TIMEBOUND: BENCHMARKING TEMPORAL VALIDITY IN LLM AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents retrieve from interaction histories in which facts may expire, be cancelled, or be superseded by later updates. Semantic relevance alone does not capture whether a retrieved memory is valid for the time referenced by a query. We introduce TIMEBOUND, a benchmark for this setting with 1000 long-history examples spanning eight temporal-memory phenomena. Each memory is annotated with observation time, event time, validity interval, status, gold answer, and gold evidence. We also introduce TimeBound-RAG, a simple diagnostic retriever that combines semantic relevance with query-time validity and memory status. Across four open-weight readers, TimeBound-RAG reaches mean relaxed accuracy of 0.719, compared with 0.475 for Semantic RAG and 0.537 for Full History. Semantic RAG obtains slightly higher evidence F1 (0.577 vs. 0.541), but retrieves substantially more invalid memories (0.374 vs. 0.152). The results separate validity-sensitive retrieval failures from temporal computation, recurrence, scheduling, and long-memory localization failures that require mechanisms beyond retrieval.

Keywords: LLM agents; temporal memory; retrieval-augmented generation; long-term memory; temporal validity

1 INTRODUCTION

Large language model (LLM) agents increasingly operate over persistent interaction histories rather than isolated prompts. Such histories evolve: meetings are rescheduled, access codes are replaced, commitments are cancelled, and preferences change. Users may also ask about a previous state rather than the current one. In these settings, semantic relevance alone is not sufficient: a retrieved memory must also be valid for the temporal frame of the query.

Retrieval-augmented generation generally selects evidence according to semantic relevance or estimated usefulness Guu et al. (2020); Lewis et al. (2020); Asai et al. (2024); Yan et al. (2024). Long-term memory systems extend LLMs with persistent storage and retrieval Wang et al. (2023); Zhong et al. (2024); Gutiérrez et al. (2024; 2025). Temporal QA instead evaluates ordering, dates, durations, and time-sensitive answers Ning et al. (2020); Chen et al. (2021); Tan et al. (2023); Chu et al. (2024). These settings leave an important gap: a memory may be highly related to the query while being expired, cancelled, superseded, or valid only at another point in time.

We introduce TIMEBOUND, a benchmark for studying this validity-changing memory setting. TimeBound-Long contains 1000 interaction histories distributed evenly across eight temporal phenomena. Each memory distinguishes observation time, event time, validity interval, and status, allowing semantic relevance and temporal validity to be evaluated separately.

We also introduce TimeBound-RAG as a deliberately simple diagnostic retriever. It combines semantic similarity with query-time compatibility and memory status. Across four local open-weight readers, TimeBound-RAG reaches mean relaxed accuracy of 0.719, compared with 0.475 for Semantic RAG and 0.537 for Full History. At the retrieval level, Semantic RAG attains slightly higher evidence F1 (0.577 vs. 0.541), but retrieves substantially more invalid memories (0.374 vs. 0.152). Thus, evidence overlap and temporal validity can diverge.

Our contributions are:

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.

- We introduce TIMEBOUND, a controlled benchmark for validity-changing memory in long LLM-agent histories.
- We evaluate temporal validity as an explicit retrieval signal across four open-weight readers and six context-construction regimes.
- We separate validity-selection failures from temporal computation failures, identifying where retrieval is useful and where arithmetic, recurrence, scheduling, or long-memory localization remains necessary.

2 RELATED WORK

Temporal reasoning. Temporal QA benchmarks study ordering, duration, time-dependent answers, and temporal generalization Ning et al. (2020); Qin et al. (2021); Chen et al. (2021); Tan et al. (2023); Chu et al. (2024); Wang & Zhao (2024). Temporal KGQA represents time-indexed facts and intervals explicitly Saxena et al. (2021); Mavromatis et al. (2022); Chen et al. (2023); Qian et al. (2024). TimeBound instead treats the underlying object as an evolving natural-language memory whose operative status may change after observation.

Retrieval and long-term memory. RAG grounds generation in retrieved evidence Guu et al. (2020); Lewis et al. (2020), while later methods add reflection, correction, structured retrieval, and time-sensitive retrieval Asai et al. (2024); Yan et al. (2024); Zhang et al. (2025). LongMem, MemoryBank, HippoRAG, LoCoMo, and LongMemEval study persistent or long-horizon memory Wang et al. (2023); Zhong et al. (2024); Gutiérrez et al. (2024); Maharana et al. (2024); Wu et al. (2025). Our focus is complementary: not only whether a memory can be retrieved, but whether it is operative at the query time.

Planning and state change. Planning benchmarks expose failures involving ordering, state transitions, durations, and constraints Valmeekam et al. (2023a;b); Zheng et al. (2024). These capabilities overlap with temporal memory but require operations beyond retrieval. A broader map of related work is provided in Appendix **??**.

3 TIMEBOUND

3.1 TEMPORAL MEMORY REPRESENTATION

Figure 1 summarizes the TimeBound pipeline. An interaction history is represented as a sequence of memory events,

$$\mathcal{H} = (e_1, \dots, e_n), \qquad e_i = (x_i, \tau_i^{\text{obs}}, \tau_i^{\text{evt}}, v_i, s_i, r_i),$$

where x_i is the memory text, τ_i^{obs} is its observation time, τ_i^{evt} is the time of the described event, $v_i = [\tau_i^{\text{start}}, \tau_i^{\text{end}}]$ is its validity interval, s_i is its status, and r_i optionally links it to a memory that it updates, cancels, confirms, or supersedes.

For query q, let τ_q denote its temporal reference point. Temporal admissibility is defined as

$$\begin{aligned}\text{valid}(e_i, \tau_q) = \mathbf{1}\{\tau_i^{\text{start}} \le \tau_q \le \tau_i^{\text{end}}\} \\ \cdot \mathbf{1}\{s_i \notin \{\text{cancelled}, \text{superseded}\}\}.\end{aligned}$$

Observation time and event time remain distinct, allowing a recently observed memory to describe an earlier event.

3.2 VALIDITY-AWARE RETRIEVAL

TimeBound-RAG combines semantic and temporal signals:

$$S(q, e_i, \tau_q) = \alpha\ \text{sim}(q, x_i) + \beta\, R_{\text{time}}(e_i, \tau_q) - \gamma\, O(e_i, \tau_q),$$

where sim denotes semantic similarity, R_{time} measures compatibility with the query-time frame, and O penalizes obsolete or otherwise incompatible memories. The retrieved context is

$$C_k(q) = \text{TopK}_k\left(\{S(q, e_i, \tau_q) : e_i \in \mathcal{H}\}\right).$$

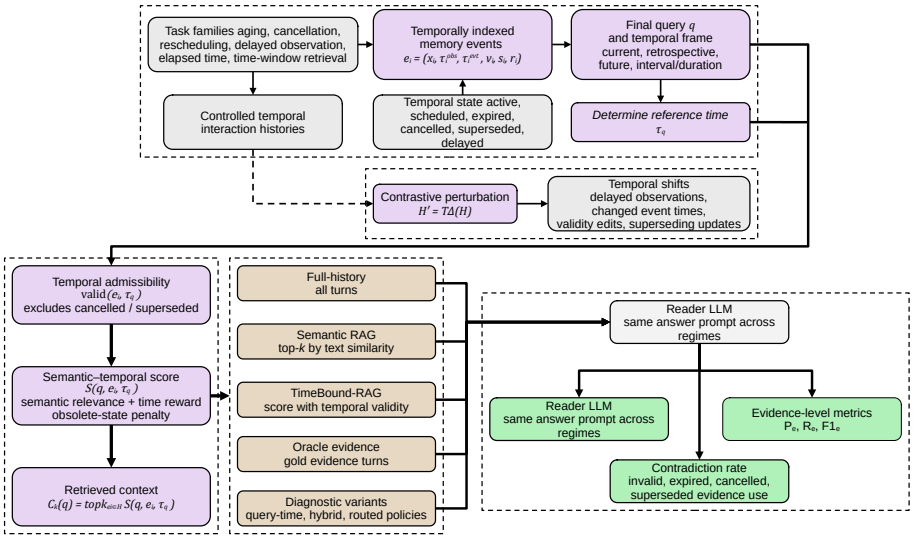


Figure 1: TimeBound represents interaction histories as temporally annotated memories. Retrieval constructs a compact evidence context, which is evaluated at both the evidence and answer levels.

Table 1: The eight balanced TimeBound-Long task families. Each family contains 125 examples.

Family	Temporal phenomenon	n	Primary diagnostic role
Aging facts	Expiration	125	Reject expired current-state memories
Cancellation	Explicit invalidation	125	Reject cancelled memories
Conflicting updates	Competing states	125	Resolve semantically similar updates
Delayed observations	Observation/event mismatch	125	Separate reporting time from event time
Elapsed-time reasoning	Duration	125	Compute over retrieved temporal evidence
Periodic events	Recurrence	125	Interpret repeated schedules
Rescheduling	Future-state replacement	125	Recover the operative update
Time-window retrieval	Retrospective validity	125	Ground validity at the queried time

The retriever is intentionally diagnostic. It does not learn a temporal world model or solve structured constraints. Its purpose is to isolate the effect of explicit temporal validity on evidence selection. Recency alone cannot represent the same distinction: a recent observation can describe an old event, while an expired memory may still be correct for a retrospective query.

3.3 Benchmark Design

TimeBound-Long contains 1000 examples, with 125 examples in each of eight task families. Table 1 summarizes the benchmark structure and the temporal phenomenon isolated by each family.

Histories contain 15–46 memory turns, with a mean of 29.8, while gold evidence contains 1.5 turns on average. Examples are generated from structured temporal states and rendered as natural-language histories. Gold answers and evidence are derived from those structured states rather than generated by an LLM.

Detailed generation statistics, memory-state counts, and validation checks are reported in Appendix A, Table 6.

Table 2: Relaxed accuracy on TimeBound-Long across four readers and six context regimes. Δ is TimeBound-RAG minus Semantic RAG.

Reader	Full Hist.	Semantic	Recency	Sem.+Rec.	TimeBound	Oracle	Δ
Mistral-7B-v0.3	0.550	0.502	0.163	0.384	**0.714**	0.722	+0.212
Qwen2.5-7B	0.513	0.404	0.236	0.341	**0.711**	0.677	+0.307
Qwen2.5-Coder-1.5B	0.437	0.404	0.365	0.376	**0.664**	0.539	+0.260
Qwen2.5-Coder-7B	0.649	0.591	0.164	0.466	**0.785**	0.866	+0.194
Mean	0.537	0.475	0.232	0.392	**0.719**	0.701	+0.243

4 Experimental Setup

We evaluate four local open-weight readers: `Mistral-7B-Instruct-v0.3`, `Qwen2.5-7B-Instruct`, `Qwen2.5-Coder-1.5B-Instruct`, and `Qwen2.5-Coder-7B-Instruct`.

Each reader receives the same answer-generation prompt under six context regimes: Full History, Semantic RAG, Recency RAG, Semantic+Recency RAG, TimeBound-RAG, and Oracle Evidence. Oracle Evidence uses annotated gold evidence and serves only as a diagnostic condition.

The complete evaluation contains 82 runs: six retrieval-only comparisons, 18 retrieval ablations, 24 four-reader experiments, and 34 external diagnostic runs. The run structure is detailed in Appendix B, Table 7.

We report strict and relaxed answer accuracy. Evidence precision, recall, and F1 are

$$P_e = \frac{|E_{\text{pred}} \cap E_{\text{gold}}|}{|E_{\text{pred}}|}, \qquad R_e = \frac{|E_{\text{pred}} \cap E_{\text{gold}}|}{|E_{\text{gold}}|},$$

$$F1_e = \frac{2P_eR_e}{P_e + R_e}.$$

We additionally measure invalid retrieval, contradiction rate, and retrieved context size. Invalid retrieval measures the fraction of selected evidence that is temporally incompatible with the query.

5 Results

5.1 Reader-Level Results

Table 2 reports the complete reader-by-regime comparison. TimeBound-RAG has higher relaxed accuracy than Semantic RAG and Full History for each of the four readers. Mean relaxed accuracy is 0.719, compared with 0.475 for Semantic RAG and 0.537 for Full History. Qwen2.5-Coder-7B reaches the highest TimeBound-RAG score, 0.785.

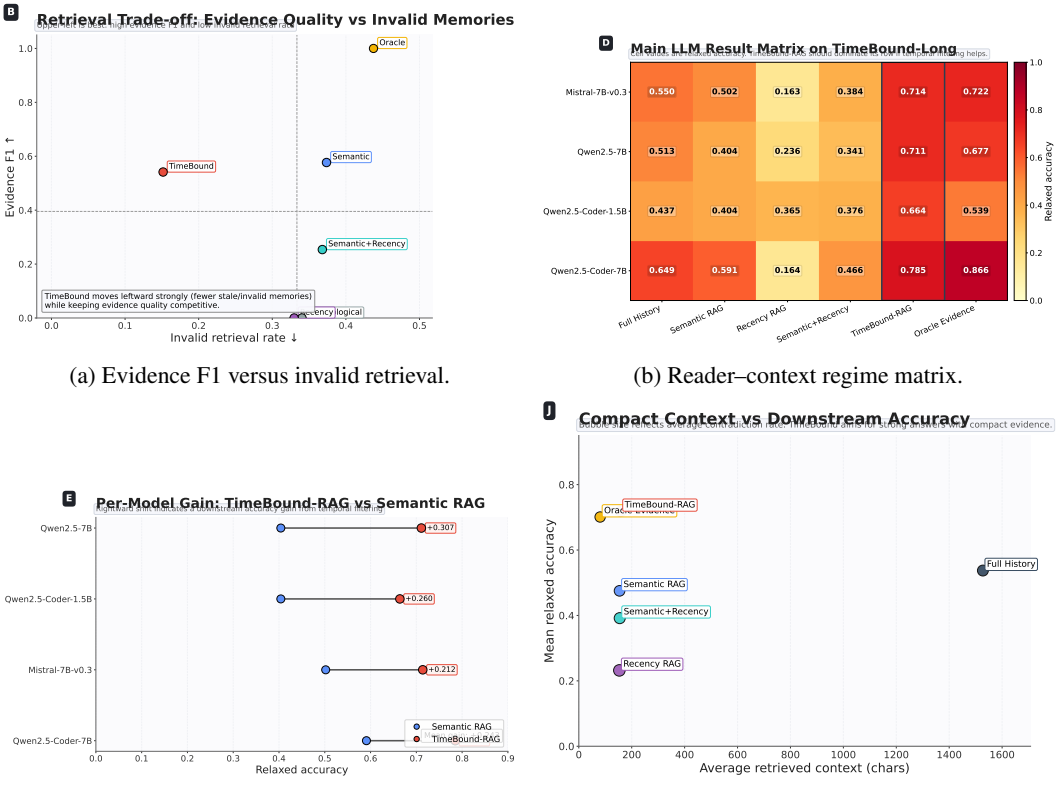
The comparison is also informative about the source of the effect. Recency RAG reaches mean relaxed accuracy of 0.232, while Semantic+Recency reaches 0.392. Thus, preferentially selecting recent memories does not reproduce the behavior of validity-aware retrieval.

Figure 2 gives a compact view of the main experimental result. Panel (a) shows the evidence-validity trade-off, panel (b) compares all reader–regime combinations, panel (c) shows the TimeBound–Semantic difference for each reader, and panel (d) relates context size to answer accuracy.

5.2 Evidence Overlap Is Not Temporal Validity

The retrieval-only results in Table 3 clarify the reader-level pattern. Semantic RAG obtains slightly higher evidence F1 than TimeBound-RAG (0.577 vs. 0.541), but selects substantially more invalid evidence (0.374 vs. 0.152). Thus, higher overlap with annotated evidence does not necessarily imply a temporally safer context.

This distinction is also visible in the retrieval-budget and component ablations reported in Appendix D. Table 9 shows that the highest evidence F1 is obtained at top-1, while Tables 10 and 11



(a) Evidence F1 versus invalid retrieval.

(b) Reader–context regime matrix.

(c) TimeBound-RAG relative to Semantic RAG.

(d) Context size versus relaxed accuracy.

Figure 2: Main TimeBound-Long results. Temporal validity changes the retrieved evidence, the effect is consistent across the four readers, and the result does not require full-history context.

Table 3: Retrieval-only comparison on TimeBound-Long. Lower invalid retrieval is better.

Retriever	Evidence F1 ↑	Invalid retrieval ↓
Oracle Evidence	1.000	0.438
Semantic RAG	**0.577**	0.374
TimeBound-RAG	0.541	**0.152**
Semantic+Recency	0.254	0.368
Recency	0.000	0.330
Chronological	0.000	0.340

show that semantic overlap can increase when status information is removed even as invalid retrieval becomes more common.

5.3 Which Temporal Phenomena Drive the Result?

Table 4 decomposes the controlled long-history result by task family. The clearest differences occur in validity-changing cases. For conflicting updates, relaxed accuracy changes from 0.448 under Semantic RAG to 0.720 under TimeBound-RAG. For rescheduling, it changes from 0.336 to 0.944. Aging facts also show a large reduction in contradiction rate, from 0.848 to 0.192.

The same table also identifies the boundary of retrieval. Elapsed-time reasoning receives almost unchanged evidence F1 despite low answer accuracy, indicating that evidence selection alone does not solve the arithmetic. Periodic-event questions similarly depend on recurrence interpretation. Time-window retrieval exposes a different issue: validity must be evaluated relative to the requested historical time rather than the current state.

Table 4: Task-wise controlled long-history comparison between Semantic RAG and TimeBound-RAG.

Family	Relaxed Acc. Sem → TB	Evidence F1 Sem → TB	Contrad. Sem → TB	Interpretation
Aging facts	.488 → .656	.246 → .341	.848 → .192	Expired evidence is filtered more reliably
Cancellation	.904 → .904	.566 → .554	.136 → .136	Relaxed answer accuracy is already high
Conflicting updates	.448 → .720	.083 → .405	.000 → .000	Update selection is the main bottleneck
Delayed observations	.352 → .376	.000 → .000	.000 → .000	Event/report-time distinction remains difficult
Elapsed-time reasoning	.064 → .096	.571 → .570	.000 → .000	Retrieved evidence still requires arithmetic
Periodic events	.896 → .904	.000 → .058	.000 → .000	Recurrence is primarily reader-side
Rescheduling	.336 → .944	.043 → .449	.000 → .000	Operative update selection dominates
Time-window retrieval	.840 → .160	.563 → .461	.000 → .000	Retrospective validity requires query-time grounding

Table 5: Mean relaxed accuracy, contradiction rate, and context size across the four readers.

Regime	Relaxed Acc. ↑	Contrad. ↓	Avg. chars ↓
Full History	0.537	0.104	1527.4
Semantic RAG	0.475	0.092	154.0
Recency RAG	0.232	0.137	153.2
Semantic+Recency	0.392	0.112	154.7
TimeBound-RAG	**0.719**	0.089	158.2
Oracle Evidence	0.701	**0.081**	**80.6**

5.4 Context Efficiency

Table 5 compares answer quality with average context size. Full History supplies 1527.4 characters on average, whereas TimeBound-RAG uses 158.2. Semantic RAG is similarly compact at 154.0 characters, indicating that the TimeBound result is not explained by context compression alone.

5.5 Qualitative Example

Figure 3 illustrates the evidence-selection problem directly. Semantically related memories can occupy different temporal states—active, expired, superseded, or scheduled—under the same query. The selected context therefore depends on more than topical similarity.

Additional examples comparing Semantic RAG, Full History, and different readers are provided in Appendix E, Figures 7– 9.

6 Discussion and Limitations

The central empirical distinction is between semantic relevance and temporal operativity. A stale memory may be highly similar to a query while being incorrect for the current state. Conversely, a memory that is no longer active can still be correct for a retrospective question. Retrieval over agent memory therefore depends jointly on content and temporal reference.

The results also suggest that longer context is not a substitute for temporal selection. Full History exposes the reader to all relevant updates but also retains stale alternatives. TimeBound-RAG uses approximately one tenth of the Full History context while reaching higher mean relaxed accuracy. At the same time, the comparison with Semantic RAG shows that compactness itself is not sufficient.

TimeBound is intentionally controlled. Explicit temporal metadata enables clean separation of retrieval, validity selection, update resolution, and reader-side computation, but real memory streams can contain uncertain timestamps, implicit updates, duplicate information, missing observations, and cross-tool dependencies. Future benchmarks should incorporate these sources of uncertainty.

TimeBound-RAG is likewise not a complete memory architecture. It does not learn update rules, maintain a probabilistic temporal state, or perform constraint solving. The task-level analysis shows that arithmetic, recurrence, retrospective grounding, scheduling, and long-history evidence localization remain distinct challenges. External datasets probing these boundaries are reported in Appendix F.

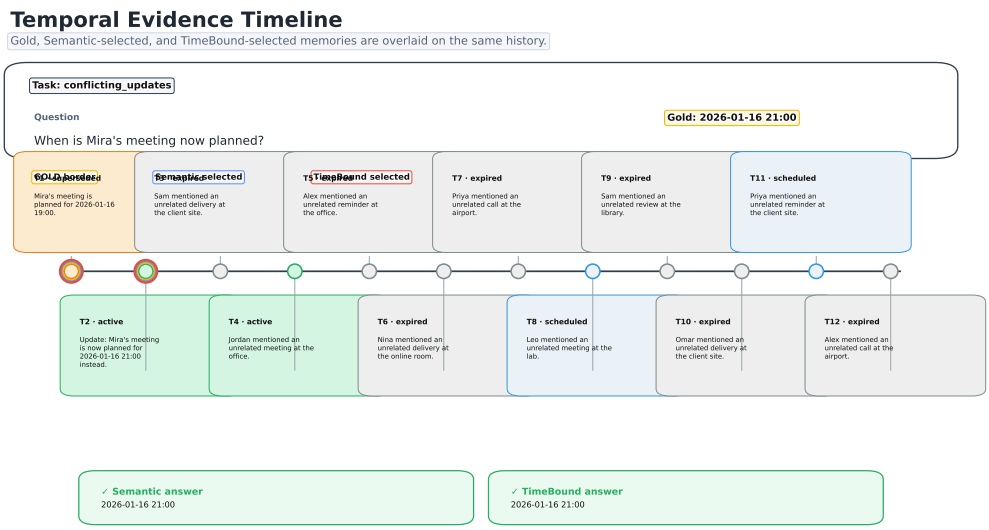


Figure 3: Temporal evidence timeline. Gold, Semantic-selected, and TimeBound-selected memories are shown over the same interaction history, illustrating the difference between semantic relatedness and query-time validity.

Finally, relaxed answer normalization is useful for equivalent dates, durations, and short paraphrases, but it can obscure some partial errors. For this reason, answer accuracy is reported together with evidence F1, invalid retrieval, contradiction rate, and context size.

7 Conclusion

We introduced TimeBound, a benchmark for studying temporal validity in long-term LLM-agent memory. The central premise is that semantic relatedness and query-time validity are distinct retrieval criteria.

Across four open-weight readers, TimeBound-RAG reaches mean relaxed accuracy of 0.719, compared with 0.475 for Semantic RAG and 0.537 for Full History. Retrieval-only analysis shows why this comparison is not captured by evidence overlap alone: Semantic RAG has slightly higher evidence F1, whereas TimeBound-RAG retrieves substantially fewer invalid memories.

The result is deliberately scoped. Validity-aware retrieval is most relevant when several plausible memories encode different operative states. Arithmetic, recurrence, retrospective reasoning, global scheduling, and long-memory localization remain separate problems. TimeBound provides a controlled setting for measuring this distinction.

AI use statement

Generative AI tools were used only for language editing, including grammar, clarity, and conciseness. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final manuscript.

Reproducibility statement

The TimeBound benchmark, generation scripts, evaluation code, aggregate statistics, paper-ready tables, qualitative figures, and completed experimental outputs are available at `https://github.com/cataug/TimeBound`.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

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*, 2021. URL https://arxiv.org/abs/2108.06314.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. doi: 10.18653/v1/2023.acl-long.637. URL https://aclanthology.org/2023.acl-long.637/.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weijia Peng, and Yang Liu. TIMEBENCH: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.66. URL https://aclanthology.org/2024.acl-long.66/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Su Yu. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hkujvAPVsg.

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 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

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

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Kai-Wei Chang, and Yiming Yang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Costas Mavromatis, Prasanna Subramanyam, Vassilis N. Ioannidis, Simon Adeshina, and George Karypis. TempoQR: Temporal question reasoning over knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/20526.

Qiang Ning, Hao Wu, and Dan Roth. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1158–1172, 2020. doi: 10.18653/v1/2020.emnlp-main.88. URL https://aclanthology.org/2020.emnlp-main.88/.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. TimeR^4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6942–6952, 2024. doi: 10.18653/v1/2024.emnlp-main.394. URL https://aclanthology.org/2024.emnlp-main.394/.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. doi: 10.18653/v1/2021.acl-long.549. URL https://aclanthology.org/2021.acl-long.549/.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. doi: 10.18653/v1/2021.acl-long.520. URL https://aclanthology.org/2021.acl-long.520/.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. doi: 10.18653/v1/2023.acl-long.828. URL https://aclanthology.org/2023.acl-long.828/.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models: A critical investigation. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2302.06706.

Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances in Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=YXogl4uQUO.

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*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ebd82705f44793b6f9ade5a669d0f0bf-Abstract-Conference.html.

Yuqing Wang and Yun Zhao. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6389–6415, 2024. doi: 10.18653/v1/2024.findings-acl.382. URL https://aclanthology.org/2024.findings-acl.382/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chen, et al. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation, 2024. URL https://arxiv.org/abs/2401.15884.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. MRAG: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3080–3118, 2025. doi: 10.18653/v1/2025.findings-emnlp.167. URL https://aclanthology.org/2025.findings-emnlp.167/.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Jie Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. NATURAL PLAN: Benchmarking LLMs on natural language planning, 2024. URL https://arxiv.org/abs/2406.04520.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, and He Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Table 6: Detailed TimeBound-Long dataset statistics. Memory-state counts are aggregated across all generated histories.

Dataset structure					
Examples	1000	Task families	8	Examples/family	125
History min.	15	History mean	29.8	History max.	46
Gold evidence mean	1.5	Validation failures	0		
Memory states					
Active	10081	Scheduled	9581	Expired	9520
Superseded	375	Delayed	125	Cancelled	125

A Benchmark Statistics and Validation

Table 6 gives the full TimeBound-Long dataset profile. Although each answer depends on only 1.5 gold turns on average, the mean history contains 29.8 turns. The histories therefore contain many temporally plausible alternatives that are not part of the gold evidence.

The benchmark generator begins from structured temporal states and applies operations including expiration, cancellation, supersession, delayed reporting, recurrence, and rescheduling. Gold answers and evidence are computed from the underlying state.

Before evaluation, generated examples are checked for invalid intervals, missing evidence identifiers, unsupported answers, broken update links, inconsistent status labels, and answer leakage. All 1000 final TimeBound-Long examples pass these checks.

B Experimental Protocol

The study contains 82 completed runs. Table 7 shows how these runs are divided between the primary benchmark, retrieval ablations, reader-level evaluation, and external diagnostics.

Table 7: Complete experimental suite.

Block	Runs	Purpose
Retrieval-only	6	Evidence quality and invalid retrieval
Retrieval ablations	18	Score, metadata, and top-k
Four-reader experiments	24	Four readers $\times$ six regimes
External diagnostics	34	Boundary-condition probes
Total	82	

The four reader models are `Mistral-7B-Instruct-v0.3`, `Qwen2.5-7B-Instruct`, `Qwen2.5-Coder-1.5B-Instruct`, and `Qwen2.5-Coder-7B-Instruct`. All receive the same answer-generation prompt.

The six context regimes are Full History, Semantic RAG, Recency RAG, Semantic+Recency RAG, TimeBound-RAG, and Oracle Evidence. Oracle Evidence is included only as a diagnostic condition because it relies on gold evidence annotations.

C Controlled Short-to-Long Diagnostic

The earlier controlled experiment in Table 8 provides an additional view of history length. Differences between retrieval strategies are relatively small in the short setting and become more visible when the interaction history grows.

Table 8: Controlled short-to-long diagnostic. The final column reports evidence F1 and contradiction rate.

Setting	Regime	Method	n	Strict Acc.	Relaxed Acc.	Evidence F1 / Contrad.
Short	Overall	Full History	830	0.542	0.923	0.852 / 0.170
Short	Overall	Semantic RAG	830	0.525	0.916	0.838 / 0.164
Short	Overall	TimeBound-RAG	830	0.529	0.914	0.849 / 0.161
Short	Overall	Oracle Evidence	830	0.530	0.930	0.897 / 0.167
Long	Overall	Full History	1000	0.303	0.614	0.529 / 0.079
Long	Overall	Semantic RAG	1000	0.193	0.541	0.259 / 0.123
Long	Overall	TimeBound-RAG	1000	0.295	0.595	0.355 / 0.041
Long	Overall	Oracle Evidence	1000	0.583	0.887	0.795 / 0.066

D Retrieval Ablations

The retrieval-budget analysis in Table 9 shows that increasing the number of retrieved memories does not monotonically increase evidence quality. Top-1 has both the highest evidence F1 and the lowest invalid retrieval rate among the evaluated TimeBound settings.

Table 9: Top-k sensitivity for TimeBound retrieval.

k	Evidence F1 ↑	Invalid retrieval ↓
1	**0.668**	**0.096**
3	0.541	0.152
5	0.413	0.161
7	0.330	0.155

Table 10 isolates the score components. Semantic and Semantic+Temporal retrieval produce higher raw evidence overlap, but also admit substantially more invalid evidence. Status-only retrieval provides the opposite extreme.

Table 10: TimeBound score-component ablation.

Variant	Evidence F1 ↑	Invalid retrieval ↓	Interpretation
Semantic+Temporal	**0.644**	0.349	High overlap, many invalid memories
Semantic only	0.611	0.359	No explicit validity filtering
Full TimeBound	0.541	0.152	Joint semantic and temporal selection
Temporal only	0.528	0.327	Limited semantic discrimination
Semantic+Status	0.439	0.050	Conservative selection
Temporal+Status	0.420	0.155	Weak semantic discrimination
Status only	0.336	**0.000**	Safe but weakly informative

Table 11 examines the same trade-off by removing individual memory fields. Removing status raises evidence F1 from 0.541 to 0.644, but invalid retrieval increases from 0.152 to 0.349.

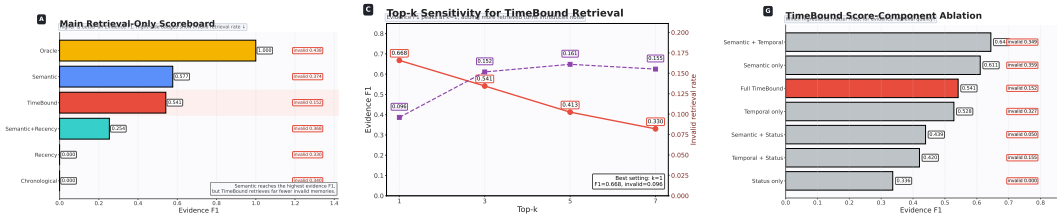
The three main retrieval ablations are visualized together in Figure 4.

Figure 5 visualizes the metadata-removal experiment from Table 11.

Finally, Figure 6 shows the larger retrieval-configuration grid. The figure illustrates that higher evidence recall can be obtained by broadening retrieval, but this need not correspond to higher downstream accuracy.

Table 11: Metadata-removal ablation.

Variant	Evidence F1 ↑	Invalid retrieval ↓
No status	**0.644**	0.349
No relation links	0.541	0.152
No observation time	0.541	0.152
No event time	0.541	0.152
Full metadata	0.541	0.152
No valid interval	0.439	**0.050**
No query time	0.439	**0.050**



(a) Retrieval-only comparison. (b) Top-k sensitivity. (c) Score-component ablation.

Figure 4: Additional retrieval-level diagnostics.

E Additional Qualitative Analysis

Figure 7 shows cases in which Semantic RAG retrieves a plausible but temporally incompatible candidate, while TimeBound-RAG selects the operative memory.

Figure 8 shows that access to the entire history does not guarantee successful update resolution: stale information remains visible even when a newer state is also present.

Figure 9 separates variation among readers from residual error types that are not directly resolved by temporal retrieval.

F External Diagnostic Probes

The external datasets test operations adjacent to validity-changing memory. They are not treated as a common leaderboard because their inputs and bottlenecks differ substantially. Table 12 summarizes their scale and intended role.

Table 12: External diagnostic datasets.

Dataset	n	Avg. turns	Primary bottleneck
TempReason	10000	1.0	Date arithmetic
Complex-TR	329	4.0	Compact interval reasoning
TCP	600	21.0	Global scheduling constraints
LoCoMo	500	46.2	Long-memory localization

Figure 10 shows the local-reader diagnostic results. Performance remains low and heterogeneous, consistent with these datasets requiring mechanisms that differ from validity filtering among competing memory states.

The strongest observed local-reader relaxed accuracies are 0.136 on TempReason, approximately 0.033 on Complex-TR, approximately 0.013 on TCP, and 0.044 on LoCoMo. These values support a limited interpretation: validity-aware retrieval targets state-selection errors, whereas arithmetic, interval reasoning, global scheduling, and long-memory localization require additional mechanisms.

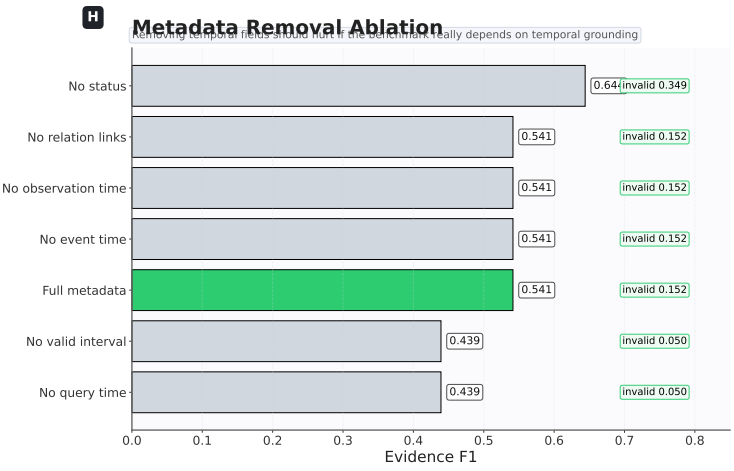


Figure 5: Effect of removing individual temporal metadata fields.

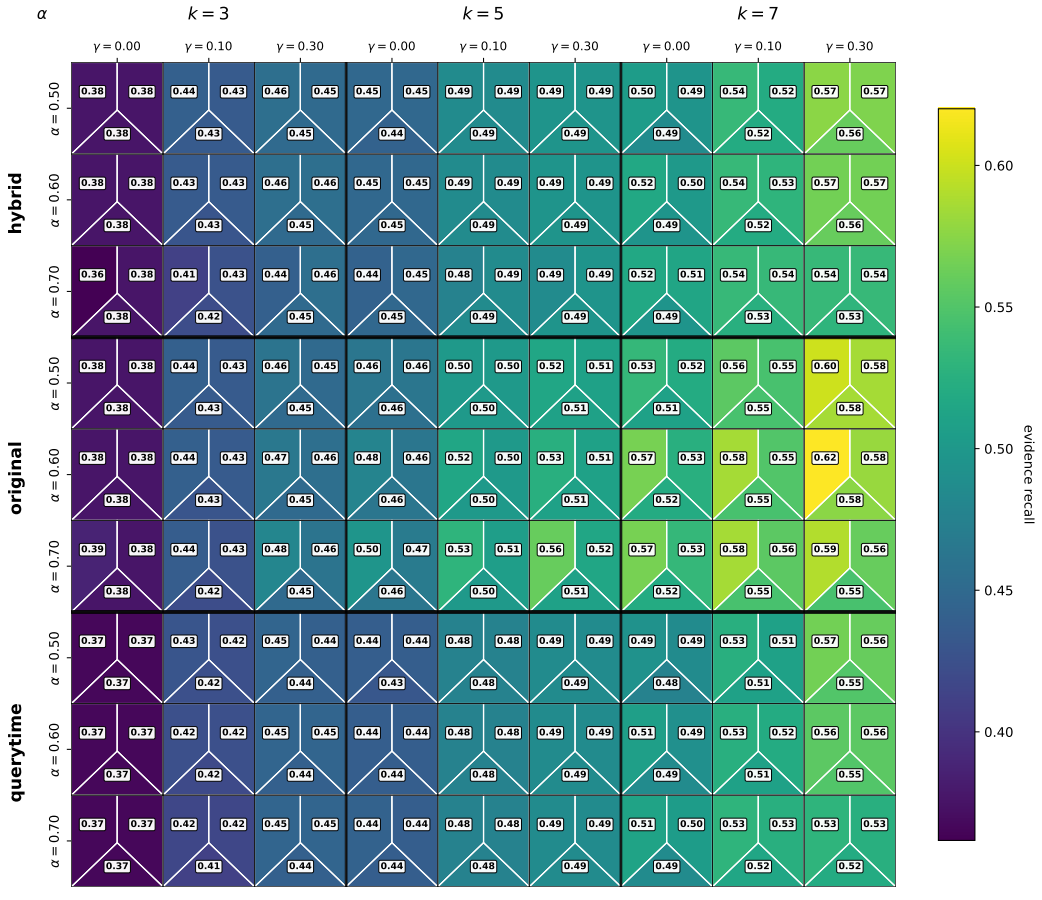


Figure 6: Grid search over retrieval policy, semantic weight, top-k, and status penalty. Cells report evidence recall.

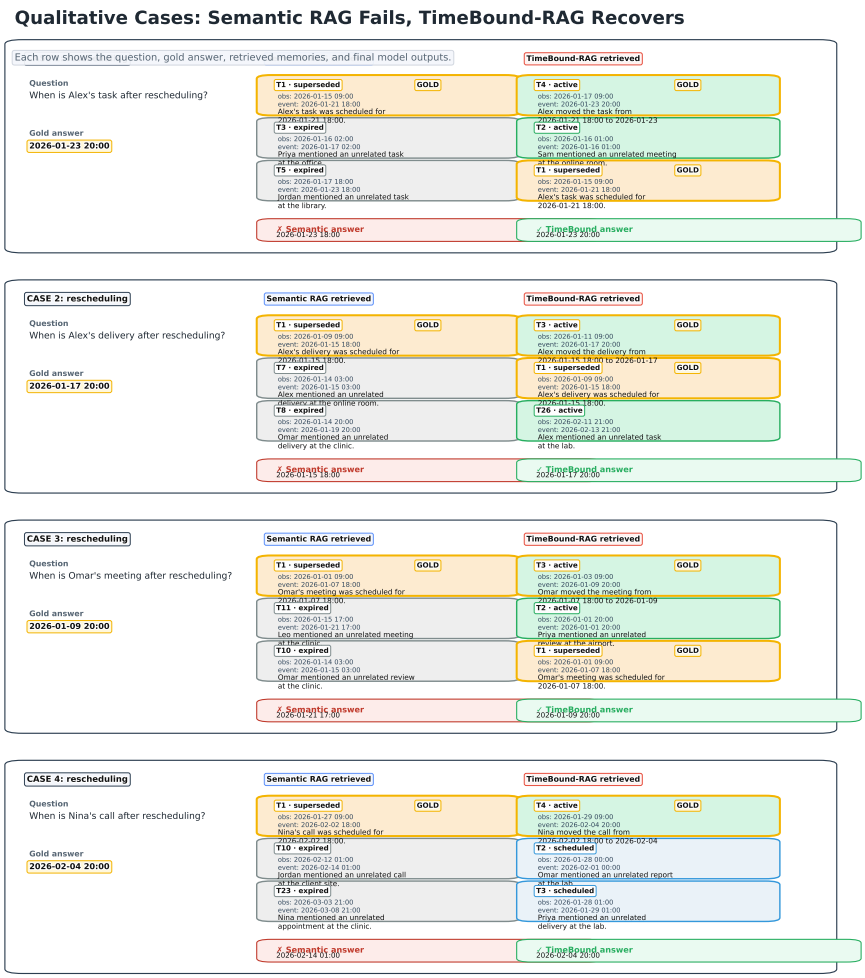


Figure 7: Examples comparing Semantic RAG and TimeBound-RAG on validity-changing memories.

G Extended Quantitative Results

Figure 11 collects the quantitative plots that complement Figure 2 in the main paper. They provide the retrieval scoreboard, top-k sensitivity, full reader-regime ranking, and metadata/component ablations.

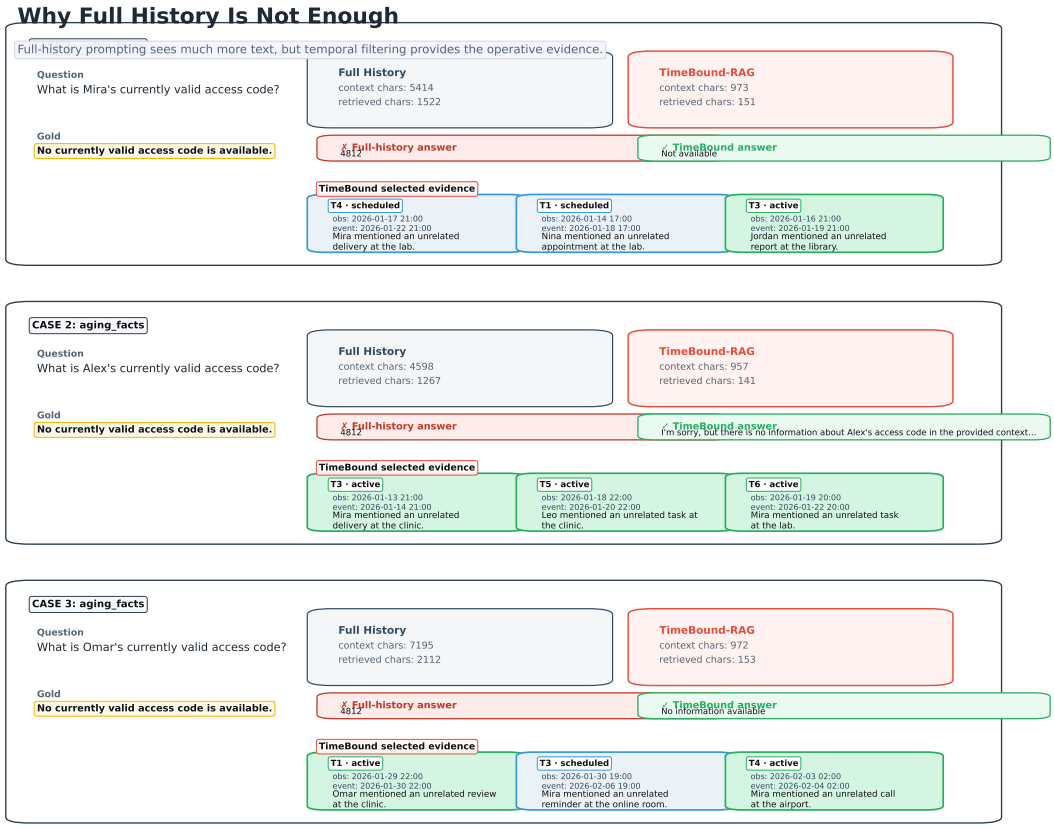
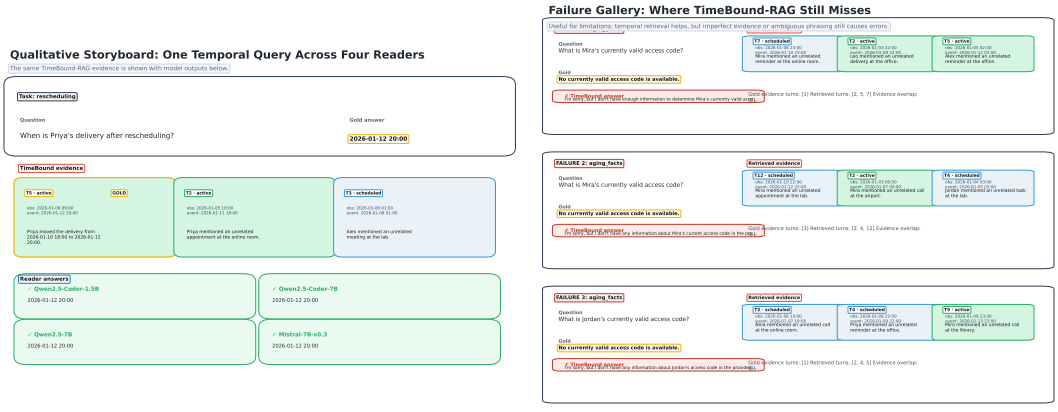


Figure 8: Examples comparing Full History and TimeBound-RAG. Full context can contain both the operative state and semantically plausible stale alternatives.



(a) One case across four readers.

(b) Residual failure modes.

Figure 9: Additional qualitative diagnostics. Reader-side arithmetic, recurrence, retrospective grounding, and localization errors remain after retrieval.

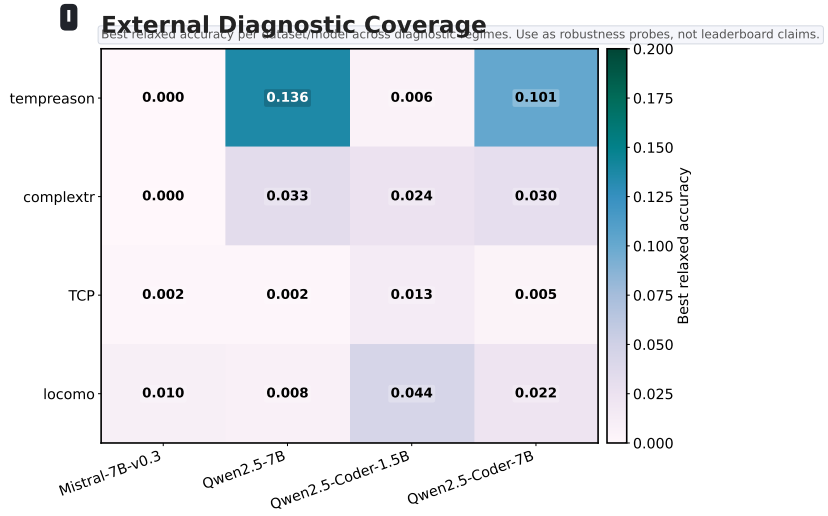
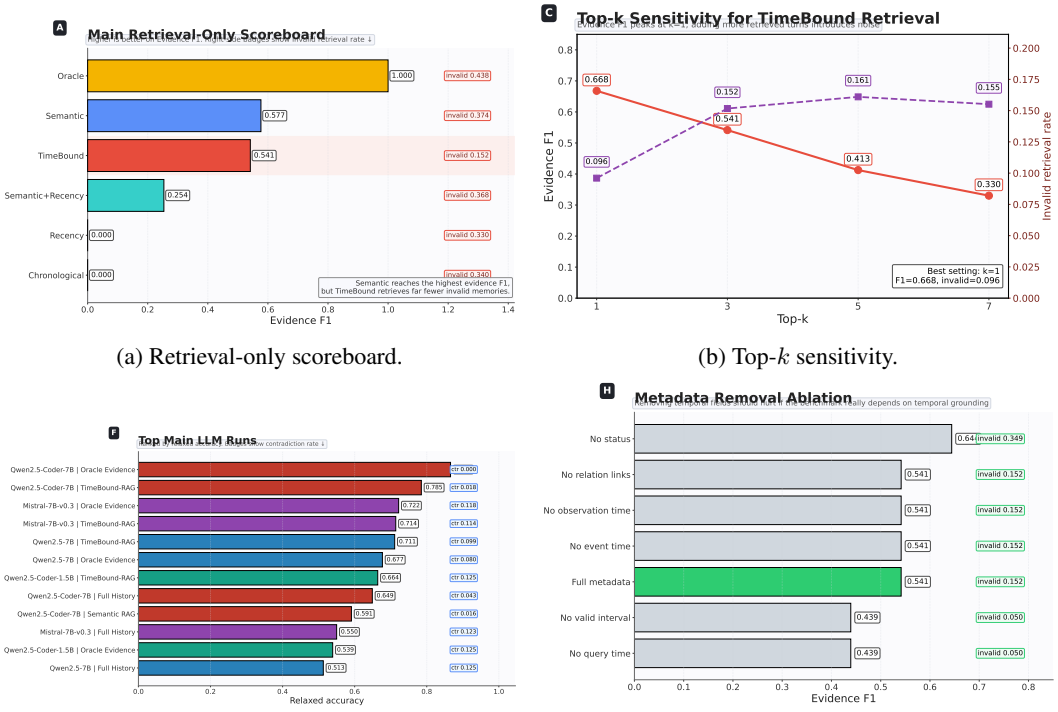


Figure 10: External diagnostic results across local readers. These datasets are used as boundary probes rather than leaderboard comparisons.



(a) Retrieval-only scoreboard.

(b) Top-k sensitivity.

(c) Reader–regime configurations.

(d) Metadata ablation.

Figure 11: Extended quantitative analysis of TimeBound-Long.