50.41	93.42	80.00
	2	405	55.80	54.81	52.35	44.20
	3	255	59.61	62.35	25.88	21.96
	≥ 4	194	45.88	45.36	5.67	7.22
Qwen2.5-7B	1	365	54.25	44.11	75.62	67.40
	2	405	58.02	53.09	11.11	8.40
	3	255	64.71	55.29	3.92	3.92
	≥ 4	194	46.39	45.88	—	—
GLM-4-9B	1	365	58.63	56.16	83.84	77.53
	2	405	62.96	60.25	15.56	12.35
	3	255	69.80	65.10	7.45	5.10
	≥ 4	194	56.70	54.64	1.03	0.52
Hard						
Llama-3.1-8B	1	357	54.62	34.45	82.91	62.46
	2	195	37.95	34.36	41.03	28.72
	3	98	48.98	42.86	28.57	15.31
	≥ 4	87	31.03	33.33	3.45	2.30
Qwen2.5-7B	1	357	35.57	24.37	45.38	31.65
	2	195	29.74	23.08	8.72	4.62
	3	98	40.82	33.67	3.06	2.04
	≥ 4	87	34.48	26.44	—	—
GLM-4-9B	1	357	51.26	35.57	59.66	43.42
	2	195	40.51	37.44	8.72	10.26
	3	98	52.04	40.82	5.10	5.10
	≥ 4	87	59.77	56.32	1.15	1.15

Across models and both strata, AR decreases sharply as the number of required evidence units increases. For example, on Easy questions under Immediate-Memory, the AR of Llama-3.1-8B decreases from 93.42% for $k = 1$ to 5.67% for $k \geq 4$. Similar patterns are observed for Qwen2.5-7B and GLM-4-9B, as well as on Hard questions. This dependence is partly expected from the definition of AR, which requires every gold evidence unit to be identified and therefore imposes a stricter criterion as the reference evidence set grows.

Answer accuracy shows a weaker and non-monotonic relationship with evidence cardinality. In contrast to AR, Answer Accuracy does not decrease monotonically with k. For instance, on Easy questions, Llama-3.1-8B achieves 57.53%, 55.80%, 59.61%, and 45.88% under Immediate-Memory for $k = 1$, 2, 3, and $k \geq 4$, respectively. Similar non-monotonic patterns are observed for the other models. The change from Immediate-Memory to Delay-4K also varies across cardinality groups rather than increasing systematically with the number of required evidence units.

These results indicate that evidence cardinality has a strong effect on complete reference-evidence coverage, whereas its association with final answer correctness is substantially weaker. Therefore, lower AR on questions requiring more evidence should not be interpreted as a proportional reduction in answering performance. This further supports treating evidence completeness and answer correctness as complementary rather than interchangeable measures.

E.4 Performance across Temporal Relations

We next disaggregate answer accuracy by the eight Wikidata relations defined in Appendix B. This analysis tests whether the aggregate Immediate-Memory–Delay-4K comparison is shared across relation types or concentrated in particular parts of the dataset. Table 14 reports relation-level answer

accuracy on the full final release. Easy and Hard questions are reported separately, with identical question IDs across models and conditions within each pair of relation and stratum.

Table 14: Answer accuracy (%) by temporal relation on the full final release. Each cell uses all n questions in its stratum–relation group, with identical IDs across models and conditions. Relation semantics are defined in Table 4. A dash denotes zero; *n/a* denotes an empty group.

Relation	n	Llama-3.1-8B		Qwen2.5-7B		GLM-4-9B	
		Immediate-Memory	Delay-4K	Immediate-Memory	Delay-4K	Immediate-Memory	Delay-4K
Easy							
P39	614	44.63	40.07	44.79	38.27	55.21	51.47
P54	257	62.65	64.98	71.21	63.42	72.76	66.93
P937	97	68.04	69.07	69.07	61.86	70.10	74.23
P108	126	60.32	65.87	61.90	61.90	62.70	68.25
P488	73	89.04	76.71	76.71	64.38	76.71	68.49
P131	27	55.56	51.85	51.85	25.93	33.33	25.93
P1037	19	100.00	100.00	78.95	78.95	100.00	89.47
P4791	6	16.67	16.67	—	16.67	—	16.67
Hard							
P39	403	41.94	28.29	28.04	19.35	48.39	39.45
P54	132	53.03	37.12	50.00	31.06	57.58	38.64
P937	89	48.31	46.07	41.57	37.08	47.19	41.57
P108	59	61.02	49.15	47.46	38.98	57.63	42.37
P488	5	60.00	80.00	80.00	60.00	60.00	40.00
P131	42	40.48	47.62	11.90	16.67	26.19	21.43
P1037	7	85.71	57.14	28.57	42.86	57.14	85.71
P4791	—	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*

Aggregate results are dominated by frequent relations. Position-held questions (P39) account for 1,017 of the 1,956 questions, followed by sports-team membership (P54) with 389. These two relations therefore contribute most to the question-averaged results. For P39, all three models have lower accuracy after continued reading in both strata. On Hard questions, Llama-3.1-8B decreases from 41.94% to 28.29%, Qwen2.5-7B from 28.04% to 19.35%, and GLM-4-9B from 48.39% to 39.45%. Hard P54 questions show the same direction of change, with accuracies decreasing from 53.03% to 37.12%, 50.00% to 31.06%, and 57.58% to 38.64%, respectively.

The direction and size of changes vary across relations. The aggregate decrease does not hold for every relation–stratum group. On Easy employer questions (P108), Llama-3.1-8B increases from 60.32% to 65.87%, GLM-4-9B from 62.70% to 68.25%, and Qwen2.5-7B remains at 61.90%. Easy work-location questions (P937) also show increases for Llama and GLM, but a decrease for Qwen. These variations are compatible with differences in the supplied context, reference composition, and model responses; relation-level accuracy alone cannot assign a specific failure mechanism. In particular, the continuation is not independently certified to contain only irrelevant information.

Small groups limit relation-specific conclusions. The Hard chairperson group (P488) contains only five questions, so one answer changes its accuracy by 20 percentage points. Hard director-or-manager questions (P1037) contain seven examples; commander questions (P4791) contain six Easy examples and no Hard examples. Empty groups are marked as unavailable rather than assigned zero accuracy. We report the observed proportions and group sizes without ranking rare relations or interpreting isolated changes as robust effects. Questions can also share source articles, so these rows should not be interpreted as collections of independent article-level observations. These limitations preclude interpreting the relation analysis as an intrinsic ordering of temporal reasoning difficulty.

F Case Studies

We present three examples to illustrate the diagnostic patterns observed in the quantitative results, particularly the discrepancy between evidence completeness and answer correctness under continued knowledge ingestion and evidence reintroduction.

Selection and reporting. The examples are selected from the final question list for interpretable temporal contrasts and recoverable execution records. We reproduce the original questions, reference answers, model answers, and cited knowledge-unit IDs. The tables list all distinct cited IDs; the accompanying text displays the source passages and citation quotations needed to interpret each case. Reference passages remain in the delayed inputs of these full-context base-model runs, so absence from the final citations does not establish that information was deleted from memory. We verify that each delayed input extends the immediate prefix and that Refresh-4K appends exact copies of the Direct-Evidence units to the same unqueried Delay-4K input. Repeated copies count as equivalent sources. These cases illustrate observed outputs rather than reveal the model's internal reasoning process.

F.1 Evidence Complete and Answer Correctness

Question and input. For `easy_1154`, Qwen2.5-7B receives the question "For which team did A. J. Soares play in 2010?" The reference answer is *Orange County Blue Star*, supported by one reference unit:

> `m_007`: In 2010 Soares made one appearance for Orange County Blue Star in the USL Premier Development League.

Immediate-Memory contains units `m_001`–`m_007`. Continued reading expands the input to 116 units, including the following later event:

> `m_008` (not a reference unit): Soares was chosen by New England Revolution as the sixth overall pick in the 2011 MLS SuperDraft.

Refresh-4K appends `m_117`, an exact copy of `m_007`, yielding 117 input units. The question is unchanged across all three conditions.

Table 15: A. J. Soares (Qwen2.5-7B): evidence and answer recovery. ER is reported in percent; coverage is computed over distinct reference units, treating reinjected copies as equivalent sources.

Condition	Model answer	Cited IDs	Coverage	ER (%)	Outcome
Immediate-Memory	Orange County Blue Star	`m_001`, `m_007`	All	100	Correct
Delay-4K	New England Revolution	`m_008`	None	—	Wrong
Refresh-4K	Orange County Blue Star	`m_007`	All	100	Correct

Observed transition. Table 15 shows a correct immediate answer with complete reference coverage. Under Delay-4K, the model instead answers "New England Revolution" and cites only `m_008`, reproducing its 2011 sentence. This citation concerns the same person but a different year from the requested 2010. Following evidence reintroduction, the answer returns to "Orange County Blue Star", and the model again quotes the complete sentence from `m_007`. It cites the original occurrence rather than the newly appended copy; either occurrence is sufficient under the equivalence rule.

This case illustrates a joint recovery of reference coverage and answer correctness. The observed error involves selection of a later event, not absence of every mention of the subject or team. Because `m_007` remains in the delayed input, the citation change supports an evidence-selection interpretation but does not distinguish storage failure, attention allocation, or temporal misapplication. The non-reference identity citation `m_001` in the immediate response is paraphrased rather than verbatim; the diagnostic reference quotation from `m_007` is faithful to the supplied text.

F.2 Complete Evidence with an Incorrect Answer

Question and input. For `easy_0653`, Qwen2.5-7B is asked "Which position was held by Ernst Kuntscher from 1965 to 1969?" The reference answer is *member of the German Bundestag*. Its sole reference unit explicitly covers the queried interval:

> `m_004`: Kuntscher was a member of the German Bundestag since its first election in 1949 until 1969.

The immediate input ends at m_004. The delayed input contains 125 units, including an account of a committee chairmanship:

> m_006 (not a reference unit): From 9 July 1954 to 1961, he was chairman of the Bundestag Committee for Expellees, and from 1961 to 1965 of the Bundestag Committee for Equalization of Burdens.

Refresh-4K adds m_126, an exact copy of m_004.

Table 16: Ernst Kuntscher (Qwen2.5-7B): an incorrect answer with complete reference coverage. ER is reported in percent; coverage is computed over distinct reference units, treating reinjected copies as equivalent sources.

Condition	Model answer	Cited IDs	Coverage	ER (%)	Outcome
Immediate-Memory	member of the German Bundestag	m_004	All	100	Correct
Delay-4K	Bundestag Committee for Equalization of Burdens	m_004, m_006	All	100	Wrong
Refresh-4K	Bundestag Committee for Equalization of Burdens	m_006	None	—	Wrong

Observed mismatch. The immediate response is correct. Under Delay-4K, the model quotes the entire reference sentence from m_004 and additionally quotes "from 1961 to 1965 of the Bundestag Committee for Equalization of Burdens." Nevertheless, its answer is "Bundestag Committee for Equalization of Burdens" (Table 16). Thus, EH, ER, and AR all equal one on this question while the answer is incorrect. Both displayed citation quotations match their corresponding input passages.

The error is more specific than incomplete reference coverage: the selected committee is associated with a period ending in 1965, whereas the question concerns 1965–1969. Its cited passage does not support that committee chairmanship over the requested interval; the concurrently cited parliamentary-membership passage does. This is an observable failure to align the answer with the available temporal evidence, although the output does not reveal the internal computation responsible. Refresh-4K leaves the incorrect answer unchanged and retains only the committee citation, so complete reference coverage in the delayed response is not preserved after reinjection. The principal diagnostic observation here is the Delay-4K response itself: faithful quotation of all reference evidence co-occurs with an incompatible answer.

F.3 Evidence Recovery without Answer Recovery

Question and input. For easy_1066, GLM-4-9B receives "From 2009 to 2010, which team did the player Laurence Ovens belong to?" The reference answer is *Newcastle Falcons*. The annotated evidence consists of two units that describe successive affiliations:

> m_002: After five seasons with Bath Rugby (2004–2009), he signed for Newcastle Falcons.
>
> m_003: In July 2010, Ovens signed for the Bedford Blues, giving him an opportunity to play every week for a promising championship side.

Immediate-Memory contains three units; Delay-4K contains 118. Refresh-4K appends m_119 and m_120, which reproduce m_002 and m_003, respectively. The reference answer and evidence mapping are unchanged.

Which evidence is covered matters. The immediate response answers "Newcastle Falcons" and quotes m_002. The delayed response instead answers "Bedford Blues" and cites m_003, with the shortened quotation "In July 2010, Ovens signed for the Bedford Blues." Both responses have ER of 50%, but they cover different reference units and have different answer outcomes (Table 17). The immediate citation identifies the reference team; the delayed citation identifies the subsequent team. Equal coverage fractions therefore need not represent equally useful evidence for the requested answer.

Table 17: Laurence Ovens (GLM-4-9B): complete evidence recovery without answer recovery. ER is reported in percent; coverage is computed over distinct reference units, treating reinjected copies as equivalent sources.

Condition	Model answer	Cited IDs	Coverage	ER (%)	Outcome
Immediate-Memory	Newcastle Falcons	m_002	Part	50	Correct
Delay-4K	Bedford Blues	m_003	Part	50	Wrong
Refresh-4K	Bedford Blues	m_119, m_120	All	100	Wrong

Full recovery of citations does not repair the answer. After reinjection, the model quotes both appended units in full, yielding EH$=1$, ER$=1$, and AR$=1$ under the equivalent-source convention. Its answer nevertheless remains "Bedford Blues". The gain from 50% to 100% ER and from zero to one AR is therefore accompanied by no answer recovery on this same question. The returned answer selects the July 2010 destination rather than the affiliation that preceded that transfer. Because the question uses year-level endpoints and the transfer occurs within 2010, this interpretation follows the benchmark's interval-target convention; it does not assert that Ovens had only one affiliation during the entire calendar year.

This case makes the aggregate discrepancy in Appendix E.2 concrete: the model can reproduce both the preceding affiliation and its temporal boundary while still returning the subsequent affiliation. Together, the three examples distinguish missing reference citations, incorrect answering despite faithful reference quotations, and evidence recovery without answer recovery. They support inspecting citation content and temporal alignment alongside AnswerAcc and coverage metrics, without treating any one output pattern as a direct measurement of an internal memory or reasoning mechanism.