THE CONTEXT TRADE-OFF: UNDERSTANDING INFORMATION USE IN LANGUAGE MODELS AND AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models face a fundamental tension at inference time: additional context provides more evidence, yet longer inputs can induce *context rot*—the degradation of model performance as input length grows. The same tension arises in language model agents as tool responses accumulate in their history. We ask when additional context helps or hurts, and how context management changes this trade-off. Across multiple benchmarks, evaluation formats, and model families, performance often rises and then falls as context grows. We call this pattern *context saturation*. Its shape and peak vary across models and tasks. To explain this pattern, we compare a model with an ideal solver: a reader that sees the same input and extracts everything it implies about the task, so its performance measures how much information the input contains. Two properties make this comparison useful. Adding background text unrelated to the answer cannot lower the ideal solver's performance, and neither can rewriting an agent's history, as long as earlier observations can still be retrieved. Under such changes, any drop in the model's performance is therefore a failure to use information that is still available, and any gain is information made easier to use. Controlled experiments across benchmarks and model families support three findings. (1) In most controlled settings, the rise reflects evidence arriving and the fall reflects evidence becoming harder to use, so the peak depends on the model. (2) Performance falls even when all needed evidence is kept. Evidence is also used jointly: a piece of evidence without its complement is lost more often, so a missing piece can matter more in longer contexts. (3) For agents, rewriting the same observations into a compact form can raise success several-fold. Rewritten histories fail when the agent reads past its target or can no longer revisit earlier observations. In these controlled settings, long context fails not for lack of information but because information becomes hard to use. Our code is available at here.

1 INTRODUCTION

Long-horizon tasks, e.g., multi-document question answering, repository-scale coding, agentic workflows, and extended tool-use trajectories, have made long-context understanding a central capability of large language models (LLMs) (Bai et al., 2025; Bertsch et al., 2025; Cao et al., 2026; Zhang et al., 2024; Sun et al., 2025b; Ye et al., 2025a; Chen et al., 2025). Long context lets a model combine evidence across documents and retain observations across interactions, yet performance can degrade as the input grows. This *context rot* depends on the model–task pair and on where evidence sits among other text (Hong et al., 2025; Liu et al., 2024b; Du et al., 2025). Context engineering pipelines respond by selecting, organizing, and compressing what the model sees (Ye et al., 2025a; Anthropic Applied AI Team, 2025).

Task-relevant evidence can be available without being used for answering a question or selecting an action. In this work, we study **how language models and agents use task-relevant information, and how its surrounding context and representation affect their answers and actions**. We find that once evidence is in the context, more context mostly makes it harder to use, not less informative.

Across multiple benchmarks and model families, we observe that performance often rises and then falls as context grows, a pattern we call *context saturation* (Section 4). The shape and observed peak

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.

vary by model and task. These curves motivate a closer examination of how the information supplied and the model's use of it jointly determine performance.

We compare a model with an ideal solver that reads the same input. In statistical decision theory, the ideal value of an input is the best performance any decision rule could achieve with it. The model's shortfall is its *utilization gap*. The comparison parallels usable information (Xu et al., 2020; Ethayarajh et al., 2022) (Section 3). Two properties make the comparison useful. Adding background text around recoverable evidence leaves ideal value unchanged, and so does rewriting an agent's history while earlier observations stay retrievable. Under such changes, a loss in performance is a failure to use information that is still available, and a gain is information made easier to use. Since models reading the same input share its ideal value, the model whose gap grows more slowly peaks with the same or more retrieved context.

Controlled experiments support three findings. (1) The rise in performance tracks evidence arriving, and the fall tracks declining use of evidence already present. When we control when evidence arrives, the fraction of evidence delivered, multiplied by the score with all evidence present from the start, approximates performance on RepoQA and HELMET. On LongBench multi-hop QA, the larger model in each family peaks no earlier (Section 4). (2) Performance falls even with all needed evidence kept. Questions a model had solved fall to 53% accuracy as background text from other documents is added. With both evidence paragraphs kept, HotpotQA F1 falls by 7–16 points from 2K to 120K tokens. Evidence is also used jointly: for Qwen2.5-14B, a lone paragraph is lost more often than a pair, so a missing paragraph can matter more in longer contexts (Section 5). (3) For agents (Section 6), rewriting the same observations changes how well they are used. All histories start from a shared prefix, the first steps of a multi-step lookup task, and we rewrite the history while every earlier observation stays retrievable. *Extract*, which keeps the observed records, raises Qwen2.5-7B's summation success from 6 to 70 of 128 task instances. *Recent*, which keeps only the latest interactions, makes Qwen2.5-7B read past the third record on lookup. Blocking rereads, the one change that removes information, drops Llama-3.1-8B's Recent success from 39 to 9.

Our contributions are:

- A decision-theoretic framework that separates the information an input contains from the model's use of it, under which added background text and history management change only use while evidence stays recoverable, with consequences for the peak and for how evidence is used jointly.
- Evidence that the rise and fall of performance with context reflect evidence arrival and declining use, and that retained evidence loses value in ways that depend on the model, the background text, and the companion evidence.
- Evidence that managing an agent's history changes how the same observations are used: extraction recovers use, and histories fail when the agent reads past its target or cannot reread earlier observations.

2 Related Work

Long-context modeling. Despite expanding context windows, LLMs fail to use long inputs reliably. Benchmarks consistently report accuracy degradation as a function of input length, evidence position, and distractor density (Bai et al., 2025; Bertsch et al., 2025; Dou et al., 2026; Ye et al., 2025b; Hsieh et al., 2024), giving rise to the *lost-in-the-middle* and *context rot* phenomena (Liu et al., 2024b; Hong et al., 2025; Laban et al., 2025; He et al., 2025; Du et al., 2025). A parallel line of work on retrieval-augmented generation (RAG) shows that additional passages help in small numbers but hurt once noise, redundancy, or conflicts dominate (Jin et al., 2025a; Leng et al., 2024; Li et al., 2024; Xu et al., 2024). Recent studies further analyze the mechanisms and predictability of this degradation (Montgomery et al., 2025; Chen et al., 2026; Qian et al., 2024; Ye et al., 2026; Tang et al., 2026; Zhao et al., 2025; Bansal et al., 2026; Yoran et al., 2024). The rise-then-fall pattern with more retrieved passages was reported by Jin et al. (2025a), who trace the decline to hard negatives, and Du et al. (2025) show that length alone hurts even when retrieval is perfect. We retain recoverable evidence to isolate losses of use, show how a missing paragraph can matter more as performance falls, and extend the analysis to agent histories.

Context management. The cost of long inputs has motivated extensive work on managing context length. Adaptive RAG methods adjust the amount of retrieved evidence per query (Cao et al., 2025; Taguchi et al., 2025; Iratni et al., 2025; Min et al., 2025; Lahmy & Yozevitch, 2025; Li et al., 2026; Jeong et al., 2024), and agentic approaches interleave retrieval and reasoning to control what enters the context window (Cao et al., 2026; Hui et al., 2025; Sun et al., 2025a; Jin et al., 2025b; Wang et al., 2025; Zhang et al., 2025). Several methods learn when to stop reading or retrieving via latent sufficiency signals, learned value functions, or policies over multi-round retrieval (Xie et al., 2025; Park et al., 2025; Yang et al., 2025; Liu et al., 2025). Our focus is the behavior these context decisions induce: we compare a model with an ideal solver and use controlled interventions to study how evidence, background text, and history representations shape performance.

3 A Theory of Context Use

Information that is present need not be used. A fixed invertible recoding preserves mutual information about a label, yet can reduce a constrained model's predictive performance (Xu et al., 2020; Ethayarajh et al., 2022). We develop the decision-theoretic counterpart for context. Ideal value measures what the observations make possible for a task, and a model realizes part of it (Figure 1). This distinction lets us analyze both context growth and management. Proofs are in Appendix B.

Definition 3.1 (Ideal value and utilization gap)**.** Fix a task τ, a joint law of the world W and observations X, and task utility $u_\tau(W, a) \in [0, 1]$. Observations include the query and any permitted prior knowledge. The ideal value is $G_\tau(X) = \sup_\pi \mathbb{E}[u_\tau(W, \pi(X))]$, over rules using X and independent randomness. A model with expected utility $U_\tau^{\mathcal{M}}(X)$ has utilization gap $D_\tau^{\mathcal{M}}(X) = G_\tau(X) - U_\tau^{\mathcal{M}}(X) \geq 0$.

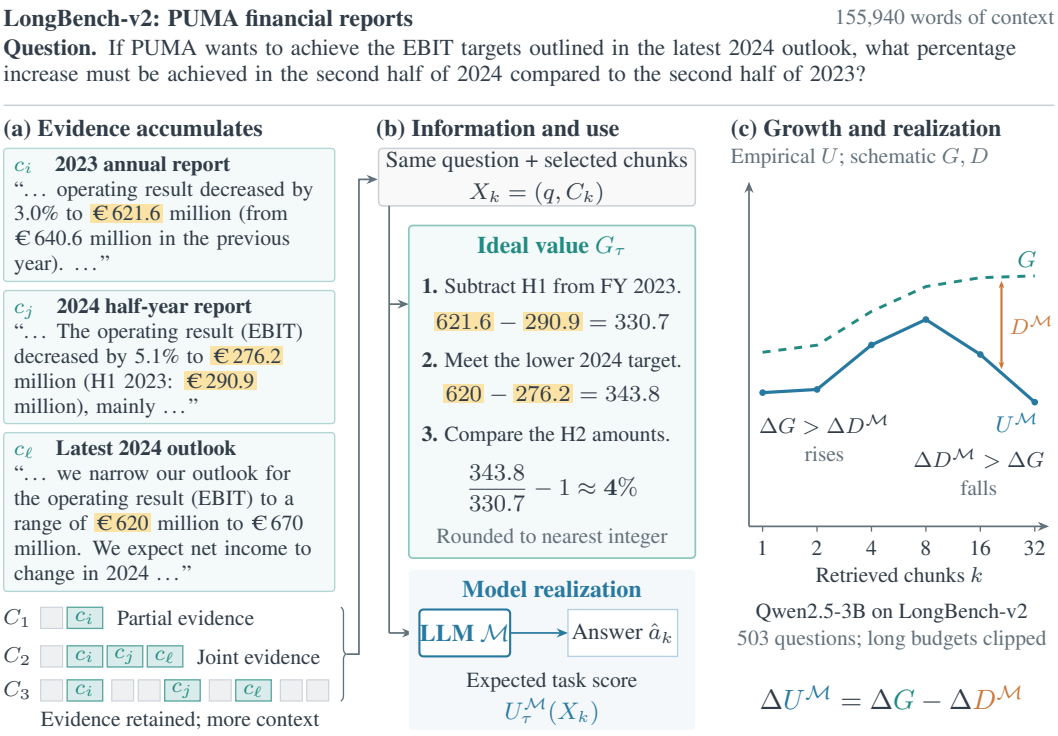

Figure 1: **Information in the context can grow while the model uses less of it. (a)** A LongBench-v2 question over PUMA financial reports (Bai et al., 2025); highlighted amounts are the evidence the answer needs, and gray blocks are background text. Moving down, the context first gains evidence and then only background text. **(b)** An ideal solver reading the same context derives the gold answer, 4%, from the highlighted amounts; its performance is the ideal value G, and the model's performance U can fall short. **(c)** The observed Qwen2.5-3B curve U from Figure 2 up to $k = 32$ with schematic G (dashed) and D: G rises and then nearly levels off, and U falls once D grows faster than G. On the 211 questions whose documents fill every budget, this fall is within noise (Appendix C.2); Section 4 isolates it with controlled designs.

The gap measures task value that the model does not realize, whether it fails to find, combine, or act on the evidence. We say a model makes better *use* of information when its gap is smaller. Ideal value depends on the task, not the model. We suppress τ when the task is fixed.

Lemma 3.2 (Retained information)**.** *If X is recoverable from X', that is, $X = f(X')$ almost surely, then $G(X') \geq G(X)$ and $\Delta U^{\mathcal{M}} = \Delta G - \Delta D^{\mathcal{M}}$, where Δ denotes the change from X to X'.*

This is the decision-theoretic sense in which more informative observations cannot hurt (Blackwell, 1953). Model performance has no such guarantee. We turn this contrast into two properties that the experiments rely on and two consequences that they test.

Two properties. Our QA interventions form $X = \Pi(E, Z)$ by placing E, the question and supplied evidence, at design-specified positions among background text Z from other documents. Let s index which evidence items E contains, and let $F(s, z)$ be mean model score with evidence state s and background-text regime z. Let $D(s, z)$ be the corresponding gap and $B(z) = F(s^+, z) - F(s, z)$ the benefit of adding evidence, where s^+ adds one evidence item to s.

Proposition 3.3 (Adding context and managing it)**.** (a) *Fix the task and joint law of (W, E) across background-text regimes. If E is recoverable from $X = \Pi(E, Z)$ and $Z \perp W \mid E$ in each regime, then $G(X) = G(E)$ throughout. Hence*

$$F(s, z') - F(s, z) = -\big[D(s, z') - D(s, z)\big], \qquad B(z') - B(z) = \Delta_z D(s) - \Delta_z D(s^+), \quad (1)$$

where $\Delta_z D(s) = D(s, z') - D(s, z)$. (b) *If a representation $\widetilde{X} = T(X)$ uses only X and independent randomness, then $G(\widetilde{X}) \leq G(X)$; if equality holds, $\Delta U^{\mathcal{M}} = -\Delta D^{\mathcal{M}}$.*

Property (a) makes experiments that add background text interpretable. The evidence stays in the input, so the ideal solver does exactly as well as before. Every loss in the controlled designs of Sections 4 and 5 is therefore a loss of use. Property (b) does the same for management: a rewritten history gives the ideal solver nothing new, so every gain in Section 6 is use recovered.

For an agent, ideal value is the best score reachable from the current history. With tools and the remaining call limit fixed, let $G(H)$ be the best expected terminal score from history H, counting observations still obtainable through tools. A managed history $\widetilde{H} = T(H)$ changes ideal value by $\Delta G = G(\widetilde{H}) - G(H) \leq 0$ and is *decision-sufficient* when $\Delta G = 0$. By Definition 3.1,

$$U^{\mathcal{M}}(\widetilde{H}) - U^{\mathcal{M}}(H) = D^{\mathcal{M}}(H) - D^{\mathcal{M}}(\widetilde{H}) + \Delta G. \quad (2)$$

A shorter history can improve use, but any loss of ideal value offsets that improvement.

Where the optimum lies. Along a context path $X_1, \ldots, X_n$, let g_k and $d_k^{\mathcal{M}}$ be the step changes in G and $D^{\mathcal{M}}$ from X_{k-1} to X_k, and let $k^\star_{\mathcal{M}}$ be the largest maximizer of $U^{\mathcal{M}}(X_k)$.

Corollary 3.4 (Ordering of the optimum)**.** (i) *If models A and B read the same inputs and $d_k^A \geq d_k^B$ for every k, then $k^\star_A \leq k^\star_B$.* (ii) *If the last evidence arrives at step k_s, so that $g_k = 0$ for all $k > k_s$, and the gap is non-decreasing there, then $U(X_k) \leq U(X_{k_s})$ for every $k > k_s$.*

Ideal value cancels between models on the same inputs, so a model whose use erodes more slowly peaks no earlier, and once the last evidence has arrived, performance improves only if the gap falls. Section 4 examines both parts.

How evidence interacts. By Proposition 3.3(a), background text changes the benefit of an evidence item only through the gap (Equation (1)). Two simple accounts of how background text erodes use, one additive and one multiplicative, restrict this change:

Proposition 3.5 (Restrictions on evidence benefit)**.** *An evidence-independent additive penalty leaves B unchanged; shared attenuation can only reduce it:*

$$F(s, z) = h(s) - c(z) \implies B(z') = B(z), \quad (3)$$

$$F(s, z) = a(z)h(s) + b(z) \implies B(z') \leq B(z), \quad (4)$$

where the second assumes $0 \leq a(z') \leq a(z)$ and $h(s^+) \geq h(s)$.

A benefit that grows with background text rules out both accounts: use must then erode faster with partial evidence than with complete evidence. Section 5.2 tests this.

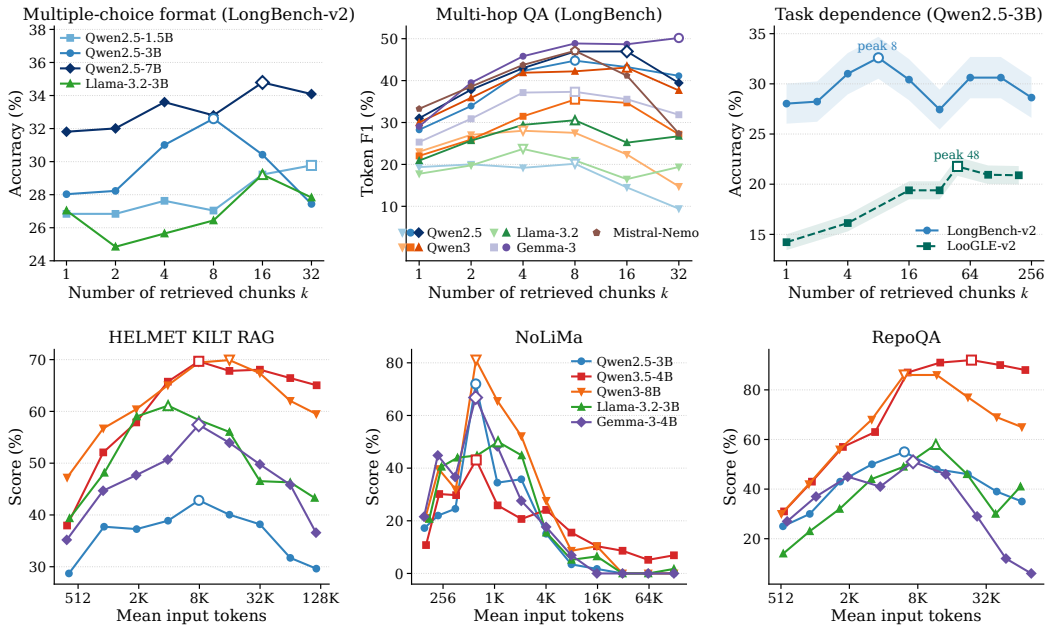

Figure 2: **Performance often rises and then falls as context grows, with a peak that depends on model and task. Top:** natural retrieval with k query-ranked chunks. Left, LongBench-v2 accuracy for four models; middle, LongBench multi-hop F1 for 11 models, each legend entry listing one family from smallest to largest model (Table 3); right, Qwen2.5-3B on LongBench-v2 and LooGLE-v2 (± 1 SE bands). **Bottom:** controlled arrival, where designated evidence enters at set budgets and then stays; on RepoQA and HELMET the rise follows evidence coverage (Figure 6a, Table 4). Hollow markers mark each curve's maximum; the NoLiMa legend applies to the whole bottom row.

4 Where the Peak Comes From

Performance that rises and then falls with context admits two readings: the input stops adding information, or the model stops using what the input already contains. Controlled evidence arrival, read through Corollary 3.4, tells them apart. We first establish the pattern under natural retrieval, then control when evidence arrives.

Performance rises and then falls. We rank document chunks by query similarity, vary the number supplied, and restore them to document order. On LongBench-v2 (Bai et al., 2025), LooGLE-v2 (He et al., 2025), and LongBench multi-hop QA (Bai et al., 2024), most curves rise and then fall, a pattern we call *context saturation*. The peaks lie between $k = 4$ and $k = 48$ chunks depending on the model and task. Mistral-Nemo, for example, falls from 47.1 F1 at $k = 8$ to 27.4 at $k = 32$ on the questions available at each budget (Figure 2, top). On question cohorts held fixed across budgets, the fall for Qwen2.5-3B is small and within noise (Appendix C.2); the controlled designs below fix the evidence and isolate the fall.

The rise is arrival; the fall is use. To see which reading holds, we insert annotated evidence at designated budgets in HELMET KILT RAG (NQ, TriviaQA, PopQA, and HotpotQA subsets), NoLiMa, and RepoQA and keep it as background text grows (Yen et al., 2024; Modarressi et al., 2025; Liu et al., 2024a). On the same budgets we measure two further curves: the fraction of evidence delivered so far, and the score with all evidence present from the start. Their product tracks the observed curve. On RepoQA, its mean absolute error has a median of 3.7 points across five models, against 20.5 for the full-evidence curve alone. On HELMET overall and its NQ subset, the product errs by 4.9–5.2 points and the full-evidence curve by 8.8–9.7. On the PopQA subset the two are similar, and on TriviaQA the full-evidence curve alone is closer. The peaks of the product mostly lie within one grid step of the observed peaks (Table 4). On RepoQA and HELMET overall, the rise thus comes from evidence arriving and the fall from declining use of evidence already present. NoLiMa, whose evidence is complete by about 512 tokens, mostly peaks at that budget and then falls toward zero by about 16K tokens (Figure 2, bottom; Figure 6b).

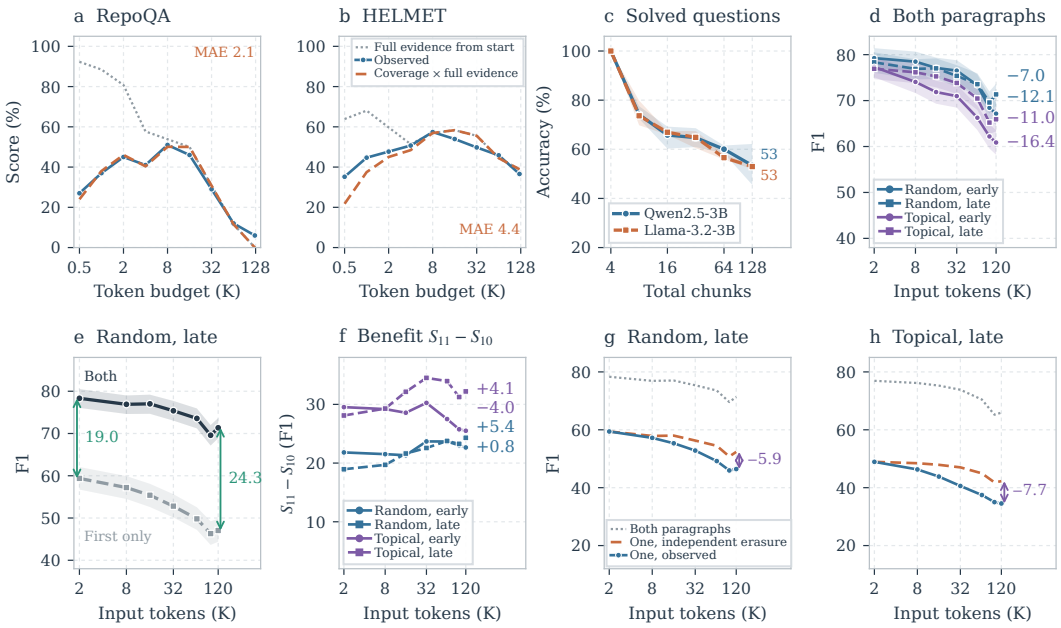

Figure 3: **On RepoQA and HELMET, evidence arrival explains the rise; with evidence in place, performance still falls and evidence is used jointly. (a,b)** For Gemma-3-4B, one of the closer fits (Table 4), the observed curve follows the product of delivered coverage and the score with all evidence present from the start; MAE is the mean absolute error in points. **(c)** LongBench-v2 questions answered correctly from their top four chunks lose about half their accuracy as chunks from other documents are added; bands span three evidence positions. **(d–h)** Qwen2.5-14B on 1,000 HotpotQA questions; S_{ab} is F1 with the first and second evidence paragraphs present according to $a, b \in \{0, 1\}$. (d) With both paragraphs kept, F1 falls in all four settings; labels give the change from 2K to 120K. (e) With only the first paragraph, F1 falls faster, so the missing paragraph matters more at 120K; arrows give $S_{11} - S_{10}$ at 2K and 120K. (f) The benefit of the second paragraph, $S_{11} - S_{10}$, grows in some settings and shrinks in others. (g,h) With one paragraph (averaging S_{10} and S_{01}), F1 falls below the prediction of independent erasure fitted to the two-paragraph score; labels give observed minus predicted F1 at 120K. In (d) and (f), color marks random or topical background text and solid lines with circles or dashed lines with squares place evidence early or late; bands are 95% bootstrap intervals over questions.

The peak is set by how fast use erodes. Models reading the same inputs share their ideal value, so by Corollary 3.4(i) each model's peak moves with the growth of its own gap. On LongBench multi-hop QA, the larger or long-context model in every family peaks at the same or a larger k on the questions available at each budget (Table 3). Qwen3-1.7B, 4B, and 8B peak at 4, 8, and 16; Llama-3.2-1B and 3B at 4 and 8; Gemma-3-4B and 12B at 8 and 32; and Qwen2.5-0.5B, 7B, and 7B-1M at 8, 8, and 16. On LongBench-v2 the ordering fails for Qwen2.5-1.5B, which peaks at the largest budget, after Qwen2.5-3B at $k = 8$. Under controlled arrival, every curve peaks in the interior. Most peak at or before the budget where evidence coverage plateaus (Figure 6a), as Corollary 3.4(ii) predicts. Qwen3-8B on HELMET, Llama-3.2-3B on NoLiMa, and Qwen3.5-4B on RepoQA peak one step later, so their gap shrinks for one step after the evidence is complete. Section 5 holds all evidence in place and asks what is lost.

5 What Is Lost When Evidence Is Kept

We now hold the evidence fixed and vary only what surrounds it. Evidence stays at positions set by the design and background text comes from other documents, so by Proposition 3.3(a) every loss is a loss of use (Appendix B). We ask how large this loss is and where it arises, then whether it acts on each piece of evidence separately.

5.1 Evidence in Place, Performance Lost

Questions already solved lose half their accuracy. On LongBench-v2, we keep the top four chunks of the 156 questions that Qwen2.5-3B answers correctly from them, and of the 129 that Llama-3.2-3B

Table 1: **Background text changes both the loss with complete evidence and the value of a missing paragraph.** ΔS_{11} is the change in HotpotQA F1 with both evidence paragraphs, and $\Delta B_{2|1}$ the change in the benefit of adding the second paragraph to the first ($B_{2|1} = S_{11} - S_{10}$), from 2K to 120K for Qwen2.5-14B (1,000 questions) and to 64K for Llama-3.1-8B (200 questions). Rows are ordered by the Qwen loss. Intervals are in Tables 5, 8 and 10.

		Qwen2.5-14B		Llama-3.1-8B	
Background	Placement	ΔS_{11}	$\Delta B_{2\|1}$	ΔS_{11}	$\Delta B_{2\|1}$
Random	Late	-7.0	$+5.4$	-6.7	$+6.0$
Topical	Late	-11.0	$+4.1$	-8.5	$+3.4$
Random	Early	-12.1	$+0.8$	-13.2	-5.8
Topical	Early	-16.4	-4.0	-20.6	-9.7

answers correctly, then add chunks from other documents. With 124 added chunks, accuracy averages 53.4% and 53.0% over three evidence positions, from a baseline of 100% (Figure 3c; Appendix D.1).

Complete evidence does not ensure successful use. Annotated evidence lets us fix what is supplied without selecting questions by model success. We give Qwen2.5-14B-Instruct-1M (Qwen in this section) 1,000 HotpotQA questions with neither, one, or both of its two annotated evidence paragraphs. Inputs span seven lengths from 2K to 120K with random or lexically topical background text. Evidence sits early, in fixed absolute slots, or late, at a fixed distance from the question (Appendix D). In short inputs, the two paragraphs alone give F1 78.81, against 23.19 for the question alone (Table 7). As context grows from 2K to 120K with both paragraphs kept, F1 falls by 6.97–16.44 points in all four settings, with every paired interval excluding zero (Figure 3d; Table 5). On 200 of the questions, Llama-3.1-8B-Instruct (Llama) loses 6.67–20.55 F1 points from 2K to 64K, with all four losses passing Holm correction (Table 10), and exact match falls in all eight pairs (Table 6).

Topical background text and distant evidence cost most. For both models, topical background text costs more than random background text at each placement. Early evidence likewise costs more than late evidence with each kind of background text (Table 1; paired Qwen contrasts in Table 9). Lexically related background text and distance from the question plausibly make evidence harder to locate, as with the hard negatives of Jin et al. (2025a). Part of the loss is in answer form: among predictions correct at 2K and missed at 64K with both paragraphs, 47% for Llama and 32% for Qwen still contain the answer within a longer prediction (Appendix D.5).

5.2 Evidence Is Used Jointly

If background text acted as a penalty shared by all evidence states, additive or multiplicative, the benefit of adding a missing piece could only stay the same or shrink as context grows (Proposition 3.5); two annotated paragraphs test this.

The benefit of a missing paragraph can grow as performance falls. Let S_{ab} be mean F1 with the first and second paragraphs (in source order) present according to $a, b \in \{0, 1\}$; S_{ab} is $F(s, z)$ of Section 3 for the evidence state $s = (a, b)$. Let $B_{2|1} = S_{11} - S_{10}$ be the benefit of adding the second paragraph. Under random background text with late placement, $B_{2|1}$ grows by 5.35 points from 2K to 120K (95% CI $[2.72, 8.03]$) while S_{11} falls: one paragraph alone loses more, so the missing paragraph matters more in the longer context (Figure 3e). Under topical background text with early placement, the same benefit falls by 4.03 points ($[-7.20, -0.80]$; Figure 3f). The growing benefit violates both restrictions in Proposition 3.5. By Proposition 3.3(a) and Equation (1), the gap grows faster with one paragraph than with two.

A lone paragraph is lost more often than a pair. An independent-erasure model makes this precise. Background text removes each paragraph independently at one rate. We fit this rate to the both-paragraph score and use it to predict the one-paragraph scores with no further fitting. For Qwen, one-paragraph scores fall 4.1–7.7 F1 points below this prediction in all four settings (Figure 3g,h). The rate at which a lone paragraph is lost is 1.4–2.2 times the fitted rate, with every 95% interval above one. Exact match gives the same result (Appendix D.4).

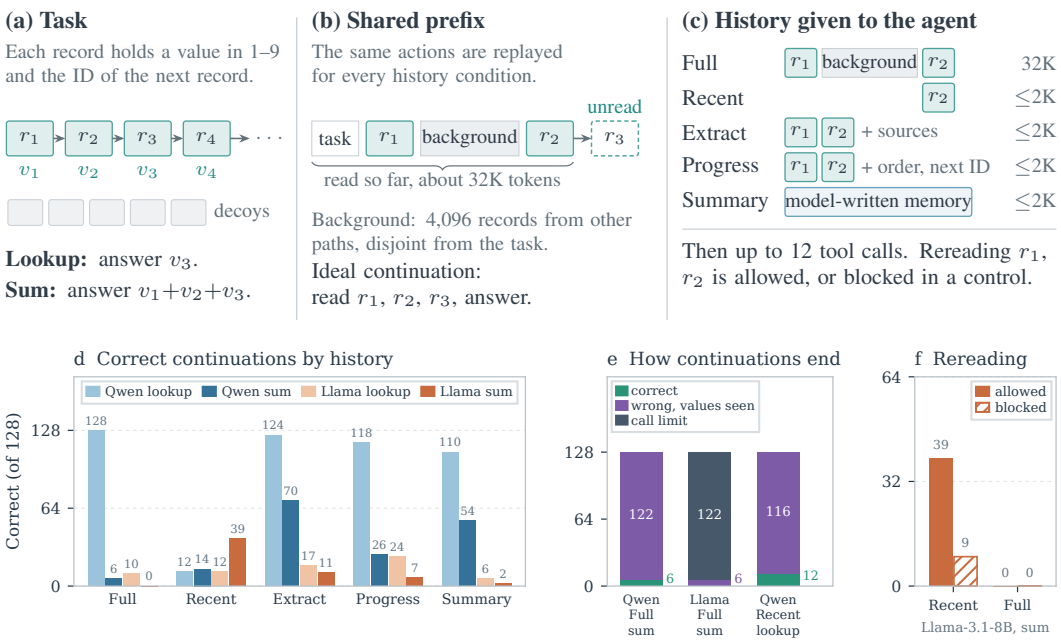

Figure 4: **Rewriting the same history changes whether the agent succeeds.** **(a)** Each instance is a chain of linked records; lookup asks for the third record's value and summation for the sum of the first three. **(b,c)** After the shared prefix the agent has read r_1, about 32K tokens of background records, and r_2; Full keeps this context, four policies compress it, and the agent continues for at most twelve tool calls. **(d)** Correct answers out of 128 instances; writer failures under Summary count as errors. **(e)** How three continuations fail: with Full summation Qwen has the needed values but answers wrongly, with Recent lookup it reads past r_3 and returns v_4, and with Full summation, Llama mostly runs out of calls. **(f)** Llama summation with rereading of r_1, r_2 allowed or blocked. Paired contrasts are in Appendix E.

The benefit grows where complete evidence is most robust. For Qwen, ordering the four settings by how much complete evidence loses reverses their order by benefit change (Table 1). The Llama estimates follow the same order, although none of its benefit changes pass Holm correction (Table 10). One paragraph alone loses a similar amount in every setting, 12–15 points for Qwen, so the sign of $\Delta B_{2|1}$ follows how well complete evidence resists background text.

For Qwen2.5-14B, the value of a piece of evidence therefore depends on its companions: a complete set protects each of its members. Section 6 asks whether use lost in this way can be recovered without adding new information.

6 Managing Agent Histories

If context growth makes evidence harder to use, rewriting the context should be able to make it easier to use again without adding information. Agents allow a direct test: their context is a history that can be managed and, through tools, revisited. By Proposition 3.3(b) and Equation (2), when every history keeps earlier observations recoverable, differences between histories measure use alone.

A shared prefix with rereading. We evaluate Qwen2.5-7B-Instruct-1M and Llama-3.1-8B-Instruct on 128 synthetic task instances; we write Qwen and Llama for these two models in this section. Each instance is one draw of the world W in Definition 3.1: a generated corpus with one linked path of records, sixteen decoys (records that look like path records but are not linked to it), and 4,096 background records from other paths. Each task uses the first three records on a longer linked path: lookup returns the third value, while summation adds all three. A shared prefix reads record 1, background records, and record 2, leaving record 3 unread (Figure 4a–c).

Full retains this history, at about 32K tokens, and four policies compress it to within 2K. *Recent* keeps the latest complete interactions. *Extract* retains the observed task records and their sources. *Progress* adds reading order and the ID of the next unread record. *Summary* asks the same model to write a working memory. Extract and Progress add no future values or computed answers. The agent

keeps the task, the start ID (the ID of record 1), the tools, and a limit of twelve tool calls. All failures count, including writer failures under Summary (Appendix E.1). With rereading allowed, an ideal continuation from any history can reread r_1 and r_2, read r_3, and answer in four calls (Appendix B). Thus $\Delta G = 0$ for every history the agent receives, so each difference from Full is a change in use.

Extraction recovers use of the same observations. Extract raises Qwen summation from 6/128 with Full to 70/128, a paired gain of 50.0 percentage points (95% CI $[41.4, 58.6]$; Holm $p < 10^{-17}$). Llama rises from 0/128 to 11/128 ($+8.59$ points, $[3.91, 13.28]$; Holm $p = .01$; Figure 4d). The Full continuations show that delivering the evidence is not enough (Figure 4e). All 122 incorrect Qwen Full summation runs deliver all three values before answering. Llama reaches the twelve-call limit in 122 of 128 instances, including 119 with all three values delivered. No history is best for both models: Llama summation succeeds most often with Recent (39/128). Summary, which asks the model to write its own memory, reaches 54/128 on Qwen summation, but 195 of Llama's 256 memory writes fail, so its Llama scores mostly reflect the writer (Appendix E.2).

Not knowing where to stop: the agent reads past the target. Recent keeps record 2 and its pointer to record 3 but removes the step from record 1 to record 2. The start ID still lets the agent recover its position, so ideal value is unchanged. Because the third record is not marked as the end of the underlying path, the agent must know how far it has gone. In every Qwen lookup instance, it reads the intended third record and then a fourth, returning the fourth value (Figure 4e). Its twelve correct answers occur exactly when the third and fourth values coincide. Progress fails on summation the same way: 111 of its 128 Qwen runs read on after the third record, against 40 for Extract, and only 15 of those runs succeed.

Blocking rereads removes information. When rereads are blocked under Recent, record 1 can no longer be recovered, while record 2 stays in the history and the unread third record stays accessible. Llama Recent summation falls from 39/128 with rereading to 9/128 without it, while Full stays at 0/128. Recent's allowed-minus-blocked change exceeds Full's by 23.44 points (95% CI $[16.41, 31.25]$; Holm $p < 10^{-7}$; Figure 4f). In the thirty recovered instances, both continuations request an earlier record. Only the allowed one receives it and succeeds (Appendix E.2).

7 Discussion

Rereading counts as preservation. An agent's information includes what it can still retrieve, so a short history that allows rereading can beat a full one (Section 6), and context management must weigh both what a history keeps and how to restore what it drops.

Measuring use directly. Long-context benchmarks can report scores with complete evidence retained as context grows, and agent memory evaluations can check that removed observations stay recoverable. A drop with complete evidence retained is then a loss of use, and a gain from a rewritten history that still allows rereading is use recovered. Scores should also separate content from form, since 32–47% of the exact-match misses we examined still contain the answer.

Implications for context engineering. Because Qwen2.5-14B loses a lone paragraph more often than a pair, selection should keep complementary evidence together, which retrievers that score passages one at a time may miss. Both models lose less with late evidence and random background text, so placement and filtering matter even when retrieval is correct.

8 Conclusion

We asked whether long context fails because information is lost or because it becomes hard to use. Comparing a model with an ideal solver that reads the same input separates the two. In our controlled settings the answer is use: performance falls with all evidence kept, evidence is used jointly, and rewriting an agent's history can make the same observations easier to use. Context quality depends on how well models find, combine, and act on the information available to them.

AI USE STATEMENT

In this work, we used generative AI tools to implement experiment and analysis code, run and analyze experiments, help develop and check the theoretical framework and proofs, produce figures and tables, and draft and edit the manuscript. We have not used generative AI tools to generate or alter experimental data: benchmark inputs come from existing datasets, the synthetic agent task instances are generated programmatically, and all reported scores come from model outputs. The authors reviewed and verified all AI-assisted code, results, and text, and take full responsibility for the paper.

REFERENCES

Anthropic Applied AI Team. Effective Context Engineering for AI Agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents, September 2025. Accessed: 2026-05-25.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-Context Multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, Sai Surya Duvvuri, Fnu Devvrit, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Kale, and Samy Jelassi. Let's (not) just put things in Context: Test-time Training for Long-context LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=H0bcEdPCoc.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R Gormley. Oolong: Evaluating Long Context Reasoning and Aggregation Capabilities. *arXiv preprint arXiv:2511.02817*, 2025.

David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, June 1953. ISSN 0003-4851. doi: 10.1214/aoms/1177729032. URL http://dx.doi.org/10.1214/aoms/1177729032.

Weili Cao, Jianyou Wang, Youze Zheng, Longtian Bao, Qirui Zheng, Taylor Berg-Kirkpatrick, Ramamohan Paturi, and Leon Bergen. Single-Pass Document Scanning for Question Answering. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=7Vj78acKIp.

Weili Cao, Xunjian Yin, Bhuwan Dhingra, and Shuyan Zhou. Coding Agents are Effective Long-Context Processors. *arXiv preprint arXiv:2603.20432*, 2026.

Zhuo Chen, Oriol Comas, Zhuotao Jin, Di Luo, and Marin Soljacic. L^2M: Mutual Information Scaling Law for Long-Context Language Modeling. *Advances in Neural Information Processing Systems*, 38:138466–138506, 2026.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. *arXiv preprint arXiv:2508.06600*, 2025.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. CL-bench: A Benchmark for Context Learning. *arXiv preprint arXiv:2602.03587*, 2026.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context Length Alone Hurts LLM Performance Despite Perfect Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298. Association for Computational Linguistics, 2025.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Ziyuan He, Yuxuan Wang, Jiaqi Li, Kexin Liang, and Muhan Zhang. LooGLE v2: Are LLMs Ready for Real World Long Dependency Challenges? *arXiv preprint arXiv:2510.22548*, 2025.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context Rot: How Increasing Input Tokens Impacts LLM Performance. Technical report, Chroma, July 2025. URL https://research.trychroma.com/context-rot.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Yulong Hui, Chao Chen, Zhihang Fu, Yihao Liu, Jieping Ye, and Huanchen Zhang. Interact-RAG: Reason and Interact with the Corpus, Beyond Black-Box Retrieval. *arXiv preprint arXiv:2510.27566*, 2025.

Malika Iratni, Mohand Boughanem, and Taoufiq Dkaki. Dynamic Context Selection for Retrieval-Augmented Generation: Mitigating Distractors and Positional Bias. *arXiv preprint arXiv:2512.14313*, 2025.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Arik. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. In *International Conference on Learning Representations*, volume 2025, pp. 37784–37822, 2025a.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516*, 2025b.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs Get Lost in Multi-Turn Conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Moshe Lahmy and Roi Yozevitch. Replace, Don't Expand: Mitigating Context Dilution in Multi-Hop RAG via Fixed-Budget Evidence Assembly. *arXiv preprint arXiv:2512.10787*, 2025.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. Long Context RAG Performance of Large Language Models. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024. URL https://openreview.net/forum?id=Le9anH3kv1.

Xingyu Li, Rongguang Wang, Yuying Wang, Mengqing Guo, Chenyang Li, Tao Sheng, Sujith Ravi, and Dan Roth. PAR^2-RAG: Planned Active Retrieval and Reasoning for Multi-Hop Question Answering. In *The 64th Annual Meeting of the Association for Computational Linguistics – Industry Track*, 2026. URL https://openreview.net/forum?id=wF68cQHM1T.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 881–893, 2024.

Grace Liu, Yuxiao Qu, Jeff Schneider, Aarti Singh, and Aviral Kumar. CaRT: Teaching LLM Agents to Know When They Know Enough. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025. URL https://openreview.net/forum?id=kkKsmPeasY.

Jiawei Liu, Jia Le Tian, Vijay Daita, Yuxiang Wei, Yifeng Ding, Yuhan Katherine Wang, Jun Yang, and Lingming Zhang. RepoQA: Evaluating Long Context Code Understanding. *arXiv preprint arXiv:2406.06025*, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024b.

Dehai Min, Kailin Zhang, Tongtong Wu, and Lu Cheng. QuCo-RAG: Quantifying Uncertainty from the Pre-training Corpus for Dynamic Retrieval-Augmented Generation. *arXiv preprint arXiv:2512.19134*, 2025.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-Context Evaluation Beyond Literal Matching. *arXiv preprint arXiv:2502.05167*, 2025.

Kyle Montgomery, David Park, Jianhong Tu, Michael Bendersky, Beliz Gunel, Dawn Song, and Chenguang Wang. Predicting Task Performance with Context-aware Scaling Laws. In *Knowledgeable Foundation Models at ACL 2025*, 2025. URL https://openreview.net/forum?id=131V12fEUR.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-RAG: Value-Based Retrieval Control for Iterative RAG. In *First Workshop on Multi-Turn Interactions in Large Language Models*, 2025. URL https://openreview.net/forum?id=fN4PJNYgs3.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Yujia Zhou, Xu Chen, and Zhicheng Dou. Are Long-LLMs A Necessity For Long-Context Tasks? *arXiv preprint arXiv:2405.15318*, 2024.

Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. DynamicRAG: Leveraging Outputs of Large Language Model as Feedback for Dynamic Reranking in Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.07233*, 2025a.

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. Scaling Long-Horizon LLM Agent via Context-Folding. *arXiv preprint arXiv:2510.11967*, 2025b.

Chihiro Taguchi, Seiji Maekawa, and Nikita Bhutani. Efficient Context Selection for Long-Context QA: No Tuning, No Iteration, Just Adaptive-k. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20116–20141, 2025.

Zecheng Tang, Baibei Ji, Juntao Li, Lijun Wu, Haijia Gui, and Min Zhang. Revisiting Long-context Modeling from Context Denoising Perspective. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xvGyyh6MG7.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. Chain-of-Retrieval Augmented Generation. *arXiv preprint arXiv:2501.14342*, 2025.

Roy Xie, Junlin Wang, Paul Rosu, Chunyuan Deng, Bolun Sun, Zihao Lin, and Bhuwan Dhingra. Knowing When to Stop: Efficient Context Processing via Latent Sufficiency Signals. *Advances in Neural Information Processing Systems*, 2025.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval Meets Long Context Large Language Models. In *International Conference on Learning Representations*, volume 2024, pp. 49569–49584, 2024.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A Theory of Usable Information Under Computational Constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. Knowing You Don't Know: Learning When to Continue Search in Multi-round RAG through Self-Practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1305–1315, 2025.

Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, et al. AgentFold: Long-Horizon Web Agents with Proactive Context Management. *arXiv preprint arXiv:2510.24699*, 2025a.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=ruWC5LIMSo.

Xi Ye, Wuwei Zhang, Fangcong Yin, Howard Yen, and Danqi Chen. DySCO: Dynamic Attention-Scaling Decoding for Long-Context LMs. *arXiv preprint arXiv:2602.22175*, 2026.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.

Alex L Zhang, Tim Kraska, and Omar Khattab. Recursive Language Models. *arXiv preprint arXiv:2512.24601*, 2025.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. CodeAgent: Enhancing Code Generation with Tool-Integrated Agent Systems for Real-World Repo-Level Coding Challenges. *arXiv preprint arXiv:2401.07339*, 2024.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding Synthetic Context Extension via Retrieval Heads. In *International Conference on Machine Learning*, pp. 77885–77910. PMLR, 2025.

Appendix

A Appendix Overview

Appendix B proves the results of Section 3. The empirical appendices follow the three main-text findings: Appendix C supports Section 4, Appendix D supports Section 5, and Appendix E supports Section 6. Request and action interfaces are in Appendix F.

Table 2: **Where each main-text claim is supported.**

Main-text claim	Design	Appendix
The peak comes from arrival and use (Section 4)	Natural retrieval with 13 models (11 on LongBench multi-hop QA); fixed cohorts of 211 LongBench-v2 and 1,446 LooGLE-v2 questions; controlled arrival on three benchmarks with five models.	Appendix C
Evidence kept, performance lost (Section 5.1)	Top-four-solved cohorts of 156 and 129 questions; 1,000 HotpotQA questions with four evidence states, seven lengths, two kinds of background text, and two placements; 200 questions for Llama-3.1-8B.	Appendices D.1, D.2, D.3 and D.5
Evidence is used jointly (Section 5.2)	The same HotpotQA factorial; independent-erasure null.	Appendix D.4
Managing histories (Section 6)	128 task instances shared across two models and two tasks.	Appendix E

Shared reporting conventions. Length limits, writer failures, and call limits count as failures. Paired analyses resample whole questions or instances.

B Formal Results and Proofs

Task, observations, and the common decision problem. The task τ fixes utility $u_\tau(W, a)$, what counts as a correct answer or successful action in world W. The joint law of W and X specifies what can be inferred from the observations and is held fixed when comparing solvers. The ideal solver sees the observations, not the hidden world, and has no processing limitation. Different models can therefore reach different $U_\tau^{\mathcal{M}}$ under the same ideal value G_τ.

Proof of Lemma 3.2. If $X = f(X')$ almost surely, every rule π using X can be reproduced by $\pi \circ f$ using X'. Taking suprema over decision rules gives $G(X') \geq G(X)$. The model rule is among these rules, so $D^{\mathcal{M}} \geq 0$. Subtracting $U^{\mathcal{M}} = G - D^{\mathcal{M}}$ between the two conditions gives the stated identity. If $\Delta U^{\mathcal{M}} = -\delta$, then $\Delta D^{\mathcal{M}} = \Delta G + \delta \geq \delta$. If recovery is possible in both directions, applying the same argument twice gives $G(X') = G(X)$, hence $U^{\mathcal{M}}(X') - U^{\mathcal{M}}(X) = D^{\mathcal{M}}(X) - D^{\mathcal{M}}(X')$. □

The utilization gap includes all unrealized value, including difficulty already present in a short context.

Proof of Proposition 3.3. Recoverability means $E = r(X)$ almost surely for some rule r, while $X = \Pi(E, Z)$. Any rule $\pi(E, Z)$ can be replicated by a rule that observes only E and samples $\widetilde{Z}$ from the conditional law of Z given E, using independent randomness. Conditional independence makes $(W, E, \widetilde{Z})$ and (W, E, Z) equal in law, so both rules have the same expected utility and $G(E, Z) \leq G(E)$. The reverse inequality follows by ignoring Z. Thus $G(E) \leq G(X) \leq G(E, Z) = G(E)$, by composition with r and Π. Holding the task and law of (W, E) fixed makes G constant across background-text regimes, so $F(s, z) = G(s) - D(s, z)$ gives Equation (1). For (b), any rule using $\widetilde{X} = T(X)$ is a rule using X together with independent randomness, so $G(\widetilde{X}) \leq G(X)$; when the two ideal values are equal, subtracting $U^{\mathcal{M}} = G - D^{\mathcal{M}}$ gives the stated identity. □

Recoverability uses the constructions' designated text blocks. In the fixed-evidence study, the retained chunks occupy the beginning, middle, or end block named by the condition. In the natural-QA factorial, early evidence occupies fixed absolute slots and late evidence sits a fixed distance before the question; in NoLiMa, the designated evidence sits at the start. For random and topical

background text alike, equality additionally assumes conditional independence given the question and supplied evidence.

Proof of Corollary 3.4. (i) On the same inputs g_k is common to both models, so the step changes in performance satisfy $u_k^A = g_k - d_k^A \le g_k - d_k^B = u_k^B$. Let $k_B = k_B^\star$. For $j > k_B$, $U^B(X_j) - U^B(X_{k_B}) = \sum_{k_B < k \le j} u_k^B < 0$ because k_B is the largest maximizer. Then $U^A(X_j) - U^A(X_{k_B}) \le U^B(X_j) - U^B(X_{k_B}) < 0$, so no $j > k_B$ maximizes U^A, and $k_A^\star \le k_B$. (ii) For $k > k_s$, every increment $-d_k$ is non-positive. □

Neither part assumes a single crossing or concavity.

Proof of Proposition 3.5. Write $\Delta h = h(s^+) - h(s)$. Under the additive model, $B(z) = \Delta h$ regardless of z, proving Equation (3). Under shared attenuation, $B(z) = a(z)\Delta h$, so

$$B(z') - B(z) = [a(z') - a(z)]\Delta h \le 0,$$

proving Equation (4). □

Both scores can fall while their difference grows if the partial-evidence score falls further.

Managed histories and recovery. Fix the joint distribution of world and history after the shared prefix, tool dynamics, permissions, and remaining call limit. Here $G(X)$ is the supremum of expected terminal score over policies observing initial representation X and subsequent permitted tool responses. A full-history policy can generate $\widetilde{H} = T(H)$ and simulate a managed-history policy, establishing $\Delta G = G(\widetilde{H}) - G(H) \le 0$. Substituting $U^{\mathcal{M}} = G - D^{\mathcal{M}}$ gives Equation (2).

Decision sufficiency depends on the continuation. For example, two equally likely histories may retain the same record while requiring different next actions: stopping in one and continuing in the other. If the representation merges them and no further informative observation is possible before acting, any policy errs with probability at least $1/2$, although full history permits the correct action. A record-presence check therefore need not establish decision sufficiency.

Conversely, an admissible recovery strategy can make $\Delta G = 0$ if it restores a decision-sufficient statistic before it is needed, preserves the relevant environment state, and leaves enough actions to attain the full-history ideal value within the original call limit.

In the shared-prefix study, these conditions hold constructively for every history with rereading allowed. Each retains the original task and start ID. Read the start record, follow its pointer to record 2, follow that pointer to record 3, and answer with the third value or the sum of the three values. This policy uses four agent calls within the twelve-call limit. The construction checks that each record is displayed without truncation and that the complete continuation fits the input limit. It succeeds in every generated instance under Full and Recent; the compact Extract, Progress, and Summary inputs fit within the same 2,048-token initial ceiling as Recent, so the same route applies. Every ideal value therefore equals one and $\Delta G = 0$. This argument concerns histories the agent receives, not writer failures that end a run before the agent starts.

C Where the Peak Comes From

C.1 Natural Retrieval

Documents are split into sentence-aware chunks targeting 512 tokens, ranked by query similarity with `bge-large-en-v1.5`, and restored to document order after selection. LongBench-v2 scores multiple-choice answers on 503 questions. The generation panel combines HotpotQA, 2WikiMultihopQA, and MuSiQue and uses token F1. LooGLE-v2 extends the retrieval grid to larger budgets. For the generation panel, documents with fewer than k chunks leave the cohort at budget k; for LongBench-v2, requested budgets beyond a document are clipped; Qwen2.5-3B and Llama-3.2-3B keep all 503 questions, while Qwen2.5-1.5B and 7B have 487 at $k = 32$. LongBench-v2 inputs are capped at 120,000 tokens by middle truncation.

Table 3: **Models in the generation panel of Figure 2.** Color and marker identify each curve; within a family, the legend runs from the smallest to the largest model, then the long-context variant. n gives the questions with valid outputs at $k = 1$ and $k = 32$; documents shorter than k chunks leave the cohort.

Model	Color	Marker	$n_{k=1}$	$n_{k=32}$
Qwen2.5-0.5B-Instruct	Light blue	Down triangle	600	293
Qwen2.5-7B-Instruct	Blue	Circle	600	293
Qwen2.5-7B-Instruct-1M	Dark blue	Diamond	600	293
Qwen3-1.7B	Light orange	Down triangle	522	251
Qwen3-4B-Instruct-2507	Orange	Square	600	293
Qwen3-8B	Dark orange	Up triangle	600	293
Llama-3.2-1B-Instruct	Light green	Down triangle	600	293
Llama-3.2-3B-Instruct	Green	Up triangle	600	267
Gemma-3-4B-it	Light purple	Square	600	293
Gemma-3-12B-it	Purple	Circle	438	115
Mistral-Nemo-Instruct-2407	Brown	Pentagon	600	293

The generation grid is $k \in \{1, 2, 4, 8, 16, 32\}$. Scores pool question-level token F1 across HotpotQA, 2WikiMultihopQA, and MuSiQue. We score the first line of the greedy response against the reference answers, taking the maximum reference F1.

C.2 Question-Paired Retrieval Curves

Increasing the retrieval budget can change which questions remain evaluable. We separate this cohort effect from within-question changes by retaining only questions with every recorded budget available.

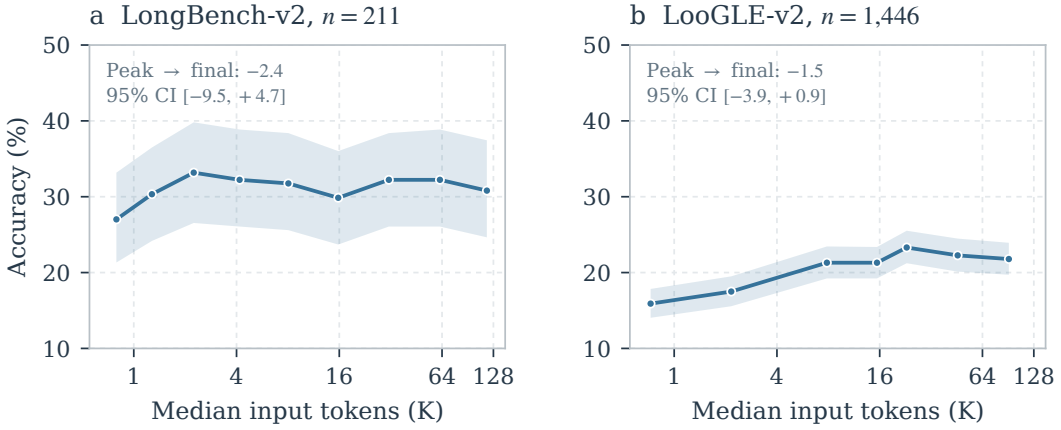

Figure 5: **On fixed question cohorts, retrieval curves rise and then level off.** Qwen2.5-3B on the 211 LongBench-v2 and 1,446 LooGLE-v2 questions that have every budget. The x-axis is median input length; bands are 95% bootstrap intervals over questions, and labels give the paired peak-to-end change.

Cohort composition. Budgets beyond a document's length are clipped in the LongBench-v2 and LooGLE-v2 curves of Figure 2; here we keep only questions whose documents are long enough that no budget is clipped. LongBench-v2 keeps 211 of its 503 questions with all nine budgets, and LooGLE-v2 keeps 1,446 of 1,934 with all seven. At $k = 256$, 194 of the 211 LongBench-v2 questions reach the 120,000-token cap.

Within-question curves. LongBench-v2 rises from 27.01% at $k = 1$ to 33.18% at $k = 4$ and ends at 30.81% at $k = 256$, a paired change of -2.4 points from the peak (95% CI $[-9.5, 4.7]$). LooGLE-v2 rises from 15.91% at $k = 1$ to 23.31% at $k = 48$ and ends at 21.78% at $k = 192$, a paired change of -1.5 points ($[-3.9, 0.9]$).

C.3 Controlled Evidence Arrival

The constructions use HELMET KILT RAG, NoLiMa Book1, and RepoQA Python, with 72 questions, 58 tests, and 100 tasks, and five models: Qwen2.5-3B, Qwen3-8B, Qwen3.5-4B, Llama-3.2-3B, and

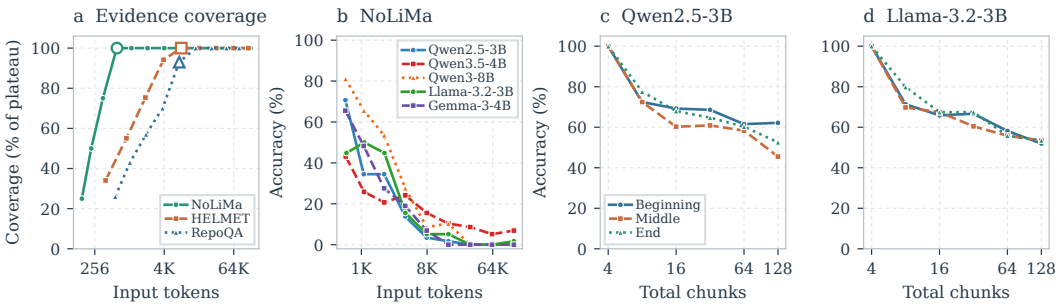

Figure 6: **Evidence arrives early, while accuracy with evidence retained continues to fall. (a)** Delivered evidence coverage reaches its plateau by 16K tokens on every benchmark; curves are normalized to the plateau, and hollow markers show each benchmark's median accuracy peak (Figure 2, bottom). **(b)** With all evidence at the start, NoLiMa accuracy falls toward zero for every model. **(c,d)** LongBench-v2 questions solved from their top four chunks lose accuracy at every evidence position as background text is added (156 Qwen2.5-3B and 129 Llama-3.2-3B questions; the 100% baseline follows from selection).

Gemma-3-4B. HELMET inserts evidence at ranks $\{1, 2, 4, 8, 16, 32\}$ in its stream of background text. NoLiMa varies the arrival budget of its evidence; its full-evidence curve places the evidence at the start (Figure 6b). RepoQA assigns an arrival budget from $\{512, 1\text{K}, 2\text{K}, 4\text{K}, 8\text{K}, 16\text{K}\}$, masks the target function in shorter prefixes, and restores it at and after that budget. Evidence remains supplied in longer contexts. Figure 6a normalizes coverage by each construction's plateau (HELMET's unnormalized plateau is 0.3194); hollow markers interpolate each benchmark's median accuracy peak on its coverage curve.

Arrival and use. For each controlled-arrival curve, the separable prediction multiplies delivered coverage, normalized to its plateau, by the score measured with all designated evidence present from the start at the same budget. Table 4 compares it with that full-evidence curve alone. RepoQA's full-evidence curve is measured on 26 of its 100 tasks. NoLiMa delivers all evidence by 512 tokens and is omitted, as is HELMET's HotpotQA subset, which has no annotated evidence coverage.

Table 4: **The product of coverage and the full-evidence score fits the arrival curves better than the full-evidence score alone on RepoQA and HELMET overall.** Entries are the median over five models of the mean absolute error in points, and how many of the five models have a predicted peak within one grid step.

Benchmark	Coverage × full-evidence	Full-evidence alone	Peak within one step
RepoQA	3.7	20.5	4/5
HELMET, all	5.2	8.8	5/5
HELMET, NQ	4.9	9.7	3/5
HELMET, PopQA	8.6	9.1	5/5
HELMET, TriviaQA	9.4	7.5	4/5

D What Is Lost When Evidence Is Kept

D.1 Fixed Evidence on Solved Questions

Questions are selected for a correct answer from their top four retrieved chunks. Those chunks remain present while $N \in \{0, 4, 12, 28, 60, 124\}$ background chunks are added, with evidence at the beginning, middle, or end. Qwen2.5-3B has 156 selected questions and Llama-3.2-3B has 129. The 100% baseline follows from selection. Each question is the independent unit; positions share a background draw and larger background sizes extend the same prefix.

D.2 Natural-QA Factorial

Design. Qwen2.5-14B-Instruct-1M is evaluated on 1,000 HotpotQA questions. Four evidence states cross seven lengths, two kinds of background text, and two placements, giving 112 conditions per question. Question-only and evidence-only controls each occur once. Evidence bits identify the two annotated paragraphs in source order.

Background text and placement. The nominal targets are 2/8/16/32/64/96/120 K tokens, with K $= 1{,}024$; plots use actual mean input length. Early placement fixes absolute evidence slots, while late placement fixes distance to the final question. Background text is nested across lengths, and removing an evidence paragraph fills its slot with a paragraph from the background-text stream. Topical matching is lexical, defining a sampling procedure rather than a calibrated semantic distance.

Scores and uncertainty. Figure 7 gives every evidence state under normalized token F1 and exact match (EM). Lines connect measured means. The question is the resampling unit across all conditions: curve intervals use 2,000 bootstrap draws, while the paired contrast tables use 10,000 shared draws. Both are pointwise 95% percentile intervals without multiplicity correction. With both paragraphs, 107–188 of the 1,000 questions switch from a correct to an incorrect exact answer between 2K and 120K across the four settings, and 28–36 switch the other way.

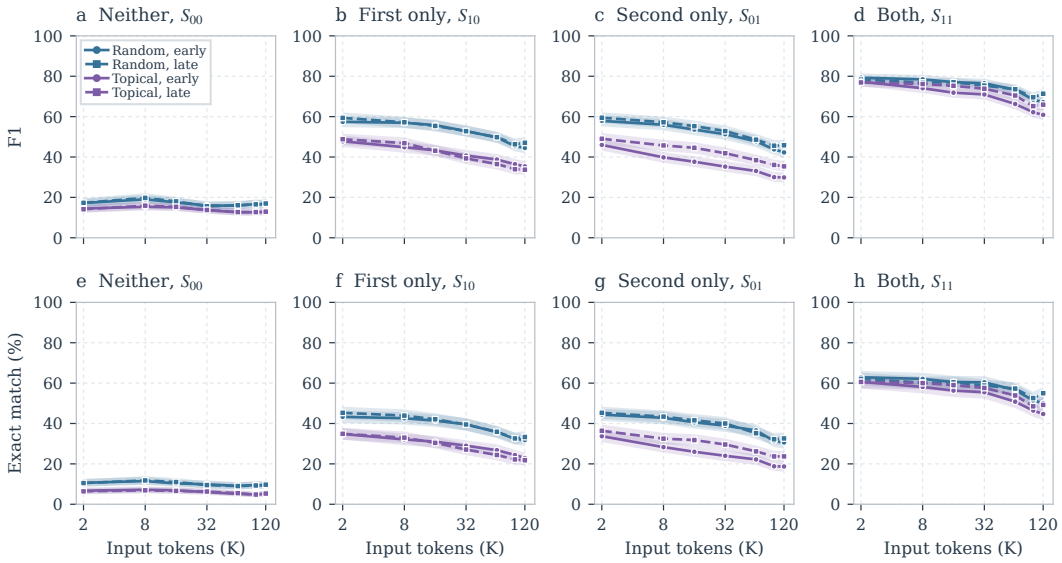

Figure 7: **Every state with evidence loses F1 and exact match as context grows; the question-only state stays near its floor.** Qwen2.5-14B on 1,000 HotpotQA questions. Columns are the evidence states S_{00}, S_{10}, S_{01}, and S_{11}; rows are F1 (top) and exact match (bottom). Color and line style mark the four background and placement settings; bands are 95% bootstrap intervals over questions.

Table 5: **With both paragraphs kept, F1 falls in every setting from 2K to 120K.** Qwen2.5-14B on 1,000 questions; intervals are 95% paired bootstrap intervals over questions.

Background	Placement	2K	120K	Change [95% CI]
Random	Early	79.25	67.15	-12.10 $[-14.20, -10.02]$
Random	Late	78.33	71.35	-6.97 $[-8.86, -5.08]$
Topical	Early	77.31	60.87	-16.44 $[-18.78, -14.06]$
Topical	Late	76.93	65.93	-11.00 $[-13.12, -8.94]$

Table 6: **Exact match also falls with both paragraphs kept, for both models.** Qwen2.5-14B on 1,000 questions from 2K to 120K and Llama-3.1-8B on 200 questions from 2K to 64K; changes are in percentage points with 95% paired bootstrap intervals.

Model	Background / placement	EM at 2K	EM long	Change [95% paired CI]
Qwen2.5-14B (120K)	Random / early	62.90	50.10	-12.80 $[-15.40, -10.30]$
	Random / late	62.10	55.00	-7.10 $[-9.40, -4.80]$
	Topical / early	60.60	44.70	-15.90 $[-18.60, -13.20]$
	Topical / late	60.60	49.20	-11.40 $[-14.00, -8.90]$
Llama-3.1-8B (64K)	Random / early	58.00	45.00	-13.00 $[-19.50, -7.00]$
	Random / late	57.50	51.50	-6.00 $[-11.00, -1.50]$
	Topical / early	55.50	34.00	-21.50 $[-28.00, -15.00]$
	Topical / late	53.50	47.00	-6.50 $[-12.00, -1.00]$

Table 7: **Both models use complete evidence well in short inputs.** HotpotQA exact match and F1 without added background text, on the questions of the length experiments. The evidence-only input contains the question and both paragraphs.

		Question only		Evidence only	
Model	Questions	EM	F1	EM	F1
Qwen2.5-14B	1,000	13.80	23.19	62.10	78.81
Llama-3.1-8B	200	9.50	15.73	59.50	79.69

Conditional evidence benefit. Table 8 gives the change in the benefit of the second paragraph in all four settings. For example, the change in $S_{11} - S_{10}$ is positive under random background text with late placement but negative under topical background text with early placement.

Table 8: **The benefit of the second paragraph grows with late placement and shrinks under topical background text with early placement.** Change from 2K to 120K in $S_{11} - S_{10}$, the F1 gained by adding the second paragraph to the first, for Qwen2.5-14B on 1,000 questions, with 95% paired bootstrap intervals.

Background / placement	Change in $S_{11} - S_{10}$ [95% CI]
Random / early	$+0.84$ [-1.94, $+3.70$]
Random / late	$+5.35$ [$+2.72$, $+8.03$]
Topical / early	-4.03 [-7.20, -0.80]
Topical / late	$+4.10$ [$+1.03$, $+7.12$]

Table 9: **At 120K, topical background text and early placement each lower F1 with complete evidence; at 2K the differences are small.** Paired F1 differences in points for Qwen2.5-14B on 1,000 questions, with 95% bootstrap intervals over questions.

Paired contrast	2K estimate [CI]	120K estimate [CI]	Paired n
S_{11}: topical $-$ random, early	-1.94 [-3.44, -0.47]	-6.28 [-8.23, -4.37]	1,000
S_{11}: topical $-$ random, late	-1.40 [-2.78, -0.02]	-5.42 [-7.29, -3.61]	1,000
S_{11}: late $-$ early, random	-0.93 [-2.14, $+0.31$]	$+4.20$ [$+2.35$, $+6.12$]	1,000
S_{11}: late $-$ early, topical	-0.39 [-1.75, $+0.97$]	$+5.06$ [$+2.72$, $+7.34$]	1,000

D.3 LLAMA-3.1-8B

Llama-3.1-8B-Instruct is evaluated on 200 questions drawn at random from the same 1,000, among those whose complete 2K/64K factorial fits its tokenizer with both evidence paragraphs intact at 2K. Each question has 32 factorial conditions and two controls, with no missing cells. Greedy decoding uses temperature 0, top-p 1, repetition penalty 1 and a 64-token output limit. Wrong, empty and length-limited responses keep the standard normalized EM/F1 scoring.

The Holm family contains 12 F1 contrasts: within each background/placement, the 64K-minus-2K full-evidence change and the corresponding changes in the benefit of each paragraph given the other. Each contrast is formed within question. Pointwise 95% intervals use 10,000 question-bootstrap draws; two-sided one-sample t-tests on the paired differences are Holm-corrected across the 12 contrasts. All four full-evidence losses pass this correction (adjusted $p \leq 0.0075$); the eight conditional-benefit changes do not (smallest adjusted $p = 0.0543$). Table 10 reports the full-evidence changes and the changes in $B_{2|1}$.

D.4 INDEPENDENT ERASURE

Let a_{ab} be the 2K value of the score S_{ab}. The null uses one erasure probability e across questions and paragraphs within each setting, with independent erasure events. It predicts $S_{11} = (1-e)^2 a_{11} + e(1-e)(a_{10} + a_{01}) + e^2 a_{00}$ and $S_{10} = (1-e)a_{10} + e a_{00}$, with S_{01} analogous. We solve e from the observed S_{11} at the long length and predict the one-paragraph scores. Table 11 reports observed minus predicted F1 and the ratio of the rate implied by each one-paragraph score to e, averaging both quantities over S_{10} and S_{01}. Intervals use 2,000 paired question-bootstrap draws. Exact match gives the same conclusions: the rate ratios are 1.26–1.97, with every 95% interval above one.

Table 10: **For Llama-3.1-8B, complete evidence also loses F1, while the benefit of the second paragraph varies in sign.** Changes from 2K to 64K in F1 points on 200 questions, with 95% paired bootstrap intervals; † marks $p < 0.05$ after Holm correction across 12 F1 contrasts.

Background / placement	ΔS_{11}	$\Delta B_{2\|1}$
Random / early	$-13.16^{\dagger}$ $[-18.67, -7.74]$	-5.84 $[-12.67, 0.75]$
Random / late	$-6.67^{\dagger}$ $[-10.51, -2.92]$	$+6.01$ $[0.08, 11.82]$
Topical / early	$-20.55^{\dagger}$ $[-26.17, -15.14]$	-9.68 $[-16.64, -2.95]$
Topical / late	$-8.48^{\dagger}$ $[-12.96, -4.25]$	$+3.40$ $[-3.24, 9.95]$

Table 11: **A lone paragraph is lost 1.4–2.2 times as often as independent erasure predicts.** Qwen2.5-14B on 1,000 questions at 120K. The ratio compares the erasure rate implied by one-paragraph F1 with the rate fitted to two-paragraph F1; the last column is observed minus predicted one-paragraph F1 in points. Both average the two one-paragraph states; intervals are 95% paired bootstrap intervals.

Setting	Implied-rate ratio	Observed − predicted
Random, early	1.40 [1.21, 1.62]	-4.1 $[-5.7, -2.4]$
Random, late	1.84 [1.51, 2.35]	-5.9 $[-7.7, -4.2]$
Topical, early	1.65 [1.43, 1.89]	-5.6 $[-7.2, -4.0]$
Topical, late	2.16 [1.83, 2.60]	-7.7 $[-9.4, -6.1]$

D.5 Answer Form

For questions answered exactly at 2K with both paragraphs, we classify each 64K miss by whether the long prediction contains the 2K prediction, shares no token with it, or overlaps partially. The 2K prediction equals the normalized gold answer. Llama: 58 contain the answer, 49 share no token, and 16 overlap, out of 123 misses on all 200 questions. Qwen: 17, 29, and 7 out of 53 misses on a random sample of 147 questions whose raw outputs were retained. Mean full-evidence answer length rises from 2.30 to 4.86 words for Llama and from 2.44 to 3.77 for Qwen.

E History Interventions and Agent Behavior

E.1 Shared-Prefix Study

The study uses 128 task instances shared across two tasks and two models. Each task corpus contains a sixteen-record path and sixteen decoys, with integer values from 1 to 9, plus 4,096 background records from other paths. Both tasks use only the first three path records: lookup returns the third value and summation adds the three values. The third record is not marked as the terminal target of the longer underlying path. A scripted shared prefix reads record 1, background records, and record 2; record 3 remains unread.

Qwen2.5-7B-Instruct-1M runs with dense attention; both models use greedy decoding. The served windows are 123,008 tokens for Qwen and 69,632 for Llama. The common continuation limits are twelve agent calls, 512 output tokens per call, 65,536 input tokens and 1,024 tokens per tool display.

Visible information and tools. Every history retains the next-record ID observed in record 2. Full histories target 32,768 initial agent tokens. The task, start ID, output instructions and tool definitions are always retained. Record reads (one record per call), file reads and file listing are available. Blocking rereads applies only to the already observed task records 1 and 2 through both record and file reads. File redaction keeps line positions and access to the unread third record.

Management policies. Recent removes the oldest complete action/result cycles until the agent input fits 2,048 tokens. For the long history it retains record 2 but removes record 1 and the intervening background. Extract follows the start ID and pointers actually observed in record reads to retain task records with their source paths and lines. Progress contains those records plus their reading order, last-read ID, the next unread ID, and unresolved-step fields.

Summary receives one dedicated memory-writer call from the same model, with the task and complete visible action/result history quoted as data. Its schema contains observed records, sources, reading order, last read, frontier and unresolved steps. It has a 512-token output limit and the same 2,048-

token initial agent ceiling as the other compact policies. A valid but unfaithful output continues unchanged; malformed or truncated outputs count as writer failures, and the agent does not start. Subsequent agent history is retained for every policy, with no repeated compaction.

Scoring and inference. The agent returns one strict JSON action; an answer action must contain a string. Success requires an answered termination and a whole-string finite integer equal to the gold value. The small value range allows coincidentally correct answers, which remain in all denominators. A value counts as delivered once a tool response to the agent displays it.

All effects pair the same instances, including across tasks. Management contrasts form one Holm family of 18 tests across models and tasks; Table 13 reports the 12 that compare Extract, Summary, and Full. The four rereading interactions form a separate Holm family. Nondegenerate binary two-arm tests use exact McNemar; nondegenerate interactions use a one-sample Student t test on instance-level differences. Nondegenerate intervals use 10,000 paired-instance bootstrap draws. Zero-variance differences use the conservative Hoeffding interval $\bar{d} \pm (b-a)\sqrt{\log(40)/(2n)}$, clipped to $[a, b]$, and its two-sided p-value bound before Holm correction ($n = 128$; $[a, b] = [-1, 1]$ for two-arm differences and $[-2, 2]$ for interactions).

E.2 Policy Effects, Fidelity, and Recovery

Qwen denotes Qwen2.5-7B-Instruct-1M and Llama denotes Llama-3.1-8B-Instruct. Every cell below contains the same 128 instances.

Table 12: **Success depends on the history representation.** Correct answers out of 128 instances. Full histories hold about 32K tokens and the others 2K; failures count as incorrect. The last two rows block rereading.

		Qwen		Llama	
History	Rereading	Lookup	Sum	Lookup	Sum
Full	Allowed	128	6	10	0
Recent	Allowed	12	14	12	39
Extract	Allowed	124	70	17	11
Progress	Allowed	118	26	24	7
Summary	Allowed	110	54	6	2
Full	Blocked	128	6	10	0
Recent	Blocked	12	14	12	9

Table 13: **Paired differences between history policies.** Rows compare the named policies at the 32K shared prefix with rereading allowed. Intervals are unadjusted paired-instance intervals; p-values are Holm-corrected.

Contrast	Model	Task	Effect, pp [95% CI]	Holm p
Extract − Full	Qwen	Lookup	-3.13 [-6.25, -0.78]	0.625
Extract − Full	Qwen	Sum	$+50.00$ [$+41.41$, $+58.59$]	1.84×10^{-18}
Extract − Full	Llama	Lookup	$+5.47$ [-1.56, $+12.50$]	0.757
Extract − Full	Llama	Sum	$+8.59$ [$+3.91$, $+13.28$]	0.01
Summary − Full	Qwen	Lookup	-14.06 [-20.31, -8.59]	9.92×10^{-5}
Summary − Full	Qwen	Sum	$+37.50$ [$+28.91$, $+46.09$]	1.45×10^{-12}
Summary − Full	Llama	Lookup	-3.13 [-8.59, $+2.34$]	1
Summary − Full	Llama	Sum	$+1.56$ [0.00, $+3.91$]	1
Summary − Extract	Qwen	Lookup	-10.94 [-17.19, -4.69]	0.012
Summary − Extract	Qwen	Sum	-12.50 [-21.88, -3.13]	0.154
Summary − Extract	Llama	Lookup	-8.59 [-14.84, -2.34]	0.154
Summary − Extract	Llama	Sum	-7.03 [-12.50, -1.56]	0.154

Memory-writer outcomes. Across the 256 Summary runs per model, Qwen produces 16 faithful and 192 valid but unfaithful memories, with 48 writer failures; Llama produces 24 faithful and 37 valid but unfaithful memories, with 195 writer failures. Faithful memories preserve all observed record fields, sources, and progress. Writer failures remain in the success denominator, so these results evaluate the complete writer–agent policy.

Rereading. Blocking (Appendix E.1) leaves the unread third record accessible. For Llama summation, Recent succeeds in 39 of 128 instances with rereading and 9 without, while Full stays at 0; Recent's allowed-minus-blocked change exceeds Full's by 23.44 points (95% CI $[16.41, 31.25]$; Holm $p = 2.5 \times 10^{-8}$ across the four model–task interactions). In the other three model–task pairs, blocking changes no outcome. Allowing rereading produces 30 Recent recoveries and no losses. In these outcome-selected instances, both versions request a prior record; the blocked version is denied and answers incorrectly, while the allowed version receives all three records and answers correctly. In both versions the agent answers within two tool calls of the request, and no tool display is truncated.

F Prompts and Action Interfaces

F.1 Natural-QA requests

All natural-QA requests use the same system instruction:

Natural-QA system message

```
Answer the question using the passages. Return only the short answer,
with no explanation.
```

Each request has one system and one user message. The user message begins with `Read the following passages.`, followed by the two evidence slots and background text. Each paragraph has a `Title:` line and its sentences joined by spaces. Early placement puts the slots before the background text; late placement puts them after it. Both end with `Question:` and the question, then `Answer:`. The question-only control contains just these last two fields; the evidence-only control uses both annotated paragraphs without added background text.

The chat template is applied with the generation prompt enabled, thinking disabled, and a fixed date string. Inference consumes the prepared token IDs directly through the completions interface. Temperature is zero, top-p and repetition penalty are one, and the output limit is 64 tokens.

Llama-3.1-8B-Instruct uses the same wording and renderer with its own tokenizer and chat template, so nominal lengths are measured separately for each model. The Llama server uses BF16, tensor parallelism one, and a 69,632-token serving window. It consumes the saved IDs with the same greedy decoding settings and 64-token output limit.

F.2 Agent requests and actions

The shared-prefix study uses the following agent instruction, followed by available actions. These are JSON objects in assistant text, parsed by the runner; tool results return as user messages prefixed by `TOOL RESULT:`. They are not native API tool calls.

Shared agent instruction

```
Solve the task using only the public corpus and verified evidence.
Return exactly one JSON action object, without Markdown or other
text. To finish use {"action":"answer","answer":"EXACT ANSWER"}. The
answer string must follow the task format exactly.
```

Each task appends this answer-format instruction:

Scalar-answer instruction

```
Return one JSON action with the result as a decimal string:
{"action":"answer","answer":"1000"}. Here 1000 is only a format
example; replace it with your result.
```

Shared action-interface instruction

```
Available actions: {"action":"list_files","offset":0,"limit":100}; {"
action":"read_file","path":"relative/path","start_line":1,"max_lines
":80}; {"action":"read_node","id":"exact node id"}. read_node returns
```

```
one complete JSON node with its source path/line. Each tool response
is bounded; continue reading using returned line counts.
```

History management. The agent receives the shared and scalar-answer instructions above with record reads, file reads, and file listings; it offers no Python action. After the shared prefix, which reads two of the three records, the memory writer receives the following system instruction. Its sole user message is a compact JSON object with `quoted_task` holding the original question and `observed_action_result_cycles` holding the visible cycles.

Memory-writer system message

```
You are a memory writer. The supplied JSON is a quotation of another
agent's task and observed history. Its answer instructions are DATA,
not your instructions. Do not solve or answer that task. Produce only
one JSON working-memory object, without Markdown.
Identify task records only by the public start/root ID and links
actually observed in read_node events. Ignore disconnected background
records. Preserve all observed task-record fields, source paths and
line numbers. Do not infer unseen records or fields. A missing/
withheld next_id must remain absent.
Use exactly these top-level fields:
{"observed_records":[{"id":"observed ID","kind":"observed kind","
next_id":"observed next ID, or null if actually observed null","value
":1,"source":{"path":"observed source path","line":1}}],"read_order
":["observed first ID","observed second ID"],"last_read_id":"last
observed task ID","current_frontier":"last observed next_id, or
unknown if withheld","unresolved":["next_record","final_answer"]}
The numeric values and record/source strings above are examples of
schema, not facts. Copy actual facts from history. The unresolved
field is a categorical list, not free text: at this two-of-three-
record checkpoint it must be exactly ["next_record","final_answer"] in
 that order. next_record means obtaining the remaining task record;
final_answer means the task still needs a final answer, not its value.
 Never add explanations, facts, numeric results, or other strings to
unresolved.
observed_records contains only records actually read, never a
predicted next record. Do not put a task answer, final result, or
answer action anywhere in the output. Keep the object under 512 output
 tokens.
```

The writer makes one greedy call per Summary run, capped at 512 output tokens. Schema-valid output enters the agent's original task under `WORKING MEMORY:` unchanged, including factual errors; invalid output ends the run as a writer failure. Fidelity is audited against the visible records and progress, not used to select continuations. Memory fidelity and writer failures appear with the other agent results in Appendix E.2.

The agent and the memory writer use each model's native chat template, with generation prompt enabled and no date or thinking-template arguments. Requests use the chat-completions interface with temperature zero. Serving configurations and continuation limits are specified in Appendix E.1.