Write Policy Determines Source Retention in External FIFO Memory: A Delayed-Copy Diagnostic

Xiangyang Liu **ChunXia Yang**
School of Artificial Intelligence (School of Future Technology)
Nanjing University of Information Science and Technology, Nanjing, China
202512492781@nuist.edu.cn chunxia@nuist.edu.cn

Abstract

External hidden-state memories can extend Transformers beyond local attention windows, but the write policy deciding which states enter finite memory is often treated as an implementation detail. We study this factor with a controlled delayed-copy diagnostic using a windowed Transformer augmented with an external FIFO hidden-state memory. Under forced masking, query-to-source attention is blocked, removing this shortcut and making source retention directly testable. We define a capacity boundary as the minimum memory size needed to reach a fixed copy-accuracy threshold. We find that write policy changes this boundary predictably under constrained FIFO memory: an idealized source-only policy, which assumes that relevant states are identifiable at write time, yields an approximately delay-invariant boundary near the source length, whereas an unfiltered prefix-all policy induces a delay-dependent boundary that grows with written non-source content. Intermediate partial-prefix write policies produce intermediate boundaries, while source-pinned controls and memory diagnostics show that prefix-all degradation is dominated by FIFO eviction of source states rather than retrieval interference from retained non-source states alone. A symbolic key–value retrieval diagnostic shows the same qualitative trend. Within this controlled setting, our results highlight write policy as a central variable in evaluating finite FIFO external-memory Transformers.

1 Introduction

Transformer-based architectures have become the dominant paradigm for sequence modeling (Vaswani et al., 2017), but their quadratic attention cost limits the length of directly accessible context. A broad line of work extends effective context length using recurrence, sparse attention, external memory, or cached hidden states (Dai et al., 2019; Zaheer et al., 2020; Wu et al., 2022a; Zhang et al., 2024; Martins et al., 2022). While these systems often emphasize how information is retrieved from memory, the write side is frequently treated as an implementation detail.

In finite-capacity memory systems, write policy determines which states enter memory and which states survive eviction. Writing additional non-target content can affect performance through two mechanisms: retained distractors may interfere with retrieval, and later writes may physically evict task-relevant source states. Long-context performance can depend strongly on the position of relevant evidence (Liu et al., 2024). In a finite external buffer, poor performance may additionally arise from interference during retrieval or physical eviction of relevant states.

We study this issue with controlled diagnostics built around an external FIFO hidden-state memory under local attention. Our main setting is a delayed-copy task with inputs of the form

$$[\texttt{source}][\texttt{sep}][\texttt{noise}][\texttt{query}].$$

We block all query-to-source attention as a diagnostic intervention, isolating memory-dependent behavior from direct query-to-source copying.

We compare an idealized *source-only* write policy, an unfiltered *prefix-all* policy, *intermediate partial-prefix* write policies, and a *source-pinned noise-FIFO* control that explicitly retains source states while admitting distractors. These idealized diagnostic controls make relevance or write schedules available by construction: source-only assumes that relevance is known at write time, partial-prefix follows a predetermined schedule, and none learns selective writing from data. Across these settings, we measure accuracy-based capacity boundaries and record memory-entry diagnostics, including source survival, attention mass, and error types. We further include a symbolic key–value retrieval diagnostic to test whether the same write-side effect appears beyond direct delayed copying.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 17748–17766
October 24-29, 2026 ©2026 Association for Computational Linguistics

Our results show that, under constrained FIFO memory, write-side contamination can dominate the measured capacity requirement by determining whether task-relevant source states survive eviction. Source-only writes yield an approximately delay-invariant boundary near the source length, while prefix-all writes induce a delay-dependent boundary that grows with the amount of non-source content written into memory. Intermediate partial-prefix policies produce intermediate boundaries, and the source-pinned control shows little degradation when source states are explicitly retained. These findings show that write policy must be controlled when interpreting capacity measurements for finite FIFO external memories.

Contributions. We make four contributions:

- We introduce delayed-copy and symbolic key–value retrieval diagnostics with forced masking to remove direct query-to-source attention under local attention.
- We define operational memory-capacity boundaries $m^*(d)$ and $m^*(d, p)$ for measuring the minimum FIFO capacity required to reach a fixed accuracy threshold.
- We complement the two endpoint policies with intermediate partial-prefix write policies, showing that the capacity boundary changes continuously with non-source write strength.
- We use the source-pinned control and memory-level diagnostics to separate FIFO source eviction from retrieval interference caused by retained distractors.

2 Diagnostic Setup

Delayed-copy task and forced masking. Each delayed-copy example has the form

$$\texttt{inp} = [\texttt{source}][\texttt{sep}][\texttt{noise}][\texttt{query}],$$
$$\texttt{tgt} = \texttt{source}.$$

The source and query segments have length L, and the noise segment has length d. The query segment contains only placeholder tokens, so the model must reproduce the source tokens at the final L query positions.

We use local windowed self-attention. In the natural setting, query positions follow the standard local attention mask. In the forced setting, we additionally block all query-to-source attention edges while leaving the rest of the local attention computation unchanged. This intervention tests whether source information is written to memory, retained under FIFO eviction, and retrieved without direct query-to-source copying.

External FIFO hidden-state memory. We compare a windowed Transformer encoder baseline with memory-augmented variants. Token embeddings are combined with learned absolute positional embeddings at the encoder input. The encoder computes final-layer hidden states, selected positions are written into a finite FIFO memory according to a write policy, and the final L query representations perform a single content-based memory-attention read after the encoder. No additional relative-position term is applied during this memory read. The memory is reset for each training example and does not persist across mini-batches. Stored hidden states are used as both keys and values; when capacity is exceeded, the oldest entries are evicted. We evaluate both a naive additive readout and a gated readout. This design uses a single inspectable post-encoder external-memory pathway and is distinct from a layer-wise autoregressive KV cache.

Write policies. Our main manipulation is which hidden states are written into memory. The *source-only* policy writes only source states and serves as an oracle-like diagnostic control, since it assumes that relevant states are identifiable at write time. The *prefix-all* policy writes the full prefix before the query,

$$[\texttt{source}][\texttt{sep}][\texttt{noise}],$$

so later separator and noise states can evict earlier source states under FIFO capacity limits.

In addition to the two endpoint write policies, we introduce an *intermediate partial-prefix write policy*. This policy writes all source states, the separator, and only the first $\mathrm{round}(pd)$ noise states, where $p \in [0, 1]$ controls the non-source write strength. Thus, $p = 0$ corresponds to source-plus-separator writes and $p = 1$ recovers prefix-all writes. Finally, the *source-pinned noise-FIFO* control, abbreviated as *source-pinned*, reserves slots for all source states and applies FIFO eviction only within a separate noise buffer. This diagnostic removes source eviction by construction while still varying the amount of non-source content available to the readout.

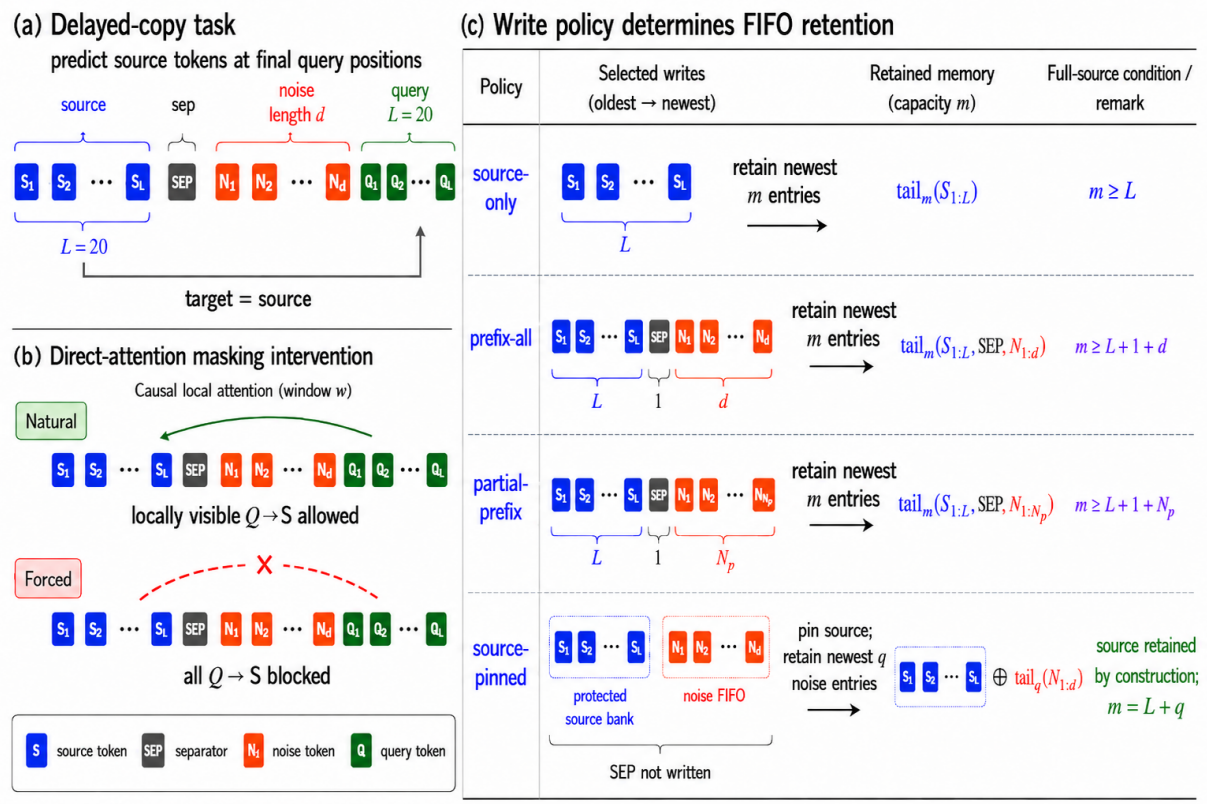


Figure 1: Diagnostic setup. (a) Delayed-copy input structure, where S, N, and Q denote source, noise, and query tokens, respectively; the model reproduces the source sequence at the final query positions. (b) Direct-attention masking intervention. Under the natural causal-local mask, locally visible query-to-source entries remain allowed. The forced mask blocks all query-to-source entries while leaving the other locally permitted entries unchanged. (c) Policy-specific selected writes and retained memory states. For a sequence X, $\text{tail}_k(X)$ denotes its newest $\min(k, |X|)$ entries. Source-only, prefix-all, and partial-prefix use a FIFO memory of capacity m, with $N_p = \text{round}(pd)$ for partial-prefix. Source-pinned instead reserves L protected source slots and uses a separate noise FIFO of capacity q, giving total capacity $m = L + q$. The rightmost column reports the condition for retaining all source states, or the corresponding construction guarantee.

Diagnostics and symbolic retrieval variant. For analysis, we record the segment origin of each memory entry, allowing us to measure source survival, attention mass over source and non-source entries, and prediction-error categories. These labels are used only for diagnostics and are never provided to the model.

We also evaluate a lightweight symbolic key–value retrieval variant. Each input contains a target key–value record, intervening distractor records whose keys differ from the target key, and a later query containing the target key but not its value. The model reconstructs the complete target record at the final query positions, and evaluation measures token accuracy over all positions in that record. We compare source-only, prefix-all, and source-pinned configurations in this structured association task.

3 Experimental Setup

Tasks and settings. We evaluate two controlled diagnostic tasks introduced in Section 2: delayed copying and symbolic key–value retrieval. The delayed-copy task uses source/query length $L = 20$ and delay lengths $d \in \{10, 20, 40, 80\}$, unless otherwise stated. The symbolic key–value retrieval variant uses the same external FIFO memory mechanism and compares source-only, prefix-all, and source-pinned configurations. Given a query containing a target key but not its value, the model reconstructs the complete target key–value record at the final query positions. This variant evaluates the write-policy capacity boundary in a structured retrieval setting. Unless otherwise stated, forced-setting experiments use local attention window size $w = 8$.

Training. All models use a 2-layer Transformer encoder with 4 attention heads and model dimension 64. We train with Adam (Kingma and Ba, 2015), learning rate 10^{-3}, and batch size 16. All delayed-copy experiments (Experiments A–C), including the main boundary experiments, memory diagnostics, and intermediate partial-prefix experiments, are trained for 3001 optimization steps. The symbolic key–value retrieval experiment (Experiment D) uses a fixed budget of 9001 optimization steps for every policy, delay, capacity, readout, and seed. The longer budget was selected after preliminary 3001-step runs indicated slower convergence on the symbolic task and is applied uniformly to all reported Experiment D runs. At every optimization step, we generate a fresh synthetic minibatch online; no fixed training dataset is reused. Results are averaged over seven random seeds $\{0, 1, 2, 3, 4, 5, 6\}$, unless otherwise stated. Memory-augmented results are reported for both naive additive readout and gated readout variants. For the intermediate partial-prefix experiment, the gated readout is used as the main diagnostic, while the naive additive readout is reported as a readout-stability analysis.

Evaluation. We measure token-level accuracy on the final query positions and report `acc_mean_tail`, the mean query-token accuracy over the final 500 training steps. Because each of these steps uses a freshly generated minibatch, this average reduces sensitivity to any single minibatch and summarizes end-of-training behavior. It is an operational training-tail measure, not a held-out generalization metric. All capacity boundaries are computed exclusively from `acc_mean_tail`. For the post-training source-survival diagnostic in Figure 3, we additionally report `eval_acc`, defined for each seed and configuration as mean query-token accuracy over 16 fresh same-distribution minibatches evaluated with a separate deterministic random stream. This diagnostic accuracy is not used to compute m^*. In the forced setting, success is defined as reaching a fixed accuracy threshold

$$\overline{\mathrm{Acc}} \geq \tau, \quad \tau = 0.95,$$

where $\overline{\mathrm{Acc}}$ denotes the seed-averaged training-tail accuracy for the corresponding task, delay, memory capacity, readout variant, and write-policy configuration.

Capacity boundary. For a write policy π and delay length d, we define the memory-capacity boundary as

$$m^*_\pi(d) = \min\left\{m \in \mathcal{M}_\pi : \overline{\mathrm{Acc}}_\pi(d, m) \geq \tau\right\}.$$

Here, $\mathcal{M}_\pi$ is the policy-specific capacity grid. If no scanned capacity reaches the threshold, we report a censored boundary:

$$m^*_\pi(d) > \max(\mathcal{M}_\pi).$$

For the intermediate partial-prefix policy family, where p controls the non-source write strength, we analogously define

$$m^*(d, p) = \min\left\{m \in \mathcal{M} : \overline{\mathrm{Acc}}(d, p, m) \geq \tau\right\}.$$

We use finer capacity points around the expected FIFO retention thresholds; full forced-setting grids and run configurations are provided in Appendix C.

Memory diagnostics. To analyze the mechanism behind the observed boundaries, we record source survival, memory-attention mass over source and non-source entries, and prediction-error categories after training. These diagnostics are used only for analysis and are never provided to the model. They allow us to distinguish physical FIFO eviction of source states from retrieval interference caused by retained non-source entries.

Experiment groups. We organize the experiments into four groups. First, we run natural-setting sanity checks and forced-setting capacity-boundary experiments for source-only, prefix-all, and source-pinned configurations. Second, we perform memory-level diagnostics to relate copy accuracy to source survival, attention mass, and error types. Third, we evaluate intermediate partial-prefix writes with

$$p \in \{0, 0.25, 0.5, 0.75, 1.0\}$$

and delays $d \in \{20, 40, 80\}$, testing whether the boundary changes smoothly with the amount of non-source content written into memory. Finally, we evaluate the qualitative write-policy boundary on symbolic key–value retrieval.

Complete data-generation, masking, external-memory and readout, write-policy, training, capacity-grid, evaluation, and diagnostic specifications are provided in Appendix C.

4 Results

Write policy changes the forced-memory capacity boundary. We first evaluate delayed-copy performance under forced masking. Figure 2 reports

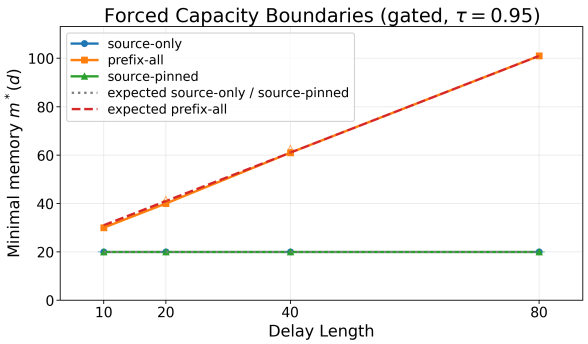


Figure 2: Forced-setting capacity boundaries under the gated readout. Source-only writes remain approximately delay-invariant near the source length; the source-pinned control does likewise across the tested admitted-noise budgets $q \leq 80$, whereas prefix-all writes require larger memory as the delay grows. Filled markers and solid lines report aggregate boundaries obtained by thresholding the seven-seed mean `acc_mean_tail`. For finite aggregate boundaries with at least two finite seed-wise boundaries, slightly offset thin capped bars span the minimum–maximum range of the observed finite (uncensored) seed-wise boundaries. Hollow triangles indicate that one or more seed-wise boundaries are right-censored. The displayed ranges are descriptive, conditional on the finite seed-wise boundaries, and are not confidence intervals; horizontal offsets are visual only. Exact conditional means, sample standard deviations, ranges, censoring counts, and scan maxima are provided in Appendix A.3. Naive-readout results and threshold-stability analyses are provided in Appendix A.

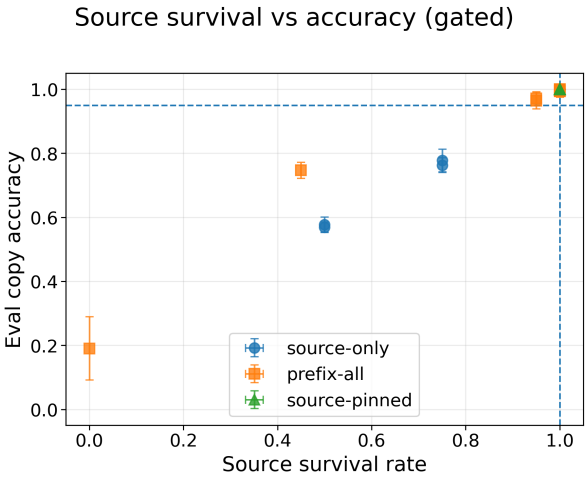


Figure 3: Source survival versus post-training evaluation accuracy under the gated readout. Each point plots a seven-seed mean, with each seed-level evaluation averaged over 16 fresh same-distribution minibatches. Horizontal and vertical error bars show one standard deviation across seeds in source survival and `eval_acc`, respectively. Accuracy rises with source survival across memory configurations and write policies.

the accuracy-based capacity boundary $m^*(d)$ under the gated readout. The source-only policy reaches the success threshold once the memory capacity is close to the source length $L = 20$, and its boundary remains approximately delay-invariant. In contrast, prefix-all writes require substantially larger memory as the delay length increases, because separator and noise states are written after the source and can evict earlier source states under FIFO replacement. Across the tested admitted-noise budgets $q \leq 80$, the source-pinned control remains near the source-length boundary and shows little degradation when source states are explicitly retained.

Accuracy tracks source survival. Two mechanisms can shift the boundary: interference from retained non-source entries and FIFO eviction of source entries. To separate these effects, we record the segment origin of each memory entry after training. Figure 3 shows that post-training evaluation accuracy (`eval_acc`) closely tracks the fraction of source states that survive in memory. Configurations with low source survival have low post-training evaluation accuracy, while configurations with full or near-full source survival have high evaluation accuracy. This pattern supports the interpretation that, in the constrained FIFO setting, prefix-all degradation is dominated by physical eviction of task-relevant source states rather than by retrieval interference alone.

Figure 4 provides a complementary memory-level diagnostic for prefix-all writes. Below the retention boundary, source states are partially or fully missing from memory, so the model cannot reliably attend to them. Around the boundary, attention mass increasingly favors source entries over non-source entries. Together, the two diagnostics associate the boundary transition with changes in source retention; Section 5 develops the eviction interpretation.

Intermediate partial-prefix policies produce intermediate boundaries. Source-only and prefix-all represent two endpoints of a write-selectivity spectrum. We therefore evaluate an intermediate partial-prefix policy family that always writes the source and separator, but writes only the first $\mathrm{round}(pd)$ noise states, where $p \in [0, 1]$ controls the amount of non-source content admitted into memory. The corresponding FIFO full-source retention threshold is

$$L + 1 + \mathrm{round}(pd).$$

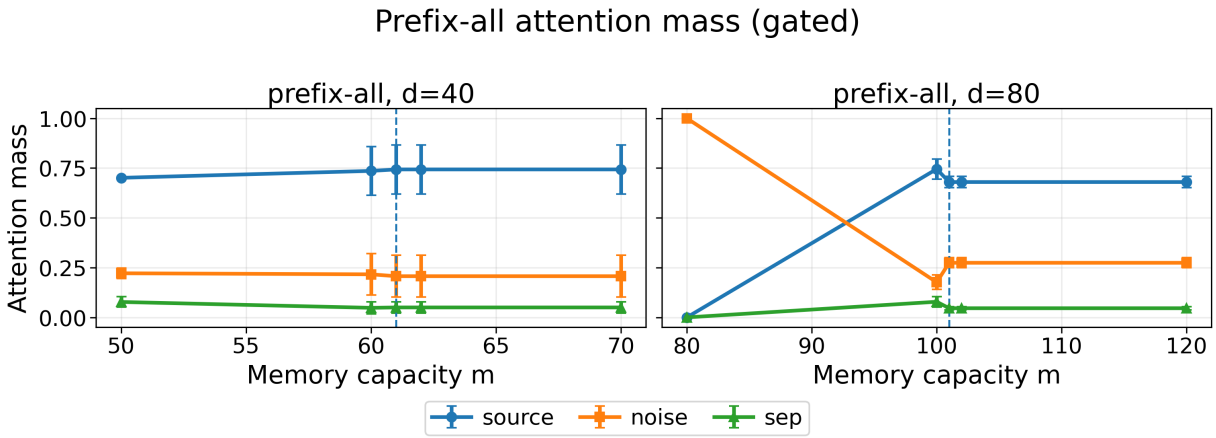


Figure 4: Prefix-all memory attention mass under the gated readout. Curves show seven-seed means, and vertical error bars show one standard deviation across seeds. Around the retention boundary, attention increasingly favors source entries over non-source entries. Additional diagnostics, including naive-readout variants, source-pinned attention, and error-type rates, are provided in Appendix A.

As shown in Figure 5a, the measured boundary $m^*(d, p)$ increases with the amount of written non-source content and closely follows this deterministic FIFO prediction. Intermediate policies therefore show that the apparent memory-capacity requirement changes systematically with write strength. The naive additive readout follows the same overall trend but is noisier, so we report it as a readout-stability analysis in Appendix A.

Symbolic key–value retrieval shows the same qualitative effect. Finally, we test whether the same write-policy effect appears beyond direct delayed copying. In the symbolic key–value retrieval variant, a later query provides the target key but not its value, and the model reconstructs the complete target key–value record after intervening distractor records. Figure 5b shows the same qualitative pattern. Source-only writes and the source-pinned control remain approximately delay-invariant, whereas prefix-all writes require larger memory as more distractor content is written into the FIFO buffer. The symbolic task yields noisier and sometimes more conservative boundaries than delayed copying, especially under the naive additive readout.

5 Analysis

In the forced setting, the measured capacity boundary is governed largely by the write policy: which states enter the finite FIFO memory and whether task-relevant source states survive eviction. This section relates the empirical boundaries to FIFO source retention and uses the diagnostics below to distinguish physical eviction from retrieval interference.

FIFO retention explains the boundary. Under source-only writes, the memory stores only the source segment of length L, so retaining the full source requires $m \geq L$, independent of delay. This explains the approximately delay-invariant source-only boundary. By contrast, prefix-all writes store the entire prefix,

$$[\texttt{source}][\texttt{sep}][\texttt{noise}],$$

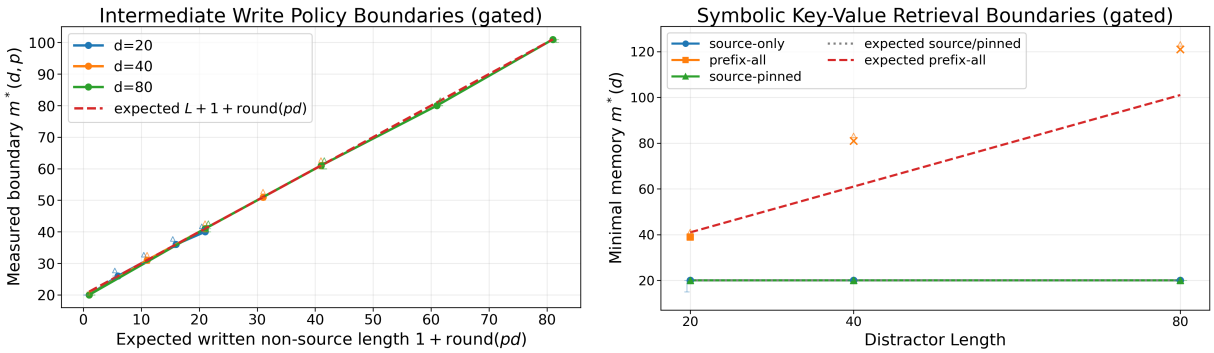
so retaining the full source requires capacity large enough to keep the source after the separator and d noise states have also been written:

$$m \geq L + 1 + d.$$

Thus, prefix-all writes create a delay-dependent boundary because increasing d writes more later non-source states that can evict earlier source states. The source-pinned control removes this eviction pathway by retaining all source states by construction; its approximately delay-invariant behavior supports the interpretation that prefix-all degradation is primarily caused by source eviction. The full deterministic FIFO calculation is given in Appendix B.

Intermediate partial-prefix policies show a continuous effect. The intermediate partial-prefix experiment gives a continuous version of the same mechanism. For write strength p, the policy writes the source, the separator, and the first $\mathrm{round}(pd)$ noise states. The expected full-source retention condition is therefore

$$m \geq L + 1 + \mathrm{round}(pd).$$



(a) Intermediate partial-prefix writes

(b) Symbolic key–value retrieval

Figure 5: Additional boundary analyses under the gated readout. (a) Intermediate partial-prefix writes produce boundaries that track the expected FIFO threshold $L + 1 + \text{round}(pd)$. (b) The symbolic key–value retrieval diagnostic shows the same qualitative write-policy effect: source-preserving policies remain approximately delay-invariant, while prefix-all writes require larger memory as distractor content grows. Filled markers and solid lines report finite aggregate boundaries obtained by thresholding the seven-seed mean `acc_mean_tail`. For finite aggregate boundaries with at least two finite seed-wise boundaries, slightly offset thin capped bars span the minimum–maximum range of the observed finite (uncensored) seed-wise boundaries. Hollow triangles indicate that one or more seed-wise boundaries are right-censored, whereas crosses plotted at the maximum scanned capacity indicate aggregate right-censoring. The displayed ranges are descriptive, conditional on the finite seed-wise boundaries, and are not confidence intervals; horizontal offsets are visual only. Exact conditional means, sample standard deviations, ranges, censoring counts, and scan maxima are provided in Appendix A.3. Naive-readout variants are provided in Appendix A.

The observed boundaries closely follow this continuous prediction: the measured capacity boundary increases systematically with the amount of non-source content written into memory.

Diagnostics separate eviction from retrieval interference. The memory-level diagnostics further support the retention-based interpretation. Accuracy rises when source survival approaches one, while configurations with low source survival fall below the success threshold. Attention-mass and error-type diagnostics show the same qualitative pattern: prefix-all failures occur mainly when source entries are absent or only partially retained. Across the tested admitted-noise budgets, the source-pinned control shows little degradation when source states are explicitly retained. Together, these diagnostics identify write-induced source eviction as the primary explanation for prefix-all degradation under finite FIFO capacity.

Readout stability and symbolic retrieval. Both naive and gated readouts exhibit the same retention-driven trend in the main boundary experiments. The gated readout gives cleaner and more stable boundaries, while the naive additive readout is less stable under stricter thresholds and intermediate write strengths. The shared trend across readouts reflects a common retention constraint: once task-relevant source states have been evicted, the readout has no direct source information to recover. Threshold- and seed-stability analyses are reported in Appendix A.

The symbolic key–value retrieval variant shows the same qualitative ordering. Source-only writes and the source-pinned control keep the boundary approximately delay-invariant, while prefix-all writes require larger memory as more distractor content is written. Because this task requires resolving key–value associations, its boundaries are noisier and can exceed the deterministic full-source retention threshold. Thus, FIFO source survival is a necessary condition for reliable retrieval in this setting, but not a complete guarantee of success.

Accordingly, $m^*(d)$ and $m^*(d,p)$ quantify operational capacity under the specified write policy, FIFO rule, success threshold, readout, and training protocol. Under this protocol, non-target writes can raise the capacity requirement by occupying finite slots and evicting task-relevant states, in addition to creating retrieval noise.

6 Related Work

Long-context and memory-augmented Transformers. Many Transformer variants extend usable context through recurrence, sparse attention, external memory, or cached hidden states.

Transformer-XL introduces segment-level recurrence (Dai et al., 2019), Compressive Transformers compress older states into longer-term memory (Rae et al., 2020), and sparse-attention models such as Longformer and BigBird reduce attention cost using structured attention patterns (Beltagy et al., 2020; Zaheer et al., 2020). Other memory-augmented models, including Memorizing Transformers, Memformer, and Cached Transformers, store or retrieve intermediate representations for long-range modeling (Wu et al., 2022b,a; Zhang et al., 2024). These works mainly focus on extending context length or improving memory architectures, whereas our work examines how write policy and write strength affect the operational capacity boundary of a finite FIFO hidden-state memory.

KV-cache management and selective retention. Recent long-context inference methods study how to select, retain, or evict key–value cache entries in pretrained language models. H2O retains heavy-hitter tokens (Zhang et al., 2023), StreamingLLM identifies attention sinks for stable streaming inference (Xiao et al., 2024), Scissorhands exploits persistence in token importance (Liu et al., 2023), and SnapKV performs prompt-aware KV-cache selection before generation (Li et al., 2024). These methods show that retention decisions can strongly affect long-context behavior. They typically operate on layer-wise autoregressive KV caches in pretrained decoder-only language models at inference time. Our trained windowed encoder instead writes selected final-layer hidden states directly as both keys and values to an external FIFO buffer and performs a single post-encoder content-based read. This abstraction exposes write-policy effects and separates FIFO source eviction from retrieval interference while omitting the full layer-wise autoregressive dynamics of practical LLM KV caches.

Diagnostics for memory behavior and context utilization. Empirical studies show that larger context windows do not guarantee reliable use of relevant information, especially when evidence appears in difficult positions within the context (Liu et al., 2024). Synthetic algorithmic tasks have also been widely used to probe learned memory, storage, and retrieval over sequences (Graves et al., 2014, 2016; Joulin and Mikolov, 2015). Our diagnostics follow this tradition but focus on a specific mechanism: how writing non-source content into a finite FIFO memory changes source survival and the resulting capacity boundary. By combining forced masking, intermediate partial-prefix write strengths, the source-pinned control, and memory-level diagnostics, we separate temporal delay, FIFO source eviction, and retrieval interference in a controlled setting.

7 Conclusion

We presented controlled diagnostics for studying how predefined write policies affect an external FIFO hidden-state memory under local attention. Using forced masking and operational capacity boundaries $m^*(d)$ and $m^*(d, p)$, we found that source-only writes produce an approximately delay-invariant boundary near the source length, while the prefix-all boundary grows with the amount of written non-source content. Intermediate partial-prefix policies trace the corresponding continuum as write strength increases.

The source-pinned control and memory-level diagnostics associate the accuracy transition with source retention: prefix-all failures concentrate where FIFO writes evict task-relevant source states, while retaining those states yields stable performance across the tested configurations. The symbolic key–value variant shows the same qualitative policy ordering. Together, these results identify write policy as a central variable in the observed capacity boundary under the tested finite-FIFO protocol.

Future work should test learned selective writing under unknown relevance and interleaved evidence, together with retention-aware eviction, stronger read mechanisms, pretrained language models, and realistic long-context or layer-wise KV-cache settings.

Limitations

Our evaluation consists of controlled synthetic delayed-copy and symbolic key–value retrieval diagnostics; the latter tests structured associative retrieval beyond direct copying. It does not cover downstream long-context NLP benchmarks, realistic multi-document retrieval or reasoning, or pretrained decoder-only language models. The results provide mechanistic evidence for this memory setting rather than performance evidence for practical long-context LLM systems.

Our simplified memory stores selected final-encoder hidden states directly as both keys and values and performs a single post-encoder content-based read. Decoder-only LLM KV caches instead

use layer-specific projected keys and values with autoregressive updates and reads across layers and decoding steps. This architectural difference makes source retention and FIFO eviction directly measurable in our setting and limits quantitative transfer to practical LLM KV-cache systems. Raw final-encoder states are one controlled representation choice. Projected key–value states, compressed representations, and layer-wise memories remain important future comparisons.

The models use learned absolute positional embeddings at the encoder input, with no additional relative-position term during the memory read, and we evaluate delay lengths only up to $d = 80$. The positional results therefore cover this tested range; longer distances, positional extrapolation, and alternative positional encodings remain untested.

The forced mask removes direct query-to-source attention while preserving possible multi-layer routes through intervening hidden states. With two encoder layers and $w = 8$, the maximum backward reach through local attention is 14 positions; the shortest setting $d = 10$ can therefore admit limited indirect propagation, whereas $d \geq 20$ exceeds this reach. Accordingly, forced masking removes the direct shortcut but does not establish the absence of all non-memory information flow.

Our boundary measurements compare only a naive additive readout and a simple gated readout. More expressive mechanisms, including multi-layer memory reads, learned retrieval controllers, top-k retrieval, and stronger cross-attention modules, remain outside this evaluation. The conclusions therefore concern these two readouts and do not establish a universal ranking between write and read mechanisms. In this controlled FIFO setting, the evaluated readouts cannot recover task-relevant information when it is no longer represented anywhere in the available memory. The capacity boundaries $m^*(d)$ and $m^*(d, p)$ are also protocol-dependent because they use discrete memory scans, fixed training budgets, and a fixed success threshold.

The write policies are idealized diagnostic controls: source-only assumes that task-relevant states are identifiable at write time, partial-prefix follows a deterministic schedule, and source-pinned prevents source eviction by construction. The source-pinned evidence is limited to the tested admitted-noise budgets $q \leq 80$; behavior under arbitrarily large retained-distractor loads remains untested. These policies neither learn relevance nor cover settings where relevant information is interleaved with distractors or becomes identifiable only after later context or the query. They diagnose write-induced eviction without providing a deployable selective-writing method.

This study uses only synthetic sequences and involves no human data or deployment; its direct societal impact is therefore limited.

Acknowledgments

ChatGPT was used by the first author as an assistive tool for language editing, code-refactoring assistance, figure-presentation suggestions, and consistency and formatting checks. The first author designed the study, ran all experiments, and prepared the final figures. All AI-assisted changes were reviewed, tested where applicable, and revised by the authors; both authors reviewed and approved the final code and manuscript. The authors take full responsibility for the final manuscript, code, results, figures, and technical content.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing machines. *CoRR*, abs/1410.5401.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Armand Joulin and Tomas Mikolov. 2015. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems*, volume 28, pages 190–198. Curran Associates, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 22947–22970. Curran Associates, Inc.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, pages 52342–52364. Curran Associates, Inc.

Pedro Henrique Martins, Zita Marinho, and Andre Martins. 2022. ∞-former: Infinite memory transformer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5468–5485, Dublin, Ireland. Association for Computational Linguistics.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Qingyang Wu, Zhenzhong Lan, Kun Qian, Jing Gu, Alborz Geramifard, and Zhou Yu. 2022a. Memformer: A memory-augmented transformer for sequence modeling. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 308–318, Online only. Association for Computational Linguistics.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022b. Memorizing transformers. In *International Conference on Learning Representations*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pages 21875–21895.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Zhaoyang Zhang, Wenqi Shao, Yixiao Ge, Xiaogang Wang, Jinwei Gu, and Ping Luo. 2024. Cached transformers: Improving transformers with differentiable memory cache. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 16935–16943. AAAI Press.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. 2023. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 34661–34710. Curran Associates, Inc.

A Additional Results and Robustness

This appendix reports supplementary results for Section 4. The main text focuses on the primary gated-readout results. Here we provide additional memory-level diagnostics, naive-readout variants, and robustness analyses across thresholds and random seeds.

A.1 Additional memory-level diagnostics

The main text reports source survival versus accuracy and prefix-all memory attention under the gated readout. Here we provide complementary diagnostics for the source-pinned control and for prediction-error types. These diagnostics further test whether degradation is better explained by physical FIFO source eviction or by retrieval interference from retained non-source entries.

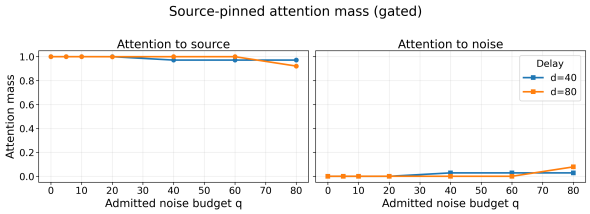


Figure 6: Attention mass in the source-pinned control under the gated readout. Because source states are retained by construction, this diagnostic tests whether admitted non-source entries dominate retrieval when source eviction is removed.

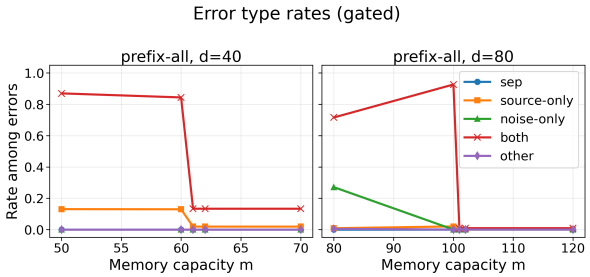


Figure 7: Error-type rates under the gated readout. Failures concentrate in configurations where source states are not sufficiently retained, supporting the source-eviction interpretation.

Figure 6 shows the attention distribution in the source-pinned control. When source states are explicitly retained, attention remains concentrated on source entries. This supports the interpretation that the degradation observed under prefix-all writes is not caused by the mere presence of non-source entries alone.

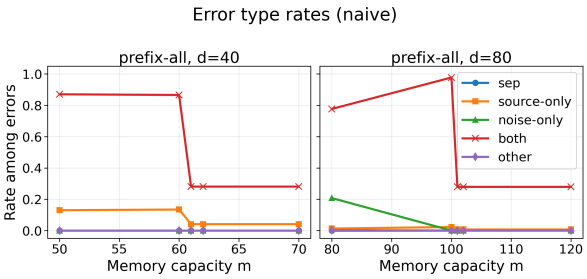


Figure 8: Error-type rates under the naive additive readout. The qualitative failure pattern is consistent with the gated variant, although the naive readout is less stable near strict success thresholds.

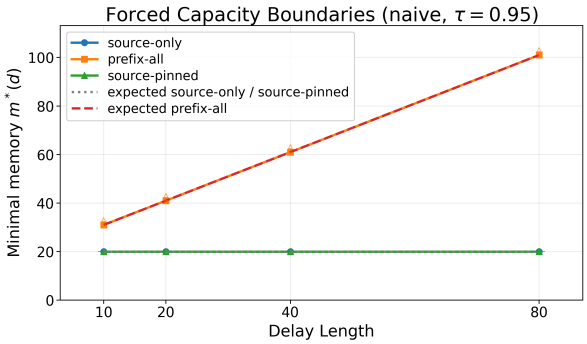


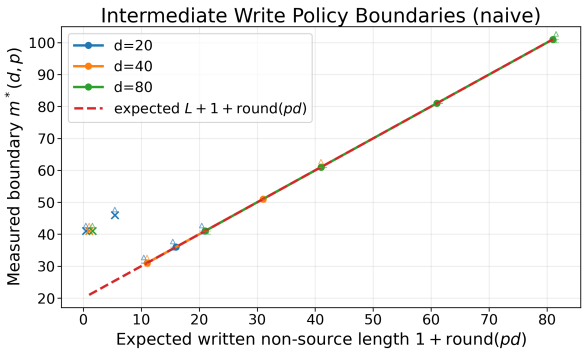
Figure 9: Forced-setting capacity boundary under the naive additive readout. The write-policy-induced boundary remains visible, but the naive readout is less stable than the gated variant. Finite-seed min–max ranges and censoring symbols follow the convention in Figure 2.

Figures 7 and 8 report error-type diagnostics for the two readout variants. Across both readouts, the dominant errors occur when source states are insufficiently retained.

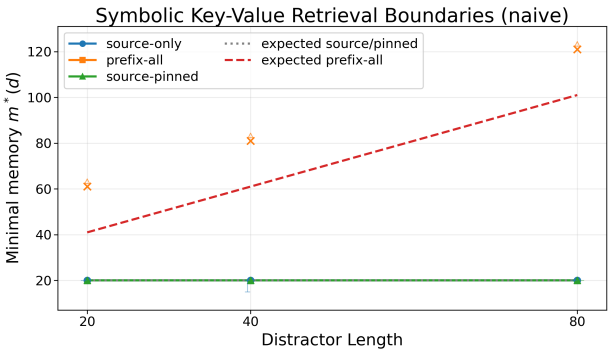
A.2 Naive-readout variants

The main text reports the primary gated-readout results. This subsection reports corresponding naive-readout variants for the main forced-setting boundary, the intermediate partial-prefix diagnostic, and the symbolic key–value retrieval diagnostic. The naive readout is less stable near strict thresholds, but the qualitative write-policy effect remains the same: source-preserving policies require capacity near the source length, while prefix-all or higher-contamination writes require larger memory as more non-source content enters the FIFO buffer.

Figure 9 shows the main forced-setting boundary under the naive additive readout. Although the transition is less stable than under gated retrieval, the same source-retention pattern remains visible: source-preserving configurations require capacity near the source length, while prefix-all writes re-



(a) Intermediate partial-prefix writes, naive



(b) Symbolic key–value retrieval, naive

Figure 10: Naive-readout variants for the intermediate partial-prefix diagnostic and the symbolic key–value retrieval diagnostic. The qualitative write-policy effect is consistent with the gated-readout results: increasing the amount of written non-source content increases the required memory capacity, while source-preserving configurations remain closer to the source-length boundary. Finite-seed min–max ranges and right-censoring symbols follow the convention in Figure 5.

quire larger capacity as delay increases.

Figure 10 shows that both readouts require increasing capacity as more non-source states are written, with cleaner boundaries under the gated readout.

A.3 Boundary stability across thresholds and seeds

Finally, we report robustness checks for the capacity-boundary metric. Table 1 reports threshold sensitivity, while Tables 2, 3, and 4 report seed-wise boundary statistics for the forced delayed-copy, intermediate partial-prefix, and symbolic key–value experiments. Across these checks, source-only writes and the source-pinned control remain near $m = L$, while prefix-all writes follow the expected FIFO source-retention threshold $L + 1 + d$. The naive readout is less stable under stricter thresholds, but preserves the same qualitative write-policy effect.

B Retention Calculation

This appendix gives the deterministic FIFO source-retention calculation used to interpret the capacity-boundary results.

Source-only writes. For source-only writes, the memory receives only the source segment of length L. Thus, retaining the full source requires

$$m \geq L.$$

Since L is fixed across delay lengths, the expected retention boundary is approximately delay-invariant.

Prefix-all writes. For prefix-all writes, the memory receives the full prefix before the query segment:

$$[\texttt{source}][\texttt{sep}][\texttt{noise}],$$

whose length is

$$L + 1 + d.$$

The source states are written first, followed by the separator and d noise states. After the full prefix has been written, FIFO memory retains only the most recent m states. Therefore, the number of retained source states is

$$r(d, m) = \min\left(L, \max(0, m - (1 + d))\right).$$

Retaining the full source requires

$$m \geq L + 1 + d.$$

Thus, prefix-all writes induce a delay-dependent retention boundary because increasing d adds more later non-source states that can evict earlier source states.

Intermediate partial-prefix writes. For intermediate partial-prefix writes, the memory receives the source segment, the separator, and only the first $\mathrm{round}(pd)$ noise states:

$$[\texttt{source}][\texttt{sep}][\texttt{noise}_{1:\mathrm{round}(pd)}],$$

where $p \in [0, 1]$ controls the non-source write strength. The number of written non-source states is therefore

$$1 + \mathrm{round}(pd).$$

Model	Policy	d	Expected	$m^*_{.90}$	$m^*_{.95}$	$m^*_{.98}$
gated	prefix-all	10	31	30	30	31
gated	prefix-all	20	41	40	40	41
gated	prefix-all	40	61	60	61	61
gated	prefix-all	80	101	100	101	101
gated	source-only	10	20	20	20	20
gated	source-only	20	20	20	20	20
gated	source-only	40	20	20	20	20
gated	source-only	80	20	20	20	20
gated	source-pinned	10	20	20	20	20
gated	source-pinned	20	20	20	20	20
gated	source-pinned	40	20	20	20	20
gated	source-pinned	80	20	20	20	20
naive	prefix-all	10	31	30	31	> 120
naive	prefix-all	20	41	40	41	> 120
naive	prefix-all	40	61	60	61	> 120
naive	prefix-all	80	101	100	101	> 120
naive	source-only	10	20	20	20	20
naive	source-only	20	20	20	20	20
naive	source-only	40	20	20	20	20
naive	source-only	80	20	20	20	20
naive	source-pinned	10	20	20	20	20
naive	source-pinned	20	20	20	20	20
naive	source-pinned	40	20	20	20	20
naive	source-pinned	80	20	20	20	20

Table 1: Threshold sensitivity of the forced-setting capacity boundary $m^*(d)$. Source-only writes and the source-pinned control remain near $m = L$, while prefix-all writes track the expected FIFO source-retention threshold $L + 1 + d$. The FIFO retention trend remains aligned across thresholds, while the naive prefix-all results at $\tau = 0.98$ are right-censored near perfect-copy accuracy.

After these states have been written, the number of retained source states under FIFO eviction is

$$r(d, p, m) = \min\Big(L, \max\big(0, m - [1 + \mathrm{round}(pd)]\big)\Big).$$

Retaining the full source requires

$$m \geq L + 1 + \mathrm{round}(pd).$$

This condition reduces to the prefix-all threshold when $p = 1$, and gives a controlled interpolation between source-plus-separator writes and prefix-all writes.

Source-pinned control. For the source-pinned control, all L source states are retained by construction. In our implementation, admitted noise states are managed by a separate FIFO region with budget q. The total memory size is

$$m = L + q.$$

This control removes source eviction by design. Across the tested budgets $q \leq 80$, stable performance therefore indicates that prefix-all degradation is mainly caused by source eviction rather than by retained noise alone.

With $L = 20$, the prefix-all full-source retention thresholds are

$$L + 1 + d \in \{31, 41, 61, 101\}$$

for delays $d \in \{10, 20, 40, 80\}$. These values motivate the near-boundary capacity points used in the prefix-all memory scans.

For the intermediate partial-prefix experiment, using

$$p \in \{0, 0.25, 0.5, 0.75, 1.0\}$$

and delays $d \in \{20, 40, 80\}$, the full-source retention thresholds are

$$L + 1 + \mathrm{round}(pd).$$

For example, when $L = 20$, the expected thresholds are

$d \backslash p$	0	0.25	0.5	0.75	1.0
20	21	26	31	36	41
40	21	31	41	51	61
80	21	41	61	81	101

These thresholds motivate the near-boundary memory grids used in the intermediate partial-prefix scans.

Model	Policy	d	Expected	Mean ± SD	Range	Cens.	Max scan
gated	prefix-all	10	31	30.14 ± 0.38	30.0–31.0	0/7	120
gated	prefix-all	20	41	40.00 ± 0.00	40.0–40.0	1/7	120
gated	prefix-all	40	61	60.17 ± 0.41	60.0–61.0	1/7	120
gated	prefix-all	80	101	100.57 ± 0.53	100.0–101.0	0/7	120
gated	source-only	10	20	20.00 ± 0.00	20.0–20.0	0/7	30
gated	source-only	20	20	20.00 ± 0.00	20.0–20.0	0/7	30
gated	source-only	40	20	20.00 ± 0.00	20.0–20.0	0/7	30
gated	source-only	80	20	20.00 ± 0.00	20.0–20.0	0/7	30
gated	source-pinned	10	20	20.00 ± 0.00	20.0–20.0	0/7	100
gated	source-pinned	20	20	20.00 ± 0.00	20.0–20.0	0/7	100
gated	source-pinned	40	20	20.00 ± 0.00	20.0–20.0	0/7	100
gated	source-pinned	80	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	prefix-all	10	31	30.25 ± 0.50	30.0–31.0	3/7	120
naive	prefix-all	20	41	40.67 ± 0.58	40.0–41.0	4/7	120
naive	prefix-all	40	61	61.00 ± 0.00	61.0–61.0	2/7	120
naive	prefix-all	80	101	100.80 ± 0.45	100.0–101.0	2/7	120
naive	source-only	10	20	20.00 ± 0.00	20.0–20.0	0/7	30
naive	source-only	20	20	20.00 ± 0.00	20.0–20.0	0/7	30
naive	source-only	40	20	20.00 ± 0.00	20.0–20.0	0/7	30
naive	source-only	80	20	20.00 ± 0.00	20.0–20.0	0/7	30
naive	source-pinned	10	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	source-pinned	20	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	source-pinned	40	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	source-pinned	80	20	20.00 ± 0.00	20.0–20.0	0/7	100

Table 2: Seed-wise stability of $m^*(d)$ under the main threshold $\tau = 0.95$. Source-only writes and the source-pinned control are stable at $m = L$, while prefix-all boundaries concentrate near the deterministic FIFO retention threshold. Mean, sample standard deviation, and range are computed over finite per-seed boundaries only; censored seeds are not imputed at the maximum scanned capacity. When only one seed has a finite boundary, only that boundary value is shown because the sample standard deviation is undefined. Cens. reports the number of right-censored seeds out of seven. Max scan reports the largest capacity in the corresponding policy-specific grid.

The retention conditions above describe full-source survival. Since the empirical capacity boundary is defined by a finite accuracy threshold $\tau = 0.95$, the measured boundary can occasionally be one scanned capacity point below the full-source retention condition when partial source survival is sufficient to meet the success threshold.

C Experimental and Reproducibility Details

This appendix provides the complete data-generation, implementation, training, evaluation, and diagnostic protocol supporting the main text. Code, configurations, result files, and plotting scripts are available at https://github.com/LiuSunshine1223/WritePolicyFIFO.

C.1 Delayed-copy data

The delayed-copy task uses synthetic sequences of the form

$$[\mathsf{source}][\mathsf{sep}][\mathsf{noise}][\mathsf{query}].$$

Unless otherwise stated, the source and query segments have length $L = 20$, and the delay/noise length is

$$d \in \{10, 20, 40, 80\}.$$

The base vocabulary contains 20 symbols. Source and noise tokens are sampled i.i.d. from this base vocabulary. The separator is a deterministic token, and the query segment consists of a fixed placeholder token repeated L times. The target is to reproduce the source tokens at the final L query positions.

C.2 Symbolic key–value retrieval variant

The symbolic key–value retrieval variant uses the same model family, masking regime, and external FIFO memory mechanism as the delayed-copy task, and compares source-only, prefix-all, and source-pinned configurations. Each source segment is a 20-token target record of the form

$$[\mathsf{FACT}][\mathsf{KEY}][\mathsf{IS}][\mathsf{VALUE}_{1:17}].$$

The intervening segment consists of 10-token distractor records of the same general form whose keys differ from the target key; records are concatenated and truncated to the specified delay length $d \in \{20, 40, 80\}$. The final 20-token query has the form

$$[\mathsf{QUERY}][\mathsf{KEY}][\mathsf{ASK}] \ldots [\mathsf{ASK}],$$

so it contains the target key but not its value. The training target is the complete 20-token target record, reconstructed at the positions of the

Model	d	p	Expected	Mean ± SD	Range	Cens.	Max scan
gated	20	0	21	20.00 ± 0.00	20.0–20.0	0/7	41
gated	20	0.25	26	25.50 ± 0.71	25.0–26.0	5/7	46
gated	20	0.50	31	30.00 ± 0.00	30.0–30.0	3/7	51
gated	20	0.75	36	35.00 ± 0.00	35.0–35.0	2/7	56
gated	20	1	41	40.00 ± 0.00	40.0–40.0	1/7	61
gated	40	0	21	20.29 ± 0.49	20.0–21.0	0/7	41
gated	40	0.25	31	30.00 ± 0.00	30.0–30.0	1/7	51
gated	40	0.50	41	40.00 ± 0.00	40.0–40.0	4/7	61
gated	40	0.75	51	50.17 ± 0.41	50.0–51.0	1/7	71
gated	40	1	61	60.17 ± 0.41	60.0–61.0	1/7	81
gated	80	0	21	20.00 ± 0.00	20.0–20.0	0/7	41
gated	80	0.25	41	40.20 ± 0.45	40.0–41.0	2/7	61
gated	80	0.50	61	60.00 ± 0.00	60.0–60.0	1/7	81
gated	80	0.75	81	80.00 ± 0.00	80.0–80.0	1/7	101
gated	80	1	101	100.57 ± 0.53	100.0–101.0	0/7	121
naive	20	0	21	21.00	21.0–21.0	6/7	41
naive	20	0.25	26	–	–	7/7	46
naive	20	0.50	31	31.00 ± 0.00	31.0–31.0	2/7	51
naive	20	0.75	36	35.33 ± 0.58	35.0–36.0	4/7	56
naive	20	1	41	40.67 ± 0.58	40.0–41.0	4/7	61
naive	40	0	21	21.00	21.0–21.0	6/7	41
naive	40	0.25	31	30.80 ± 0.45	30.0–31.0	2/7	51
naive	40	0.50	41	41.00 ± 0.00	41.0–41.0	0/7	61
naive	40	0.75	51	50.86 ± 0.38	50.0–51.0	0/7	71
naive	40	1	61	61.00 ± 0.00	61.0–61.0	2/7	81
naive	80	0	21	21.00 ± 0.00	21.0–21.0	5/7	41
naive	80	0.25	41	40.86 ± 0.38	40.0–41.0	0/7	61
naive	80	0.50	61	61.00 ± 0.00	61.0–61.0	0/7	81
naive	80	0.75	81	81.00 ± 0.00	81.0–81.0	0/7	101
naive	80	1	101	100.80 ± 0.45	100.0–101.0	2/7	121

Table 3: Seed-wise stability of the intermediate partial-prefix boundary $m^*(d, p)$ at $\tau = 0.95$. Mean, sample standard deviation, and range are computed over finite per-seed boundaries only; censored seeds are not imputed. When only one seed has a finite boundary, only that boundary value is shown because the sample standard deviation is undefined; dashes indicate that no seed has a finite boundary. Cens. reports the number of right-censored seeds out of seven, and Max scan reports the largest tested capacity for the configuration.

final query segment, and evaluation measures token accuracy over all positions in that record.

All tokens are synthetic symbols drawn from a 21-token vocabulary. The structural token IDs are 0 for FACT, 1 for QUERY, 2 for IS, and 3 for ASK; ID 4 is reserved but unused in this experiment, IDs 5–12 are eight key symbols, IDs 13–19 are seven value symbols, and ID 20 is the separator. The target key is sampled uniformly from the eight key symbols, and each of its 17 value tokens is sampled independently and uniformly from the value vocabulary. Each 10-token distractor record has the form

$$[\mathsf{FACT}][\mathsf{KEY}][\mathsf{IS}][\mathsf{VALUE}_{1:7}].$$

Its key is sampled uniformly from the seven keys other than the target key, and its value tokens are sampled independently and uniformly from the value vocabulary. Distractor records are sampled independently, so their keys and values may repeat across records; at the record level, the key-collision rule excludes only the target key from distractor-key sampling. The final query contains one QUERY token, the target key, and 18 repetitions of ASK.

We use this controlled structured retrieval diagnostic to test whether the finite-memory write-policy effect persists beyond the original random-token delayed-copy task. The symbolic experiments use the same capacity-boundary procedure as the delayed-copy experiments. Censored symbolic boundaries are reported when no scanned capacity reaches the success threshold.

C.3 Attention masks

We use local windowed self-attention with window size w. Let positions be indexed by $t, s \in \{1, \ldots, T\}$, where T is the sequence length. The standard local window mask allows each position to attend to itself and the previous $w - 1$ positions:

$$M^{\text{win}}_{t,s} = \begin{cases} 0, & \max(1, t - w + 1) \le s \le t, \\ -\infty, & \text{otherwise.} \end{cases}$$

In the natural setting, the model uses this standard local mask. In the forced setting, we additionally block all query-to-source attention edges. Let

$$\mathcal{S} = \{1, \ldots, L\}$$

denote source positions and

$$\mathcal{Q} = \{T - L + 1, \ldots, T\}$$

Model	Policy	d	Expected	Mean ± SD	Range	Cens.	Max scan
gated	prefix-all	20	41	39.00 ± 0.00	39.0–39.0	1/7	61
gated	prefix-all	40	61	60.00 ± 0.71	59.0–61.0	2/7	81
gated	prefix-all	80	101	99.60 ± 0.89	99.0–101.0	2/7	121
gated	source-only	20	20	19.29 ± 1.89	15.0–20.0	0/7	30
gated	source-only	40	20	20.00 ± 0.00	20.0–20.0	0/7	30
gated	source-only	80	20	20.00 ± 0.00	20.0–20.0	0/7	30
gated	source-pinned	20	20	20.00 ± 0.00	20.0–20.0	0/7	100
gated	source-pinned	40	20	20.00 ± 0.00	20.0–20.0	0/7	100
gated	source-pinned	80	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	prefix-all	20	41	39.00	39.0–39.0	6/7	61
naive	prefix-all	40	61	60.00 ± 1.41	59.0–61.0	5/7	81
naive	prefix-all	80	101	99.00 ± 0.00	99.0–99.0	5/7	121
naive	source-only	20	20	20.00 ± 0.00	20.0–20.0	0/7	30
naive	source-only	40	20	19.29 ± 1.89	15.0–20.0	0/7	30
naive	source-only	80	20	20.00 ± 0.00	20.0–20.0	0/7	30
naive	source-pinned	20	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	source-pinned	40	20	20.00 ± 0.00	20.0–20.0	0/7	100
naive	source-pinned	80	20	20.00 ± 0.00	20.0–20.0	0/7	100

Table 4: Seed-wise stability of the symbolic key–value retrieval boundary $m^*(d)$ at $\tau = 0.95$. Mean, sample standard deviation, and range are computed over finite per-seed boundaries only; censored seeds are not imputed. When only one seed has a finite boundary, only that boundary value is shown because the sample standard deviation is undefined. Cens. reports the number of right-censored seeds out of seven, and Max scan reports the largest tested capacity for the configuration.

Write policy	Full-source retention condition	Expected behavior
Source-only	$m \geq L$	The full source is retained once capacity reaches the source length, so $m^*(d)$ is approximately delay-invariant.
Prefix-all	$m \geq L + 1 + d$	The memory must retain the source after the separator and all delay/noise states are written, so $m^*(d)$ grows with the written prefix length.
Intermediate partial-prefix	$m \geq L + 1 + \mathrm{round}(pd)$	The boundary increases with the amount of written non-source content, yielding intermediate boundaries between source-plus-separator and prefix-all writes.
Source-pinned	Source retained by construction; $m = L + q$	Source eviction is removed by design, so increasing the admitted noise budget does not change the physical source-retention condition.

Table 5: Deterministic full-source retention conditions under different write policies. These conditions describe physical source survival under FIFO eviction. The empirical boundary m^* is measured by an accuracy threshold and may differ slightly from the full-retention condition.

denote query positions. The forced mask is

$$M^{\mathrm{forced}}_{t,s} = \begin{cases} -\infty, & t \in \mathcal{Q} \text{ and } s \in \mathcal{S}, \\ M^{\mathrm{win}}_{t,s}, & \text{otherwise.} \end{cases}$$

This forced mask removes direct query-to-source attention and is used to test memory-dependent behavior without that direct shortcut. Unless otherwise stated, forced-setting experiments use $w = 8$.

C.4 External FIFO memory and readout variants

The memory-augmented model uses a windowed Transformer encoder followed by a single external-memory read at the final query positions. Token embeddings are combined with learned absolute positional embeddings at the encoder input. Selected hidden states from the final encoder layer are written into a finite FIFO memory according to the write policy. Stored hidden states are used as both memory keys and memory values. When the memory capacity is exceeded, the oldest entries are evicted. The memory is reset for each training example and does not persist across minibatches or examples.

Let H_q denote the query-position hidden states and let K, V denote the stored memory keys and values. The memory read is computed by attention over the external memory:

$$A = \mathrm{softmax}\left(\frac{H_q K^\top}{\sqrt{d_{\mathrm{model}}}}\right), \qquad R = AV.$$

This is a content-based post-encoder read; no additional relative-position term is applied to the memory-attention logits. We evaluate two readout-

fusion variants. The naive additive readout uses

$$\widetilde{H}_q = H_q + R.$$

The gated readout computes

$$G = \sigma(W_g H_q + b_g),$$

where W_g maps each query representation to a scalar gate that is broadcast over its hidden dimension. The gated readout then fuses the original query representation with the memory read as

$$\widetilde{H}_q = (1 - G) \odot H_q + G \odot R.$$

The final prediction head is applied to $\widetilde{H}_q$ at the query positions.

C.5 Write policies

We compare three write-policy families and one source-pinned control.

Source-only. The source-only policy writes only source states into memory. It serves as an oracle-like diagnostic control, because task-relevant states are assumed to be identifiable before writing.

Prefix-all. The prefix-all policy writes the full prefix before the query,

$$[\texttt{source}][\texttt{sep}][\texttt{noise}].$$

Under finite FIFO memory, later separator and noise states can evict earlier source states.

Intermediate partial-prefix. The intermediate partial-prefix policy always writes all source states and the separator, but writes only the first $\mathrm{round}(pd)$ noise states:

$$[\texttt{source}][\texttt{sep}][\texttt{noise}_{1:\mathrm{round}(pd)}],$$

where

$$p \in \{0, 0.25, 0.5, 0.75, 1.0\}.$$

Thus, $p = 0$ corresponds to source-plus-separator writes, and $p = 1$ recovers prefix-all writes. This policy provides a controlled interpolation in the amount of written non-source content.

Source-pinned control. The source-pinned control explicitly retains all L source states. In our implementation, admitted noise states are managed by a separate FIFO buffer. If the admitted noise budget is q, then the total memory size is

$$m = L + q.$$

This control removes source eviction by construction while still allowing the readout to see a controlled amount of non-source content.

C.6 Training protocol

Unless otherwise stated, models are trained with Adam using learning rate 10^{-3}, batch size 16, model dimension 64, 2 Transformer encoder layers, and 4 attention heads.

Delayed-copy Experiments A–C, including the main forced-boundary experiments, memory diagnostics, source-pinned control experiments, and intermediate partial-prefix experiments, are trained for 3001 optimization steps. Symbolic key–value Experiment D uses a fixed budget of 9001 optimization steps across all policies, delays, capacities, readouts, and seeds, with the remaining optimization settings unchanged. This longer fixed budget was chosen after preliminary 3001-step runs indicated slower convergence on the symbolic task. Those preliminary runs are not included in the reported aggregates or capacity boundaries; all reported main Experiment D results follow the 9001-step protocol. At every optimization step, the data generator samples a fresh synthetic minibatch; training does not repeatedly draw from a fixed dataset.

All reported results are averaged over seven random seeds:

$$\{0, 1, 2, 3, 4, 5, 6\}.$$

We report `acc_mean_tail`, the mean query-token accuracy over the final 500 training steps. Since each step uses a fresh minibatch, averaging reduces sensitivity to any one minibatch and summarizes end-of-training behavior. This is an operational training-tail measure rather than a held-out generalization metric.

C.7 Capacity boundary definition

In forced-setting experiments, success is defined by a fixed accuracy threshold

$$\overline{\mathrm{Acc}} \geq \tau, \qquad \tau = 0.95,$$

where $\overline{\mathrm{Acc}}$ denotes the seven-seed average of `acc_mean_tail` for the corresponding configuration.

For source-only and prefix-all writes, the capacity boundary at delay d is

$$m^*(d) = \min\left\{m \in \mathcal{M} : \overline{\mathrm{Acc}}(d, m) \geq \tau\right\}.$$

For intermediate partial-prefix writes, the boundary is

$$m^*(d, p) = \min\left\{m \in \mathcal{M} : \overline{\mathrm{Acc}}(d, p, m) \geq \tau\right\}.$$

If no scanned capacity reaches the threshold, we report a censored boundary, e.g.,

$$m^*(d) > \max(\mathcal{M}).$$

C.8 Capacity grids

We use policy-specific capacity grids with additional points near expected FIFO retention thresholds. The natural-setting sanity check uses $m = L = 20$.

For source-only delayed-copy writes, the executed capacity grid is

$$\mathcal{M}_{\mathrm{src}} = \{5, 10, 15, 20, 25, 30\}.$$

For prefix-all delayed-copy experiments, the executed capacity grid is

$$\mathcal{M}_{\mathrm{pre}} = \{10, 20, 30, 31, 32, 40, 41, 42, 50, \\ 60, 61, 62, 70, 80, 100, 101, 102, 120\}.$$

The near-boundary values are centered around the expected FIFO full-source retention thresholds $L + 1 + d$.

The post-training memory diagnostics in Experiment B use delays

$$d \in \{40, 80\}.$$

Their executed source-only grid is

$$\mathcal{M}^B_{\mathrm{src}} = \{10, 15, 20, 25, 30\},$$

and the prefix-all grids are

$$\mathcal{M}^B_{\mathrm{pre}}(40) = \{50, 60, 61, 62, 70\},$$
$$\mathcal{M}^B_{\mathrm{pre}}(80) = \{80, 100, 101, 102, 120\}.$$

The Experiment B source-pinned grid is

$$\mathcal{M}^B_{\mathrm{pin}} = \{20, 25, 30, 40, 60, 80, 100\}.$$

For the intermediate partial-prefix experiment, we use

$$d \in \{20, 40, 80\},$$
$$p \in \{0, 0.25, 0.5, 0.75, 1.0\}.$$

For each (d, p), we scan capacities around

$$b(d, p) = L + 1 + \mathrm{round}(pd).$$

The grid is

$$\mathcal{M}_{\mathrm{int}}(d, p) = \mathrm{unique}\Big(\{10, 20, \\ b - 2, b - 1, b, b + 1, \\ b + 2, b + 10, b + 20\}\Big).$$

after removing nonpositive values, de-duplicating, and sorting.

For the source-pinned control, we vary the admitted noise budget

$$q \in \{0, 5, 10, 20, 40, 60, 80\},$$

so that the total memory size is

$$m = L + q.$$

Experiment D uses compact policy-specific grids that include sub-threshold controls and concentrate points around the corresponding source-preservation and prefix-all retention thresholds. For source-only writes, the executed grid is

$$\mathcal{M}^D_{\mathrm{src}} = \{10, 15, 20, 25, 30\}.$$

For prefix-all writes, the delay-specific executed grids are

$$\mathcal{M}^D_{\mathrm{pre}}(20) = \{10, 20, 39, 40, 41, 42, 43, 51, 61\},$$
$$\mathcal{M}^D_{\mathrm{pre}}(40) = \{10, 20, 59, 60, 61, 62, 63, 71, 81\},$$
$$\mathcal{M}^D_{\mathrm{pre}}(80) = \{10, 20, 99, 100, 101, 102, \\ 103, 111, 121\}.$$

The Experiment D source-pinned control reserves all $L = 20$ source states and therefore begins at total capacity $m = 20$; its executed total-capacity grid is

$$\mathcal{M}^D_{\mathrm{pin}} = \{20, 40, 60, 100\},$$

corresponding to admitted-noise budgets $q \in \{0, 20, 40, 80\}$.

C.9 Experiment groups

We organize the experiments into four groups. First, natural-setting sanity checks verify that the external memory pathway can solve delayed copying when source states are retained. Second, forced-setting delayed-copy boundary experiments measure $m^*(d)$ for source-only writes, prefix-all writes, and the source-pinned control under $w = 8$. Third, intermediate partial-prefix experiments measure $m^*(d, p)$, testing whether the apparent memory requirement changes continuously with the amount of written non-source content. Fourth, symbolic key–value retrieval experiments measure the corresponding capacity boundaries for structured key–value records.

C.10 Memory-level diagnostics

For trained forced-setting configurations, we record memory-level diagnostics after training. These diagnostics are used only for analysis and are never provided to the model. For the source-survival plot in Figure 3, we place the trained model in evaluation mode and compute `eval_acc` over 16 freshly generated same-distribution minibatches for each seed and configuration. The evaluation generator uses a separate deterministic stream initialized with the training seed plus 100000. The plotted configuration-level accuracy averages these seed-level values over the seven seeds. This post-training diagnostic is separate from `acc_mean_tail` and is not used to determine a capacity boundary.

We measure source survival as

$$\rho_{\mathrm{src}}(d, m) = \frac{n_{\mathrm{src}}^{\mathrm{ret}}(d, m)}{L},$$

where $n_{\mathrm{src}}^{\mathrm{ret}}(d, m)$ is the number of source entries retained in memory. For memory attention, we compute the total attention mass assigned to entries originating from source, separator, and noise positions. For example, the source attention mass is

$$\alpha_{\mathrm{src}} = \sum_{j \in \mathcal{I}_{\mathrm{src}}} A_j,$$

where A_j is the memory-attention probability assigned to memory entry j, and $\mathcal{I}_{\mathrm{src}}$ is the set of retained memory entries whose origin is the source segment.

We also categorize incorrect predictions according to whether the predicted token appears in the source segment, the noise segment, both, the separator token, or none of these categories. These diagnostics help separate physical source eviction from retrieval interference.