CONTEXT POISONING AS EXTREME-VALUE ATTENTION INTERFERENCE IN LONG-CONTEXT LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can process increasingly long prompts, yet their ability to locate and use decisive evidence may degrade as irrelevant or confusable context is added. We formulate this phenomenon, which we call *context poisoning*, as extreme-value interference in attention: the decisive-evidence score is upper-bounded, while the maximum score among effective distractors grows with their number. Under a softmax retrieval abstraction, we derive a finite-sample upper bound showing that maintaining a fixed accuracy target above base rate requires the evidence margin to scale as $\Omega(\sqrt{\log N})$, where N denotes the effective distractor count rather than necessarily the raw context length. The analysis connects long-context degradation to score aliasing, positional aliasing, and softmax dilution. Controlled experiments show that retrieval accuracy decreases as total context grows in the presence of embedded hard negatives, that the same-format condition produces the largest observed accuracy drop among the tested distractor constructions at fixed context length, and that retrieval gating can improve evidence use while its net benefit depends on preserving evidence recall. These results motivate evidence bottlenecks, alias-resistant representations, retrieve-then-reason architectures, verifier-mediated memory, and contrastive anti-poison training.

1 INTRODUCTION

Increasing a language model's context window is not the same as increasing its usable memory. Transformer attention gives every token a path to every other token (Vaswani et al., 2017), and large autoregressive models can exploit prompts for in-context adaptation (Brown et al., 2020). However, long input does not guarantee robust evidence use. Empirical studies report strong position effects and degradation when relevant information is buried in the middle of a long context (Liu et al., 2024), sensitivity to irrelevant information (Shi et al., 2023), and benchmark-level gaps between nominal and effective context length (Bai et al., 2024; An et al., 2024; Hsieh et al., 2024; Du et al., 2025). This paper uses the term *context poisoning* for a non-training-data phenomenon: the input itself becomes contaminated by distractors, aliases, and positional artifacts as the window grows.

The practical symptom is familiar. A prompt contains a small set of key facts, then many thousands of irrelevant or weakly related tokens. A model that would answer correctly from the short prompt may answer incorrectly from the long prompt, quote a distractor, or ignore the key fact. Existing long-context systems attack the computational side through recurrence, compression, sparse attention, or fast exact attention (Dai et al., 2019; Rae et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020; Choromanski et al., 2021; Katharopoulos et al., 2020; Dao et al., 2022). Those methods make long contexts feasible, but feasibility alone does not imply robust extraction. Positional methods such as ALiBi, RoPE, interpolation, and YaRN improve extrapolation behavior (Press et al., 2022; Su et al., 2024; Chen et al., 2023; Peng et al., 2023), but long-context failures can persist even within claimed windows.

Contributions. First, we give formal definitions for context poisoning, score aliasing, and effective distractor count. Second, we prove a theorem: in a simple but diagnostic attention model, the probability that a distractor captures enough attention increases with N, and maintaining accuracy

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.

requires a signal margin that grows like $\sqrt{\log N}$. Third, we derive design principles that turn the bound into system requirements: reduce N, increase the evidence margin, verify evidence before decoding, and regularize aliasing. Fourth, we report two controlled empirical studies in the main paper: a hard-negative retrieval decay curve on a production model (Empirical study: retrieval decay with embedded hard negatives) and a fixed-length distractor-confusability ablation across three models (Empirical study: distractor confusability at fixed context length). The appendix provides a retrieve-then-reason gate experiment at context lengths up to 512K tokens (Empirical Test of the Gate-Mitigation Prediction), together with the full proofs, additional implementation details, complete experimental protocols and supplemental tables, attention-distribution diagnostics, a confusable-decoy dose-response analysis, and a lost-in-the-middle reanalysis. Across these components, the paper connects a finite-sample failure bound to observable changes in distractor count, distractor composition, and retrieval recall. We do not claim that every LLM failure is explained by a single attention head or Gaussian scores. The theorem is a stress test: if a long-context architecture cannot control the extreme distractor term in this abstraction, it should not be expected to remain robust merely by scaling the raw context window.

2 Related work

Efficient long-context architectures. Long-context language modeling builds on transformers (Vaswani et al., 2017), large-scale pretraining (Devlin et al., 2019; Raffel et al., 2020), and in-context learning (Brown et al., 2020). Transformer-XL and Compressive Transformers introduced recurrence and memory compression (Dai et al., 2019; Rae et al., 2020). Longformer, Big Bird, Performer, and linear attention reduce long-sequence cost (Beltagy et al., 2020; Zaheer et al., 2020; Choromanski et al., 2021; Katharopoulos et al., 2020), while FlashAttention improves exact-attention efficiency (Dao et al., 2022). ALiBi, RoPE, positional interpolation, and YaRN address positional extrapolation and context-window extension (Press et al., 2022; Su et al., 2024; Chen et al., 2023; Peng et al., 2023). These methods make long contexts feasible, but feasibility does not ensure that decisive evidence remains distinguishable from many competitive candidates.

Selection, compression, and memory. Retrieval-augmented generation selects a smaller evidence set before decoding (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021), and prompt compression removes low-utility tokens (Jiang et al., 2023). Attention-sink methods instead preserve selected states to stabilize streaming generation (Xiao et al., 2024). These approaches expose a shared tradeoff: reducing the candidate set can lower interference, but an imperfect selector can discard required evidence. Proposition 1 expresses this tradeoff as replacing N by K effective distractors while paying a recall penalty η.

Long-context evaluation and failure analysis. Lost-in-the-middle effects show that central evidence can be underused (Liu et al., 2024), and irrelevant information can distract reasoning even when the evidence remains present (Shi et al., 2023). Positional interventions can partly recover evidence use (Zhang et al., 2024). LongBench, L-Eval, ZeroSCROLLS, and RULER broaden evaluation across tasks and context lengths (Bai et al., 2024; An et al., 2024; Shaham et al., 2023; Hsieh et al., 2024). Recent controlled work further shows degradation despite controlled or perfect retrieval (Du et al., 2025), reinforcing the distinction between nominal capacity, evidence retrieval, and evidence use.

3 Formalizing context poisoning

Let a query q request an answer $y^\star$ from a context $C_N = (x_0, \ldots, x_N)$. The index set of relevant evidence is $R \subseteq \{0, \ldots, N\}$, and $D = \{0, \ldots, N\} \setminus R$ are distractors. A model f_θ induces an answer distribution $p_\theta(y \mid q, C_N)$ and a score map $s_\theta(q, x_i, i)$ used, explicitly or implicitly, to select evidence. The score includes token semantics and position: for a transformer head one may write

$$s_i = \frac{\langle W_Q h_q, W_K(h_i + p_i)\rangle}{\sqrt{d}}, \tag{1}$$

where h_i is a contextual token representation and p_i denotes a positional code or bias.

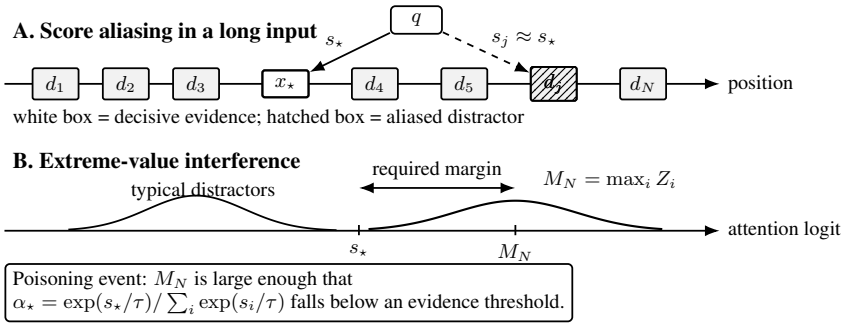


Figure 1: Score and positional aliasing in a long context. A relevant token $x_\star$ competes not with an average distractor but with the maximum of many distractors. As N_{eff} grows, the right tail creates aliased distractors whose logits can dominate the softmax denominator.

Definition 1 (Context poisoning curve)**.** *Fix a task distribution $\mathcal{D}$, a short reference length N_0, and an augmentation operator T_N that appends $N - N_0$ irrelevant or nuisance tokens while preserving the answer. The accuracy at length N is*

$$A(N) = \mathbb{P}_{(q,C_{N_0},y^\star)\sim\mathcal{D}}\left[\arg\max_y p_\theta(y \mid q, T_N(C_{N_0})) = y^\star\right]. \tag{2}$$

The poisoning curve is $P(N) = A(N_0) - A(N)$. Context poisoning occurs on $[N_0, N_1]$ when $P(N)$ is positive and non-negligible for some $N \le N_1$, even though T_N does not change the ground-truth answer.

Definition 2 (Score aliasing and effective distractor count)**.** *For a query q, relevant item $r \in R$, and distractor $i \in D$, the pair (r, i) is ϵ-score-aliased if $|s_i - s_r| \le \epsilon$. It is ϵ-position-aliased if the positional contribution obeys*

$$|\langle W_Q h_q, W_K p_i\rangle - \langle W_Q h_q, W_K p_r\rangle| \le \epsilon. \tag{3}$$

The effective distractor count $N_{\text{eff}}(q, C)$ is the number of distractors whose scores have non-negligible upper-tail probability near a relevant score. In the theorem below, $N_{\text{eff}} = N$ by construction; in a system with retrieval or gating, N_{eff} is the number of tokens or chunks surviving the gate.

Figure 1 summarizes the issue. Long contexts produce many opportunities for an irrelevant token to become close to, or exceed, the relevant score. This can happen semantically, positionally, or through attention-normalization effects. The key mathematical feature is not the average distractor score; it is the maximum distractor score.

4 A THEOREM FOR ATTENTION-MEDIATED POISONING

The theorem studies a single retrieval bottleneck. This abstraction is intentionally small: many LLM computations are multi-layer and multi-head, yet final answer quality often depends on whether decisive evidence receives enough usable representation mass. The result states that a fixed signal margin cannot protect against an unbounded number of effective distractors.

Definition 3 (ρ-faithful evidence use)**.** *Consider a model with attention mass $\alpha_{\star,N}$ assigned to the decisive evidence. For a threshold $\rho \in (0, 1)$ and base accuracy $a_0 \in [0, 1]$, the decoder is ρ-faithful if*

$$\mathbb{P}(\widehat{y} = y^\star \mid \alpha_{\star,N} < \rho) \le a_0. \tag{4}$$

Here a_0 may be the chance accuracy of a multiple-choice task, the accuracy of a prior-only model, or the accuracy of an abstaining verifier.

Theorem 1 (Extreme-value attention poisoning)**.** *Fix temperature $\tau > 0$, threshold $\rho \in (0, 1)$, base accuracy $a_0 \in [0, 1]$, and noise scale $\sigma > 0$. For each context length N, let the context contain one decisive item with attention logit $S_\star \le \gamma$ almost surely and N effective distractors with independent logits*

$$Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{5}$$

Let

$$\alpha_{\star,N} = \frac{\exp(S_\star/\tau)}{\exp(S_\star/\tau) + \sum_{i=1}^{N} \exp(Z_i/\tau)} \tag{6}$$

be the relevant attention mass, and assume the decoder is ρ-faithful. Then the expected accuracy $A_N = \mathbb{P}(\widehat{y} = y^\star)$ satisfies

$$A_N \le a_0 + (1 - a_0) \left[\Phi\left(\frac{\gamma + \tau \log \frac{1-\rho}{\rho}}{\sigma} \right) \right]^N, \tag{7}$$

where Φ is the standard normal cdf. Consequently, if $\gamma, \tau, \rho, \sigma$ are fixed and a_0 is fixed, then $\limsup_{N\to\infty} A_N \le a_0$. Conversely, for the upper bound to permit $A_N \ge 1 - \epsilon$ with $\epsilon < 1 - a_0$, it is necessary that

$$\begin{aligned} \gamma &\ge \sigma \Phi^{-1}\left(\left(\frac{1 - \epsilon - a_0}{1 - a_0} \right)^{1/N} \right) - \tau \log \frac{1-\rho}{\rho} \\ &= \sigma \sqrt{2 \log N} + O(1) \end{aligned} \tag{8}$$

for fixed $\epsilon, a_0, \rho, \tau, \sigma$ as $N \to \infty$.

Interpretation. Equation equation 7 is a formal poisoning curve. It says that the model's usable memory decays as the chance of at least one high-scoring distractor increases. The theorem also explains why "just attend over more tokens" can fail: unless architecture or training increases the relevant margin with N, the attention mass on the relevant evidence eventually becomes too small. The dependence is on N_{eff}, not necessarily raw token length. A perfect gate that reduces the candidate set from N to K changes the bound by replacing N with K.

Corollary 1 (Random relevant logit)**.** *Suppose $S_\star$ is independent of $Z_1, \ldots, Z_N$ and $\mathbb{P}(S_\star > \gamma) \le \delta$. Then*

$$A_N \le \delta + a_0 + (1 - a_0) \left[\Phi\left(\frac{\gamma + \tau \log \frac{1-\rho}{\rho}}{\sigma} \right) \right]^N. \tag{9}$$

Thus occasional high relevant logits help only through the tail probability δ; robust long-context use requires systematic margin growth or a smaller effective distractor set.

5 Mechanisms predicted by the theorem

Softmax dilution. Even when no single distractor wins, the denominator $\sum_i \exp(Z_i/\tau)$ can grow. The theorem uses a maximum event because it yields a clean lower bound on failure; denominator growth gives a second, cumulative failure route. Smaller temperature τ sharpens attention, but it also increases sensitivity to logit noise and can over-commit to an aliased distractor.

Position aliasing. Long-context extrapolation requires distinguishing positions never or rarely seen during training. Methods such as ALiBi and RoPE encode useful inductive biases (Press et al., 2022; Su et al., 2024), while interpolation and scaling methods extend the range of positional features (Chen et al., 2023; Peng et al., 2023). The definition above highlights a failure case: if two far-apart positions induce similar query-key contributions, a distractor can become indistinguishable from evidence for a given head. "Found in the middle" style interventions can be interpreted as changing positional logits to reduce the number of near-ties around relevant evidence (Zhang et al., 2024).

Retrieval versus use. The result is compatible with work showing that retrieval can be imperfect (Liu et al., 2024; Hsieh et al., 2024), but it also covers the stronger case in which retrieval succeeds while problem solving still degrades (Du et al., 2025). In the theorem, the evidence item is present, but its score advantage is bounded; failure arises as the growing candidate set increases the maximum competing score. In a multi-step reasoning problem, similar interference can occur at multiple evidence selection stages.

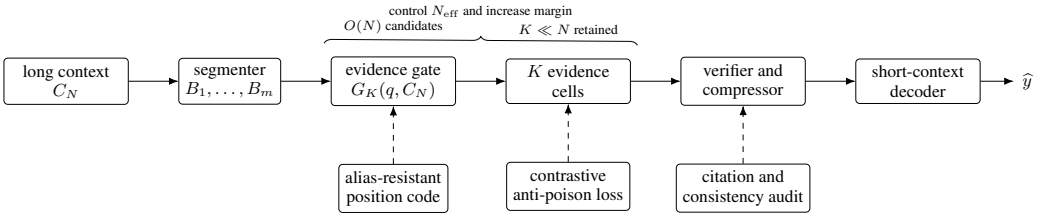


Figure 2: Poison-resilient long-context architecture. A gate converts a raw $O(N)$ context into K evidence cells, alias-resistant positional coding reduces near-ties, contrastive losses push distractors below evidence, and a verifier forces the decoder to reason over recited, auditable evidence rather than the full noisy context.

6 Architectural alternatives

The bound suggests a design principle: long-context systems should not merely increase the raw token window; they should control N_{eff}, amplify the evidence margin, and verify the evidence before answer generation. Figure 2 shows a candidate architecture.

6.1 Evidence bottleneck and retrieve-then-reason

A model-agnostic mitigation is to split long-context processing into two calls or two internal phases: retrieve and recite the evidence, then solve from the shorter evidence prompt. This resembles retrieval-augmented generation (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021), but the retrieval source may be the user's own context rather than an external corpus. The theorem predicts that, conditional on retaining the decisive evidence, the extreme-value term changes from $\Phi(t)^N$ to $\Phi(t)^K$. Evidence omissions introduce a separate penalty governed by the gate's miss probability η. The intervention is therefore useful only when reducing the candidate set does not substantially reduce evidence recall.

Proposition 1 (Gate substitution)**.** *Assume the theorem's conditions after a gate G_K returns a set of at most K distractors and includes the decisive evidence with probability at least $1-\eta$. Then*

$$A_N \le \eta + a_0 + (1-a_0)\left[\Phi\left(\frac{\gamma + \tau \log \frac{1-\rho}{\rho}}{\sigma}\right)\right]^K. \tag{10}$$

The gate improves the bound when the recall penalty η is smaller than the reduction in extreme-value risk from replacing N by K.

6.2 Alias-resistant positional and memory design

Position encodings should be evaluated by their worst-case confusability, not only by perplexity. One can regularize heads so that for sampled query states q and relevant positions r,

$$\mathcal{L}_{\text{alias}} = \mathbb{E}\left[\log \sum_{i \in D} \exp\left(\frac{s_i - s_r + m(i,r)}{\tau_a}\right)\right], \tag{11}$$

where $m(i,r)$ is a distance-aware margin. This loss penalizes the log-sum-exp of distractor gaps and directly targets the denominator and maximum terms in the theorem. For streaming applications, preserving a small set of attention sinks can stabilize decoding (Xiao et al., 2024); however, sinks should be separated from semantic evidence cells so that sink tokens do not become spurious evidence.

6.3 Verifier-mediated memory

A verifier can transform long-context use from implicit attention to explicit evidence checking. Given candidate cells $E_1, \ldots, E_K$, the verifier produces a structured state

$$z = (\text{claims}, \text{support spans}, \text{conflicts}, \text{abstain flag}). \tag{12}$$

Table 1: Mitigation mechanisms and their mathematical target.

Mechanism	Bound-level effect	Implementation sketch
Evidence bottleneck	Replace N by $K \ll N$ while paying recall error η	Retrieve, recite, then solve; cite spans
Alias-resistant positions	Reduce probability that $s_i \approx s_\star$	RoPE/ALiBi scaling plus margin regularization
Contrastive anti-poison loss	Increase γ and suppress $\max_i Z_i$	Hard negatives, shuffled positions, adversarial distractors
Verifier-mediated memory	Reduce a_0 failures and detect low-margin cases	Evidence table, consistency checks, abstention
Hierarchical memory	Bound per-level N_{eff}	Chunk summaries with back-pointers to raw spans

The decoder receives z and a compact evidence prompt. This creates two audit points: whether the key evidence was selected, and whether the answer follows from selected evidence. It also enables selective abstention when the evidence margin is below a calibrated threshold.

6.4 Training objective

For synthetic and real long-context tasks, construct triples $(q, x_\star, D, y^\star)$ with hard negatives. Let $s_\star$ be the score for the decisive evidence and s_i for distractors. A direct anti-poison objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{ans}} - \lambda \log \frac{\exp(s_\star/\tau)}{\exp(s_\star/\tau) + \sum_{i \in D} \exp(s_i/\tau)} \\ + \mu \sum_{i \in D} \max\{0, s_i - s_\star + m_i\},\end{aligned} \tag{13}$$

where m_i increases for semantically close or positionally confusable distractors. The second term controls attention mass; the third controls the maximum distractor gap. Prompt-compression methods such as LongLLMLingua are complementary when they preserve decisive evidence and remove low-utility tokens (Jiang et al., 2023). Table 1 summarizes these mitigation mechanisms and their corresponding effects on the theoretical bound.

7 Experimental setup and metrics

The theory recommends measuring the whole poisoning curve, not just the maximum accepted length. A benchmark instance should preserve the short-context answer while independently varying raw length N, effective distractor count N_{eff}, relevant position, number of evidence items, semantic similarity of distractors, and positional regime. Existing suites such as LongBench, L-Eval, ZeroSCROLLS, and RULER provide useful starting points (Bai et al., 2024; An et al., 2024; Shaham et al., 2023; Hsieh et al., 2024).

We report four quantities where available: (i) short-context accuracy $A(N_0)$, (ii) the poisoning curve $P(N) = A(N_0) - A(N)$, (iii) evidence recall, and (iv) conditional answer accuracy after successful retrieval. The last quantity separates retrieval failure from evidence-use failure (Du et al., 2025). Paired conditions reuse the same questions, and failure outputs are inspected when possible to distinguish supplied-decoy confusion from hallucination or formatting error.

Table 2 divides the hypothesis space across candidate count, distractor composition, and mitigation. We summarize controlled study properties rather than model architecture because the hosted systems do not expose comparable internal specifications. No single study validates the full theory; together they test distinct, non-redundant implications.

Table 2: Overview of the controlled studies. The first two studies are reported in the main paper; the gate intervention is reported in the appendix. The designs vary complementary axes rather than repeating the same manipulation.

Study	Model(s)	Primary intervention	Context regime	Diagnostic separation
Hard-negative decay	GPT-4o	Increase distractor count while retaining embedded decoys	$N \in \{100, 300, 500, 1000\}$ sentences	Manual audit verifies that errors select supplied decoy values
Distractor confusability	Claude Sonnet 4, GPT-4.1, Gemma 3 (12B)	Change topical, structural, and entity-name overlap at fixed N	300 records, approximately 20–21K tokens	Answer accuracy and cited-record recall distinguish extraction from retrieval
Retrieve-then-reason gate	GPT-4.1	Replace all N chunks with BM25 top-K candidates	4K–512K tokens	Recall@K and accuracy conditional on retrieving all support

Table 3: Retrieval accuracy vs. number of distractors N, GPT-4o, 10 trials per condition.

N	Trials	Accuracy	Mean prompt length (chars)
100	10	100%	8,952
300	10	80%	25,978
500	10	80%	41,954
1000	10	80%	84,668

8 Empirical study: retrieval decay with embedded hard negatives

Experimental setup and metrics introduces a minimal synthetic test of Theorem 1—key-value retrieval with a variable number of hard negatives. We carry out that test here, checking whether the theorem's central prediction, that accuracy decays toward a chance floor as the number of distractors grows, holds for a modern production model.

Setup. We embed a single target fact in a passage of N surrounding sentences and ask GPT-4o to retrieve it. The filler is drawn from a public-domain novel, so it reads as ordinary prose rather than an artificial list. Critically, the target fact is not the only sentence of its kind: several *decoy* sentences, each stating a different value in the same natural phrasing as the true fact, are woven into the passage. This makes the task a genuine test of retrieval under hard negatives rather than trivial pattern matching—the model must identify *which* value is being asked about from narrative context, not merely notice that a value is present. The target's position is randomized on each trial, and N is varied across $\{100, 300, 500, 1000\}$ with 10 independent trials per condition.

Prediction. Theorem 1 predicts that retrieval accuracy decays toward a chance floor a_0 as N grows, following

$$A(N) \approx a_0 + (1 - a_0)\, q^N, \qquad q = \Phi(b/\sigma) \in (0, 1), \tag{14}$$

unless the target's evidence margin grows with N. We fit Eq. equation 14 to the observed accuracy by nonlinear least squares, and we inspected every incorrect response individually to confirm that failures reflect genuine confusion among competing candidates (the model's answer was always one of the decoy values actually present in that trial's passage) rather than hallucination or task misunderstanding.

Results. Table 3 shows accuracy holding at 100% at $N = 100$ before dropping to a stable 80% by $N = 300$ and remaining there through $N = 1000$. Fitting Eq. equation 14 gives $\hat{a}_0 = 0.777$, $\hat{q} = 0.9962$, with $R^2 = 0.74$.

Interpretation. These results provide initial behavioral evidence consistent with the theorem's qualitative prediction: retrieval accuracy decreases as exposure to hard-negative distractors grows.

Table 4: Answer accuracy by model and distractor construction at fixed context length ($N = 300$ records, ∼20K tokens, errors excluded). The rightmost column reports entity-name lexical overlap, measured by character-trigram Jaccard similarity to the gold entity. Gemma values are approximate due to a high endpoint error rate. Full results with evidence recall, valid counts, and paired contrasts are provided in the appendix (Distractor confusability ablation: full protocol and results).

Distractor condition	Claude Sonnet 4	GPT-4.1	Gemma 3 (12B)	Sim. proxy
Unrelated filler	1.000	1.000	1.000	0.004
Same domain, different attribute	1.000	1.000	1.000	0.005
Same format, different entity	**0.889**	**0.967**	∼**0.85**	0.010
Near-alias entity	1.000	0.983	∼0.89	0.317

Every observed error selected a decoy value present in the prompt, ruling out hallucination, refusal, or task misunderstanding as the immediate failure type. Because the experiment does not observe internal attention scores, it does not establish attention interference as the unique underlying mechanism. The fitted exponential form is also exploratory rather than conclusive ($R^2 = 0.74$): accuracy drops between $N = 100$ and $N = 300$ and then plateaus through $N = 1000$, suggesting a threshold-like transition over the tested range. A denser sweep around this transition and more trials per condition are needed to estimate the curve's shape reliably.

This study varies candidate exposure but cannot separate raw count from candidate composition. The next study holds record count and approximate token length fixed while changing the form of competition.

9 Empirical study: distractor confusability at fixed context length

We instantiate the evaluation protocol with a controlled study that holds raw context length fixed while varying observable properties of distractor confusability. This design tests a behavioral implication of score aliasing (Definition 2): retrieval performance may depend not only on the number of tokens or records in the context, but also on how strongly the distractors compete with the gold evidence. Because the experiment does not directly observe attention scores, it should be interpreted as a behavioral test of this implication rather than a direct measurement of the effective distractor count N_{eff}.

Setup. Each instance is a key–value retrieval query over $N = 300$ records (∼20,000–21,000 tokens after length matching) with one gold record; the model must return the answer and cite the evidence record ID. We evaluate three models—Claude Sonnet 4, GPT-4.1, and Gemma 3 (12B)—at temperature zero, holding N fixed across four distractor constructions: unrelated filler, same-domain/different-attribute, same-format/different-entity, and near-alias records sharing the gold entity's surname. These manipulate topical, structural, and entity-name overlap separately, not as points on one similarity scale. Records are padded to a fixed length to avoid confounding with token count, and evidence position is balanced across beginning, middle, and end. Full protocol, prompts, and per-model tables appear in the appendix (Distractor confusability ablation: full protocol and results).

Results. Table 4 and Figure 3 summarize the outcome. Accuracy is near-ceiling for unrelated and same-domain distractors for all three models, then decreases in the same-format condition: Claude falls to 0.889 (an 11.1 percentage-point paired drop against baseline, 95% CI $[+2.8, +22.2]$—the only statistically significant paired contrast in the study), GPT-4.1 falls to 0.967, and Gemma falls to approximately 0.85. The character-trigram Jaccard measure serves specifically as an entity-name lexical-overlap check. It remains below 0.01 for the first three conditions and rises sharply to 0.317 only in the near-alias condition (Figure 3, right). Thus, the decrease in the same-format condition occurs despite low entity-name overlap and cannot be explained by name similarity alone.

Two findings. First, *structural overlap is more damaging in these data than entity-name overlap*. The largest observed degradation for every model occurs in the same-format condition, where all

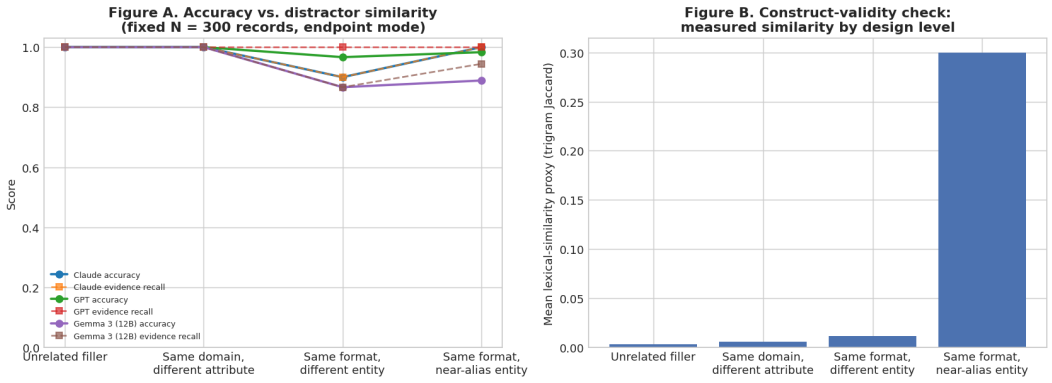


Figure 3: Distractor-confusability ablation at fixed context length ($N = 300$). **Left:** answer accuracy and evidence recall across four distractor constructions; all models show their largest decrease in the same-format condition, while the two frontier models recover toward baseline in the near-alias condition. **Right:** entity-name lexical-overlap check—the character-trigram Jaccard measure is near zero for the first three conditions and rises only in the near-alias condition. The accuracy decrease in the same-format condition therefore cannot be attributed to high entity-name overlap.

300 records share the gold record's template but refer to different entities. This pattern is behaviorally consistent with increased score aliasing, because a shared structure may create a larger set of competitive candidates. However, the experiment does not directly observe attention scores or determine which distractors fall within an ϵ-aliasing band. Second, we observe a *near-alias paradox*: despite substantially greater entity-name similarity, accuracy in the near-alias condition recovers toward baseline for the two frontier models (Claude 1.000, GPT-4.1 0.983). One possible explanation is that the first name remains a discriminative cue, whereas structurally identical records provide fewer template-level features for distinguishing the gold record. This explanation is a hypothesis rather than a directly tested mechanism. Gemma shows the largest drops overall, but its endpoint rejected 96% of these ∼20K-token calls, so its estimates are underpowered and should be interpreted as suggestive.

10 Limitations and Conclusion

The theorem isolates one long-context failure mechanism rather than modeling an entire LLM: it assumes independent Gaussian distractor logits, a single decisive item, and a threshold model of evidence use, whereas real systems contain correlated representations, multiple heads, residual pathways, memorized priors, tools, and instruction hierarchies. The result is thus a mechanistic diagnostic rather than a calibrated predictor, and the appendix attention analyses confirm that, although the Gaussian i.i.d. abstraction is not exact, the predicted extreme-value scale remains a useful reference.

The empirical evidence is similarly controlled: the retrieval-decay study uses only 10 trials per condition, the fixed-length ablation measures behavioral proxies rather than attention-score margins, and unequal endpoint reliability makes the Gemma estimates suggestive. Both studies emphasize retrieval-style tasks, so code, conversations, and temporally conflicting documents may exhibit additional failure modes. Even so, theory and experiment point in the same direction.

A fixed evidence margin becomes increasingly fragile as the effective distractor set grows, and retrieval performance depends on both candidate exposure and distractor structure, so reducing the competitive set helps only when required evidence is preserved. A larger context window increases both available information and the opportunities for distractors to compete with decisive evidence, and the relevant quantity is therefore the effective number and strength of confusable tokens rather than raw window length. Robust systems should accordingly combine evidence selection, alias control, recall-preserving retrieval, and verification, since a context window is capacity while usable memory is capacity plus reliable addressing. Future work should cross context length with distractor construction and report accuracy, recall, and conditional accuracy together, since any one metric alone can conceal failure elsewhere.

AI USE STATEMENT

In this work, we used generative AI tools to implement experimental methods, clean and reformat data, generate the synthetic fictional- character records used in the distractor-confusability ablation (Section 9 and Appendix C), and support interpretation of a subset of our results in Sections 8–9. We have not used generative AI tools to formulate mathematical claims, develop the theoretical framework, or write any part of the proofs; Theorem 1, its corollary, Proposition 1, and their proofs in Appendix A were derived and written by the authors. Additionally, we used generative AI tools to edit portions of the manuscript text for readability and concision. We have reviewed all AI-assisted work: experimental code was verified and tested for correctness by the authors before use, generated synthetic data was manually checked for factual consistency with the source fictional universes, and all AI-edited text was reviewed and revised by the authors for accuracy. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

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

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan 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*, pp. 23281–23298, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, pp. 3929–3938, 2020.

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 *Conference on Language Modeling (COLM)*, 2024.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 874–880, 2021.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2021. arXiv:2112.09118.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMlingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

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

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models, 2023.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

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

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227, 2023.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

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

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

APPENDIX

A PROOFS

Proof of Theorem 1. Define

$$b = \gamma + \tau \log \frac{1-\rho}{\rho}, \qquad M_N = \max_{1 \le i \le N} Z_i. \tag{15}$$

If $M_N \ge b$, then there exists a distractor j such that

$$Z_j - S_\star \ge Z_j - \gamma \ge \tau \log \frac{1-\rho}{\rho}. \tag{16}$$

Using only this distractor in the softmax denominator gives

$$\alpha_{\star,N} = \frac{\exp(S_\star/\tau)}{\exp(S_\star/\tau) + \sum_{i=1}^{N} \exp(Z_i/\tau)} \tag{17}$$

$$\le \frac{\exp(S_\star/\tau)}{\exp(S_\star/\tau) + \exp(Z_j/\tau)} = \frac{1}{1 + \exp((Z_j - S_\star)/\tau)} \tag{18}$$

$$\le \frac{1}{1 + (1-\rho)/\rho} = \rho. \tag{19}$$

Therefore, if $M_N > b$, then $\alpha_{\star,N} < \rho$, and hence

$$\{\alpha_{\star,N} \ge \rho\} \subseteq \{M_N \le b\}.$$

Since the Z_i are i.i.d. Gaussian, M_N has a continuous distribution, so

$$\mathbb{P}(M_N \le b) = \prod_{i=1}^{N} \mathbb{P}(Z_i \le b) = \left[\Phi\left(\frac{b}{\sigma}\right)\right]^N. \tag{20}$$

Thus

$$\mathbb{P}(\alpha_{\star,N} \ge \rho) \le \mathbb{P}(M_N \le b) = \left[\Phi\left(\frac{b}{\sigma}\right)\right]^N. \tag{21}$$

By ρ-faithfulness,

$$A_N = \mathbb{P}(\widehat{y} = y^\star, \alpha_{\star,N} \ge \rho) + \mathbb{P}(\widehat{y} = y^\star, \alpha_{\star,N} < \rho) \tag{22}$$

$$\le \mathbb{P}(\alpha_{\star,N} \ge \rho) + a_0 \mathbb{P}(\alpha_{\star,N} < \rho) \tag{23}$$

$$= a_0 + (1-a_0)\mathbb{P}(\alpha_{\star,N} \ge \rho) \tag{24}$$

$$\le a_0 + (1-a_0)\left[\Phi\left(\frac{\gamma + \tau \log \frac{1-\rho}{\rho}}{\sigma}\right)\right]^N. \tag{25}$$

This proves the finite-sample accuracy bound stated in Theorem 1. If b is fixed, then $\Phi(b/\sigma) < 1$ unless $b = +\infty$, so the exponential term tends to zero and $\limsup_{N\to\infty} A_N \le a_0$.

For the bound to permit accuracy at least $1-\epsilon$, it is necessary that

$$a_0 + (1-a_0)\Phi(b/\sigma)^N \ge 1 - \epsilon. \tag{26}$$

Rearranging gives

$$\Phi(b/\sigma) \ge \left(\frac{1-\epsilon-a_0}{1-a_0}\right)^{1/N}. \tag{27}$$

Because Φ^{-1} is increasing,

$$b \ge \sigma \Phi^{-1}\left(\left(\frac{1-\epsilon-a_0}{1-a_0}\right)^{1/N}\right). \tag{28}$$

Substituting

$$b = \gamma + \tau \log \frac{1-\rho}{\rho}$$

yields the finite-sample margin requirement stated in Theorem 1.

To obtain its asymptotic form, define

$$r = \frac{1-\epsilon-a_0}{1-a_0} \in (0,1) \qquad \text{and} \qquad \kappa = -\log r > 0.$$

Then

$$r^{1/N} = \exp\left(-\frac{\kappa}{N}\right) = 1 - \frac{\kappa}{N} + O(N^{-2}). \tag{29}$$

Using the standard Gaussian upper-quantile asymptotic,

$$\Phi^{-1}\left(1 - \frac{\kappa}{N} + O(N^{-2})\right) = \sqrt{2\log N} + O(1). \tag{30}$$

Therefore,

$$\gamma \geq \sigma\sqrt{2\log N} + O(1), \tag{31}$$

which proves the stated asymptotic margin requirement. □

Proof of the corollary. Let $E = \{S_\star \leq \gamma\}$. On E, the theorem's proof applies conditionally. On E^c, bound accuracy by one. Hence

$$A_N \leq \mathbb{P}(E^c) + \mathbb{P}(E)\left(a_0 + (1-a_0)\left[\Phi\left(\frac{\gamma+\tau\log\frac{1-\rho}{\rho}}{\sigma}\right)\right]^N\right) \tag{32}$$

$$\leq \delta + a_0 + (1-a_0)\left[\Phi\left(\frac{\gamma+\tau\log\frac{1-\rho}{\rho}}{\sigma}\right)\right]^N. \tag{33}$$

□

B Additional implementation details

A practical implementation of the retrieve-then-reason architecture shown in Figure 2 of the main paper can be built without changing a base LLM: chunk the context, retrieve top-K chunks with a lexical and dense hybrid scorer (Robertson & Zaragoza, 2009), ask the model to quote relevant spans, and start a new short-context call that contains only the quoted spans and the question. A model-internal implementation can train a gate and verifier jointly with the decoder. In both cases, the evaluation should report gate recall separately from final answer accuracy.

C Distractor confusability ablation: full protocol and results

This section provides the complete protocol and supplemental tables for the distractor-confusability study reported in the main paper (Empirical study: distractor confusability at fixed context length).

C.1 Task and distractor constructions

Each instance poses a factual question about a fictional character drawn from well-known literary universes (Harry Potter, The Lord of the Rings, The Chronicles of Narnia, A Song of Ice and Fire); the entities are used only to construct newly written, factual-style records, and no copyrighted text is reproduced. The model receives one gold record and 299 distractor records, for a total of $N_{\text{records}} = 300$ records, and must return a JSON object containing the extracted answer and the cited evidence record ID. Illustrated for the question "What is Stannis Baratheon's castle?" (gold answer: Dragonstone; gold record ID: gold-0000), the four distractor conditions are:

- **Unrelated filler (baseline).** Distractors are catalogue notes about unrelated objects (e.g. "a dragon egg"), with no lexical or semantic overlap with the gold entity or attribute. The gold record is the only topically relevant record.

- **Same domain, different attribute.** Distractors are facts from the same fictional universe but about different attributes and entity types (lineages, battles, geography), creating topical but non-responsive overlap.
- **Same format, different entity.** Every distractor uses the gold record's exact template ("Authoritative record: X's castle is Y") for a clearly different character, so the model must discriminate purely on entity name among 300 structurally identical records.
- **Near-alias entity.** Distractors use the same template and name entities that share the gold entity's surname (e.g. Robert, Renly, and Joffrey Baratheon versus Stannis Baratheon), each giving a factually correct but wrong-for-the-question answer, so a model that matches on surname alone extracts the wrong answer.

C.2 CONTROLS

The context contains $N_{\text{records}} = 300$ records in every condition. Each distractor record is padded to a fixed word count with neutral filler tokens so the similarity manipulation is not confounded with text length. Evidence position is balanced across beginning, middle, and end, and each base instance appears exactly once per similarity condition to enable within-instance paired comparisons. A model-independent lexical-similarity proxy (character-trigram Jaccard between each distractor entity name and the gold entity name) is computed per condition as a manipulation check. API errors (timeouts, rate limits, context-length rejections) are excluded from accuracy calculations; only successful responses are scored.

C.3 MODELS AND SCALE

Claude Sonnet 4 and GPT-4.1 were accessed through an OAuth2 gateway; Gemma 3 (12B) through a pay-per-token foundation-model API. Claude and GPT used 60 base instances $\times\, 4$ levels $= 240$ conditions each; Gemma used an expanded design of 460 base instances ($1{,}840$ conditions) to accumulate valid calls despite a high endpoint error rate. Each call processes roughly $20{,}000$ tokens of context. Answer accuracy uses loose string normalization (lowercased, articles stripped, whitespace collapsed); evidence recall requires the correct gold record ID; paired accuracy drops are estimated by paired bootstrap ($5{,}000$ iterations, seed 123).

C.4 CONSTRUCT VALIDITY AND LENGTH MATCHING

The entity-name lexical-overlap proxy remains small in the unrelated, same-domain, and same-format conditions (0.004–0.010) and rises to 0.317 in the near-alias condition. This verifies that the near-alias construction specifically increases entity-name overlap, whereas the same-format construction does not. Median approximate token counts are similar across conditions, so the observed accuracy differences are not explained by large changes in context length (Table 5).

Table 5: Construct-validity and length-matching check: mean lexical-similarity proxy (character-trigram Jaccard between distractor and gold entity names) and median approximate token count by distractor condition, at $N_{\text{records}} = 300$.

Distractor condition	Mean similarity proxy	Median tokens
Unrelated filler	0.004	20,508
Same domain, different attribute	0.005	20,149
Same format, different entity	0.010	21,055
Near-alias entity	0.317	21,053

C.5 API RELIABILITY

At $\sim$20K tokens, endpoint reliability varied sharply across providers (Table 6). GPT-4.1 completed all calls; Claude's error rate scaled with token count (approximately 11.7% at $\sim$9K, 35.4% at $\sim$20K, 59.6% at $\sim$35K in companion runs); Gemma's endpoint rejected most $\sim$20K-token inputs at a context-length limit. All accuracy analyses use valid (non-error) calls only.

Table 6: Endpoint reliability at ∼20K tokens. Accuracy analyses are computed on valid calls only.

Model	Total calls	Valid calls	Error rate	Primary cause
Claude Sonnet 4	240	155	35.4%	Rate limits / timeouts
GPT-4.1	240	240	0.0%	—
Gemma 3 (12B)	1,840	72	96.1%	Context-length limit

C.6 Full accuracy, evidence recall, and paired contrasts

Table 7 reports answer accuracy and evidence recall for every model and distractor condition with valid counts; Table 8 reports paired accuracy drops relative to the unrelated-filler baseline. Claude's $+0.111$ same-format drop is the only paired contrast whose reported 95% confidence interval excludes zero. GPT-4.1 has a smaller point-estimate drop in the same direction ($+0.033$), but its confidence interval includes zero. Notably, GPT-4.1's evidence recall remains at 1.000 across all conditions: when it errs, it cites the correct record but extracts an incorrect answer, indicating an evidence-use rather than a retrieval failure.

Table 7: Answer accuracy and evidence recall by model and distractor condition ($N_{\text{records}} = 300$, ∼20K tokens, errors excluded). Gemma counts and values are approximate due to the 96.1% error rate and should be interpreted with caution.

Model	Distractor condition	Accuracy	Evidence recall	Valid n
Claude Sonnet 4	Unrelated filler	1.000	1.000	51
Claude Sonnet 4	Same domain, different attribute	1.000	1.000	36
Claude Sonnet 4	Same format, different entity	0.889	0.889	36
Claude Sonnet 4	Near-alias entity	1.000	1.000	32
GPT-4.1	Unrelated filler	1.000	1.000	60
GPT-4.1	Same domain, different attribute	1.000	1.000	60
GPT-4.1	Same format, different entity	0.967	1.000	60
GPT-4.1	Near-alias entity	0.983	1.000	60
Gemma 3 (12B)	Unrelated filler	1.000	1.000	∼22
Gemma 3 (12B)	Same domain, different attribute	1.000	1.000	∼20
Gemma 3 (12B)	Same format, different entity	∼0.85	∼0.85	∼16
Gemma 3 (12B)	Near-alias entity	∼0.89	∼0.89	∼14

Table 8: Paired accuracy drop relative to the unrelated-filler baseline (positive = harder; paired bootstrap, $5{,}000$ iterations, errors excluded). Gemma is omitted: sparse matched pairs ($n = 4$–5) preclude reliable paired testing.

Model	Comparison condition	Accuracy drop Δ	95% CI	Significant?
Claude Sonnet 4	Same domain, different attribute	0.000	$[0.000, 0.000]$	No
Claude Sonnet 4	Same format, different entity	$+0.111$	$[+0.028, +0.222]$	Yes
Claude Sonnet 4	Near-alias entity	0.000	$[0.000, 0.000]$	No
GPT-4.1	Same domain, different attribute	0.000	$[0.000, 0.000]$	No
GPT-4.1	Same format, different entity	$+0.033$	$[0.000, +0.083]$	No
GPT-4.1	Near-alias entity	$+0.017$	$[0.000, +0.050]$	No

Across the four condition-level means, the descriptive Spearman correlations between the lexical-overlap proxy and mean accuracy are $\rho = -0.258$ for Claude, $\rho = -0.738$ for GPT-4.1, and $\rho = -0.738$ for Gemma. Because these coefficients are based on only four condition-level observations and include tied accuracy values, they are descriptive summaries rather than inferential evidence.

C.7 Limitations and adversarial-RAG implications

Several caveats qualify these results. Claude's 35.4% error rate raises a survivorship concern—valid calls may under-represent the hardest instances—so the effect should be replicated on a higher-

throughput endpoint. Gemma is underpowered ($72/1{,}840$ valid); its aggregate pattern is suggestive but not conclusive, and a dedicated run at shorter context ($N \approx 64$–128) would provide more reliable estimates. The near-alias manipulation tops out at Jaccard 0.317, which frontier models handle well; exact-paraphrase or same-entity/different-date distractors could reveal stronger effects. Because the entities are drawn from popular fiction, pretraining memorization may make the task easier than novel-entity retrieval in production RAG, so the surprising same-format-over-near-alias ordering warrants replication with non-fictional entities. Finally, the present analysis evaluates the four distractor conditions at $N = 300$ and therefore does not estimate an interaction between context length and distractor construction. The observed same-format drop suggests a potential vulnerability: injecting structurally similar records containing plausible alternative answers may degrade accuracy even when the gold evidence remains present. In this experiment, the same-format condition produced a larger point-estimate drop than the near-alias condition, although that difference was not tested directly. The results motivate defenses that diversify record formats, attach provenance metadata, and de-rank suspiciously repetitive templates. The present study does not estimate how this effect changes with context length.

D Attention-distribution diagnostics and dose-response: from assumption to mechanism

Theorem 1's bound rests on a specific structural assumption: that distractor attention logits are independent, identically distributed Gaussian noise. This section tests that assumption directly on an open-weight model and then asks two follow-up questions that an internal diagnostic cannot answer on its own: does the assumption violation correspond to behavioral failure, and, if raw context length alone does not explain failure, what does? These findings complement, rather than replace, the empirical studies reported in the main paper (Empirical study: retrieval decay with embedded hard negatives and Empirical study: distractor confusability at fixed context length).

We use Gemma-3-4B-IT (Gemma Team, 2025) with `output_attentions` enabled, run locally via Hugging Face `transformers`. Each trial builds a needle-in-haystack prompt: one evidence sentence stating a random secret code, embedded among N distractor sentences, followed by a question asking for the code. For each of the 34 layers and 8 attention heads (272 layer-head pairs), we sum the final answer-generating token's post-softmax attention weights over the tokens belonging to each distractor sentence, producing a sentence-level attention-mass value. We analyze the logarithm of this quantity after robust median/MAD standardization. Because sentence aggregation occurs before the logarithm, this statistic is a sentence-level log-attention score rather than a recovered raw attention logit. We sweep $N \in \{20, 50, 100\}$ with 10 trials each for the attention analysis (limited by the quadratic memory cost of storing full attention weights), and separately test purely behavioral accuracy (no attention extraction, hence much cheaper) at larger N and under varying distractor content.

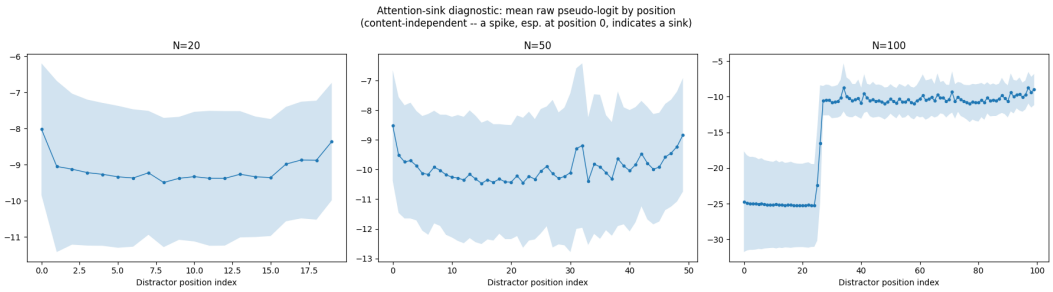


Figure 4: Mean attention by sentence position only, independent of content, pooled across trials/layers/heads.

Figure 4 diagnoses a positional confound at every N tested: attention depends measurably on position alone. At $N = 20$ and $N = 50$ this reproduces the U-shaped "lost in the middle" effect reported by Liu et al. (2024); at $N = 100$ it instead shows a sharp step, consistent with Gemma-3's architecture interleaving sliding-window and global-attention layers.

Table 9: Attention distribution tests by N (34 layers $\times$ 8 heads, 10 trials per N).

	$N = 20$	$N = 50$	$N = 100$
Gaussianity (marginal distribution)			
Shapiro–Wilk p	4.0×10^{-26}	3.5×10^{-33}	$< 10^{-300}$
Anderson–Darling stat.	205.9	1298.2	35768.7
Excess kurtosis (95% CI)	2.70 [2.35, 3.09]	6.57 [6.04, 7.03]	9.12 [8.83, 9.39]
Independence (lags 1–10)			
Ljung–Box Q	3.74	11.50	490.60
Ljung–Box p	0.958	0.320	$< 10^{-99}$
Extreme value (max statistic vs. Gumbel)			
KS statistic	0.204	0.622	0.318
KS p	1.2×10^{-99}	$< 10^{-300}$	2.2×10^{-244}
Empirical mean max (std.)	2.31	4.18	3.02
Leading-order scale $\sqrt{2 \log N}$	2.45	2.80	3.04

The Gaussianity tests reject normality at every N, and the KS tests reject an exact Gumbel fit at every N for Gemma-3 (Table 9). The Ljung–Box test does not reject independence at $N = 20$ or $N = 50$ ($p = 0.958$ and $p = 0.320$, respectively), but rejects it strongly at $N = 100$ ($p < 10^{-99}$). Thus, the Gaussian i.i.d. abstraction is not quantitatively exact, although the nature of the departure changes with context length. Heavy tails may increase extreme-score risk, whereas dependence can either amplify or suppress it. We therefore treat Theorem 1 as a diagnostic baseline rather than a calibrated quantitative bound for these models.

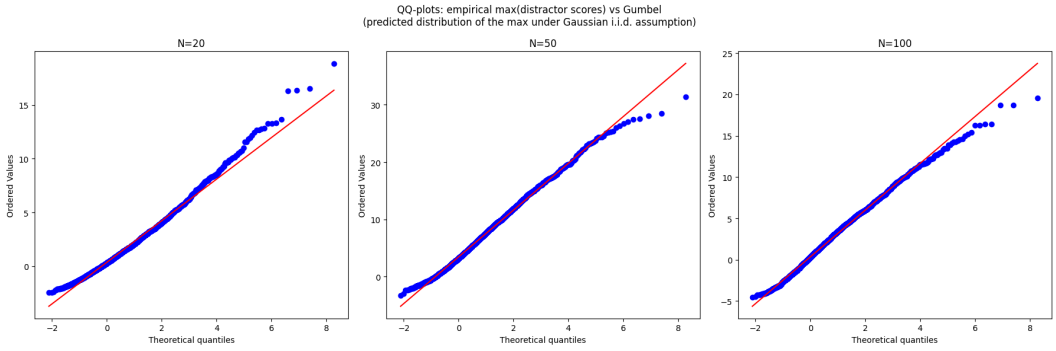


Figure 5: Q–Q plots of the empirical maximum standardized distractor scores against the chosen asymptotic Gumbel approximation, by N.

Figure 5 compares the empirical maximum standardized distractor scores with an asymptotic Gumbel approximation. Under an exact i.i.d. standard-Gaussian model, the finite-N maximum has CDF $F_N(x) = \Phi(x)^N$; a Gumbel distribution arises only asymptotically after appropriate centering and scaling. Rejection of the selected Gumbel approximation at $N \in \{20, 50, 100\}$ therefore does not by itself constitute a direct finite-sample test of the Gaussian i.i.d. hypothesis.

Despite this, the theorem's own margin-growth prediction (Figure 6) tracks the data reasonably closely at two of three points—2.31 vs. 2.45 at $N = 20$ and 3.02 vs. 3.04 at $N = 100$—showing the abstraction captures the right order of magnitude even where its distributional assumptions do not hold exactly.

A natural question is whether the findings above reflect a general property of transformer attention or an artifact specific to Gemma-3's architecture, which interleaves sliding-window (local) attention layers with global-attention layers. We repeated the identical experiment on Llama-3-8B-Instruct (Grattafiori et al., 2024), which uses full causal (global) attention in every layer, with no sliding-window mechanism. Table 10 summarizes the two architectures.

Figure 7 shows a smooth positional profile at all three N, with no sharp discontinuity at $N = 100$. The absence of the step in Llama-3 is consistent with an architecture-specific explanation for the

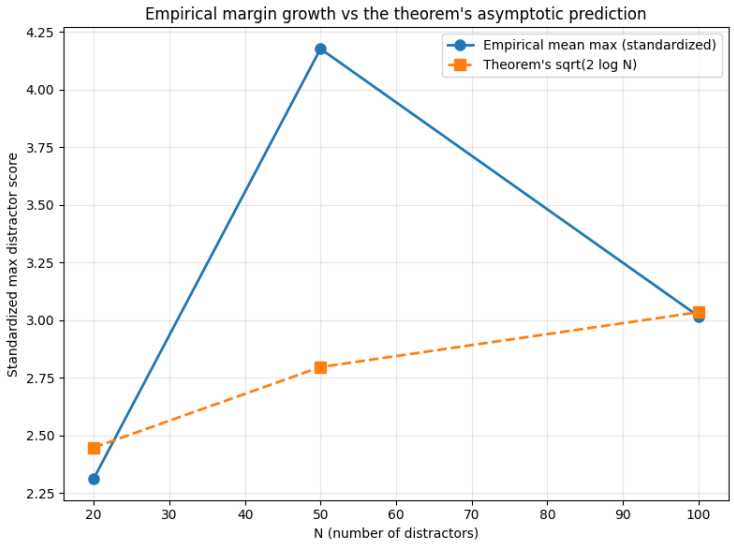


Figure 6: Empirical mean of the standardized maximum distractor score vs. the theorem's own asymptotic prediction, $\sqrt{2 \log N}$.

Table 10: Architectural comparison of the two models tested.

	Gemma-3-4B-IT	Llama-3-8B-Instruct
Parameters	∼4B	∼8B
Layers	34	32
Attention heads	8	32
Attention pattern	Mixed: sliding-window (local) + global layers	Full causal (global) attention, every layer

Gemma-3 pattern. However, this two-model comparison does not isolate sliding-window/global layer mixing as the cause, because the models also differ in size, training, tokenization, and other architectural details.

Table 11: Attention distribution tests on Llama-3-8B-Instruct, by N (32 layers $\times$ 32 heads, 10 trials per N).

	$N = 20$	$N = 50$	$N = 100$
Gaussianity (marginal distribution)			
Excess kurtosis (95% CI)	2.32 [2.01, 2.71]	0.98 [0.93, 1.03]	0.89 [0.86, 0.92]
Independence (lags 1–10)			
Ljung–Box Q	3.92	25.29	104.63
Ljung–Box p	0.951	0.0048	$< 10^{-6}$
Extreme value (max statistic vs. Gumbel)			
KS statistic	0.195	0.228	0.273
Empirical mean max (std.)	2.29	2.63	2.96
Leading-order scale $\sqrt{2 \log N}$	2.45	2.80	3.04

The cross-architecture comparison reveals model-specific departures from the Gaussian i.i.d. abstraction (Tables 9 and 11). Gemma-3's excess kurtosis increases with N (2.70 $\rightarrow$ 9.12), whereas Llama-3's decreases (2.32 $\rightarrow$ 0.89). At $N = 100$, the Ljung–Box test rejects independence for both models. Because the analyses pool different numbers of layer-head pairs, however, the raw Q statistics are not directly comparable as measures of violation severity. For Gemma-3, the reported KS p-values reject an exact Gumbel fit. Table 11 reports KS statistics for Llama-3 but not the corresponding p-values, so we do not make a formal rejection claim for Llama-3 without those values.

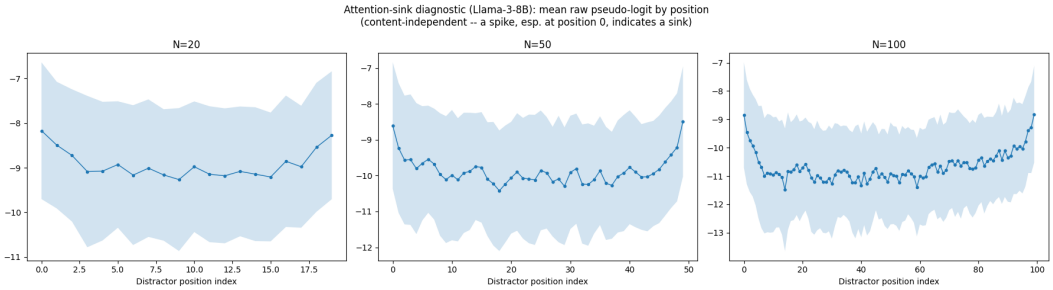


Figure 7: Attention-sink diagnostic on Llama-3-8B-Instruct (see Figure 4 for Gemma-3).

Despite these departures, the empirical mean maximum for Llama-3 is 93%, 94%, and 97% of the $\sqrt{2 \log N}$ prediction at $N = 20$, 50, and 100, respectively (Figure 8). Thus, the asymptotic scale remains a useful order-of-magnitude reference, even though the exact distributional assumptions are not satisfied.

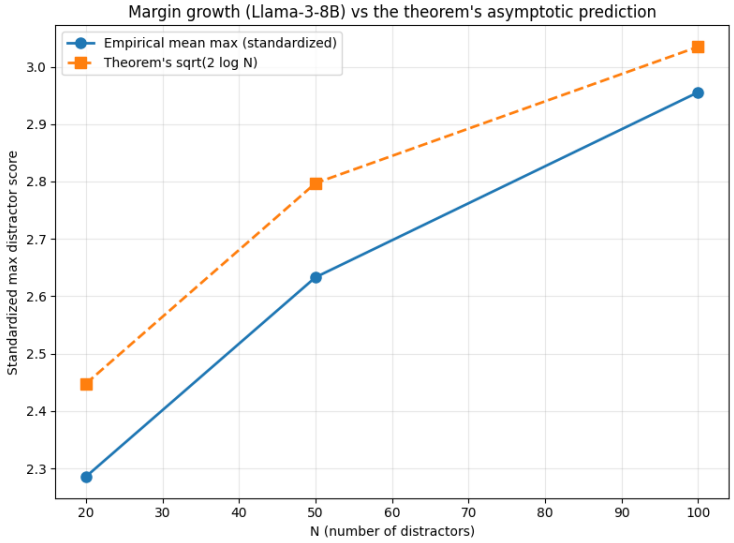


Figure 8: Margin growth on Llama-3-8B-Instruct vs. the theorem's asymptotic prediction (cf. Figure 6 for Gemma).

Does this internal violation translate into behavioral failure? For each trial we additionally generated the model's real answer and checked whether it contained the true code. Retrieval accuracy was 100% at every N (30/30 trials total; Figure 9), while excess kurtosis rose monotonically over the same range. For this task—simple single-fact retrieval against generic filler—the internal assumption violation is not sufficient on its own to produce behavioral failure.

If neither the assumption violation nor N alone predicts failure, does raw context length matter at all? Extending purely behavioral testing (no attention extraction, hence tractable at much larger N) with the same generic-filler distractors to $N = 100, 200, 400$ (up to ∼5,400 tokens), accuracy remained 100% throughout. Increasing the number of generic-filler distractors from 100 to 400 did not reduce observed accuracy for this model on this task. This result applies only to the tested context range and distractor construction.

We therefore varied *what* the distractors said rather than how many there were, holding $N \in \{20, 50, 100\}$ fixed. Low-similarity distractors (generic, topically unrelated filler) and medium-similarity distractors (the same numeric-code format, distinguished by a different label such as "backup code") both left accuracy at 100% at every N tested. High-similarity distractors—identical phrasing to the evidence sentence, differing only in the number—collapsed accuracy to near zero

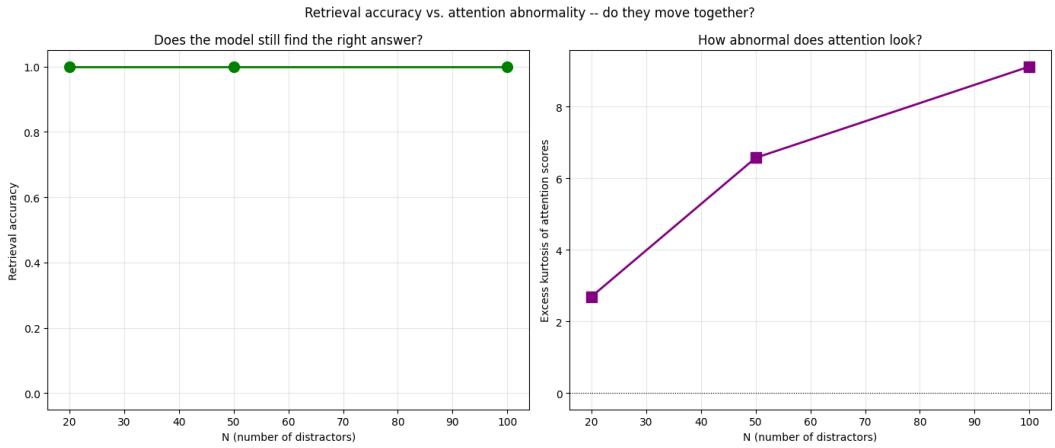


Figure 9: Retrieval accuracy (left) vs. excess kurtosis of the attention distribution (right), by N.

at every N tested instead (0.00, 0.07, and 0.00 at $N = 20$, 50, and 100 respectively). This approximately N-invariant pattern over the three tested values supports the paper's distinction (Definition 2) between raw distractor count N and effective distractor count N_{eff}: over the tested range, confusability is more predictive of failure than raw distractor count. Because this three-condition low/medium/high comparison cannot determine how much confusability is required, we then fixed $N = 50$ total distractors and swept k, the number of high-similarity decoys mixed into an otherwise generic-filler haystack, from 0 to 50 (Table 12, Figure 10).

Table 12: Retrieval accuracy versus the number k of confusable decoys, with $N_{\text{dist}} = 50$ total distractors fixed.

k	0	1	2	3	5	10	20	35	50
Accuracy	1.00	0.67	0.33	0.40	0.07	0.13	0.07	0.07	0.00

Each point is based on 15 trials, so one additional correct response changes the reported accuracy by $1/15 \approx 0.067$. Accordingly, the adjacent changes from $k = 2$ (0.33) to $k = 3$ (0.40), and from $k = 5$ (0.07) to $k = 10$ (0.13), each correspond to only one additional correct response. These local reversals should therefore not be interpreted as evidence of a non-monotonic underlying effect. The robust feature is the overall steep initial decline followed by a low plateau.

Accuracy drops sharply within the first two confusable decoys ($k = 0$: 1.00, $k = 1$: 0.67, $k = 2$: 0.33). A single confusable decoy among fifty distractors reduces observed accuracy by approximately one third, and two reduce it by approximately two thirds. From $k = 5$ through $k = 35$, observed accuracy remains in the 7%–13% range across a sevenfold increase in k. At $k = 50$, when every distractor is confusable, observed accuracy is $0/15$. Thus, most of the observed degradation occurs after introducing a small number of confusable decoys, while additional decoys produce comparatively little further change.

In qualitative inspection, some failures under high similarity or larger k included hedging behaviors that were not observed at $k = 0$, such as listing multiple candidate codes or generating an unrequested confidence statement or rationale. Because these behaviors were not coded systematically, we treat them as anecdotal indications of uncertainty rather than as a measured result.

Taken together, these results provide a coherent but task-specific picture. The Gaussian i.i.d. assumption is not exact, and the form of its violation differs across the two architectures. Nevertheless, the theorem's margin-growth prediction remains close in order of magnitude to the observed maxima. In this retrieval task, behavioral failure tracks the number of textually confusable decoys more closely than raw context length over the tested range. Only a small number of such decoys is needed to produce most of the observed degradation. These findings support interpreting N in the bound as an

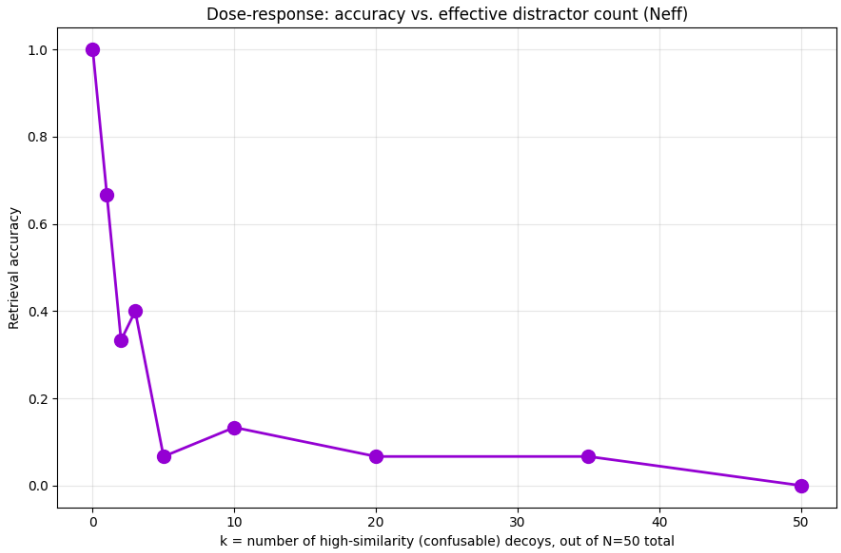


Figure 10: Retrieval accuracy versus the number k of confusable decoys, with $N_{\text{dist}} = 50$ total distractors fixed.

effective distractor count N_{eff}, while not establishing that raw context length is irrelevant for other tasks, models, or longer-context regimes.

E LOST-IN-THE-MIDDLE REANALYSIS

Definition 1 defines the poisoning curve $P(N) = A(N_0) - A(N)$: the drop in accuracy as context length N grows. This has so far been supported partly by citing prior work (Liu et al., 2024); here we re-run a version of their setup directly on our own model to report a concrete, reproducible number.

We use the publicly released data from Liu et al. (2024), which pair real questions from the Natural Questions dataset (Kwiatkowski et al., 2019) with Wikipedia passages retrieved using Contriever-MSMARCO (Izacard et al., 2021). Each example has one gold passage containing the true answer and several distractor passages. We sampled 40 fixed questions, reused across every condition so results are paired rather than confounded by question difficulty, and evaluated them on Gemma-3-4B-IT (Gemma Team, 2025) (Hugging Face `transformers`, greedy decoding). A model's answer is scored correct if it contains the known gold answer string.

Two conditions were tested. In the position condition, $N_{\text{docs}} = 20$ documents are fixed and the gold passage is moved to zero-indexed position 0, 4, 9, 14, or 19. In the scaling condition, the gold passage is held at zero-indexed position 0, which had the highest observed accuracy in the first condition, while $N_{\text{docs}} \in \{10, 20, 30\}$ is varied. The second condition therefore measures the effect of document count while holding the gold passage's position fixed.

Accuracy is highest when the gold passage is first (52.5%) and ranges from 32.5% to 40.0% at the other tested positions, a decrease of 12.5–20.0 percentage points (Figure 11). This provides descriptive evidence of a position effect in this sample and is consistent with Liu et al. (2024).

Table 13: Accuracy and poisoning curve $P(N) = A(10) - A(N)$, gold passage fixed at position 0.

N	10	20	30
$A(N)$	0.550	0.525	0.550
$P(N)$	0.000	0.025	0.000

With the gold passage fixed at position 0, which had the highest observed accuracy, $P(N)$ remains close to zero across $N \in \{10, 20, 30\}$ (Table 13, Figure 12). We therefore observe no monotonic

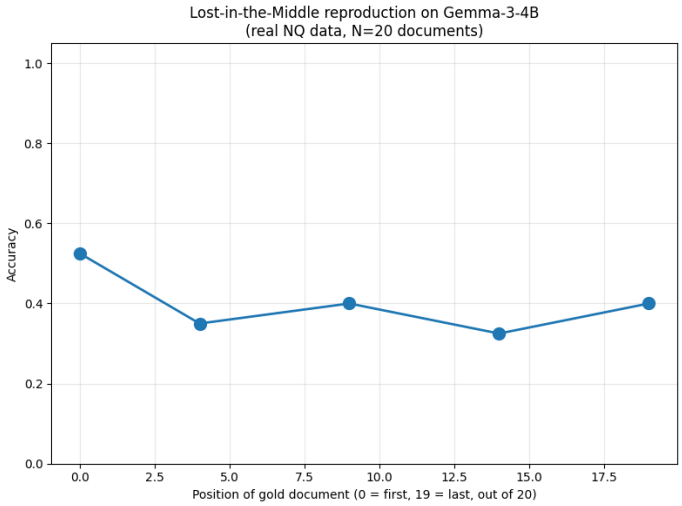


Figure 11: Accuracy vs. position of the gold document, $N = 20$, 40 real NQ questions.

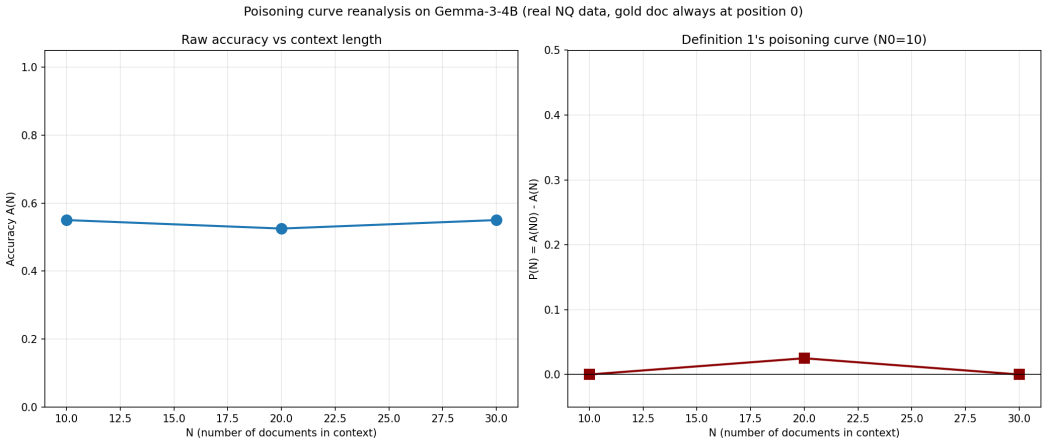


Figure 12: Accuracy and $P(N)$ vs. context length, gold passage fixed at position 0.

decrease in accuracy with document count over this range. Together with the position experiment, these results indicate that position has a larger observed effect than raw document count for this model and within the tested range. They do not establish that position remains the dominant factor at larger context sizes or for other models. The result is compatible with Theorem 1's emphasis on effective distractor exposure and Definition 2, but this experiment does not measure N_{eff} directly.

Two limitations are worth noting. First, the study uses only 40 questions, so the estimates have substantial uncertainty. Second, the scaling condition tests only up to $N = 30$ documents. Whether $P(N)$ remains near zero at larger values of N, or whether position and length effects interact at greater scale, remains unresolved.

F Empirical Test of the Gate-Mitigation Prediction

Proposition 1 predicts that replacing a raw candidate set of N context chunks with a bounded retrieved set of $K \ll N$ chunks trades context-length interference against a retrieval-error penalty $\hat{\eta} = 1 - \text{Recall@}K$. We test this prediction directly, following the reporting protocol in Experimental setup and metrics: we measure answer accuracy and evidence recall separately, and we isolate the model's use of retrieved evidence from the quality of retrieval itself. We emphasize the scope of

the experiment: it tests the qualitative tradeoff described by the proposition, but it does not validate the theorem's Gaussian-logit assumptions or its $\sqrt{\log N}$ asymptotic rate.

F.1 DESIGN

Benchmark. We use RULER's `qa_2` task (Hsieh et al., 2024), which is built on HotpotQA (Yang et al., 2018). Each instance is a multi-hop question whose two gold supporting paragraphs are embedded among distractor paragraphs presented as numbered `Document` n`:` blocks. We construct instances directly from the HotpotQA distractor development set to preserve HotpotQA's gold supporting-fact titles, which allows us to measure evidence recall rather than relying on an answer-string proxy.

Starting from each question's native two gold and eight distractor paragraphs, we add distractor paragraphs drawn from a global pool until a target token length is reached. We evaluate seven lengths, $\{4, 8, 16, 32\}$K and $\{128, 256, 512\}$K tokens, using the same 100 questions at every length so that comparisons are paired across both conditions and lengths. Lengths are measured with the `o200k_base` tokenizer. The median raw candidate count N grows from 29 chunks at 4K to 3,711 chunks at 512K (Table 14).

Conditions. Each instance is answered under two conditions that differ only in the documents supplied. The **raw** condition passes all N documents in source order. The **gated** condition scores each document against the question using BM25 (Robertson & Zaragoza, 2009), retains the top-K documents, and restores them to their original source order before answering. Restoring source order ensures that the intervention changes context length without also changing document ordering.

Both conditions use the same prompt template, system instruction, and decoding configuration: temperature 0, top-p 1, a maximum of 64 generated tokens, and one generation call. We deliberately omit a quote-the-evidence-first step from the gated condition because such a step would confound retrieval gating with an additional inference procedure.

Gate and K selection. Each `Document` n`:` block is treated as one retrieval chunk. Queries and documents use identical lowercasing and tokenization for BM25 scoring. Because the benchmark is multi-hop, the gate succeeds only when it retrieves all annotated supporting documents:

$$\mathrm{Hit}_i(K) = \mathbf{1}\{G_i \subseteq R_i(K)\}\,. \tag{34}$$

The estimated retrieval penalty is

$$\hat{\eta} = 1 - \frac{1}{M}\sum_{i=1}^{M} \mathrm{Hit}_i(K). \tag{35}$$

On a disjoint 100-question development split, we sweep $K \in \{2, 4, 8, 16\}$ and select the smallest value achieving all-support Recall@$K \geq 0.90$ at every length through 32K. This procedure selects $K = 16$. Retrieval and selection of K are performed offline and require no test-set model calls.

Model. All answers are generated by GPT-4.1 (`gpt-4.1_2025-04-14`) at temperature 0. We verified that the serving endpoint accepts inputs of at least 520K tokens, and no raw prompt was truncated. The study consists of 1,400 generations, forming 700 paired raw–gated comparisons (100×7 pairs), with zero endpoint errors.

F.2 METRICS

We report normalized exact match (EM) and token F1 using the official HotpotQA scorer. The primary effect is the paired difference

$$\Delta\mathrm{EM}(N, K) = A_{\mathrm{gate}} - A_{\mathrm{raw}}, \tag{36}$$

with 95% confidence intervals computed from 10,000 paired bootstrap samples and paired binary outcomes evaluated using an exact McNemar test.

Table 14: GPT-4.1 on RULER qa_2/HotpotQA: raw long context versus a BM25 top-K retrieval gate with $K = 16$, frozen on a disjoint development split. Each length uses the same 100 questions at temperature 0. The retrieval penalty is $\hat{\eta} = 1 - \text{Recall@}K$. ΔEM is the paired gated-minus-raw difference with a 10,000-sample paired-bootstrap 95% confidence interval. Δ_{use} is the gated-minus-raw difference restricted to instances for which all annotated supporting evidence was retrieved. McNemar p is the exact two-sided test.

Ctx	Median raw N	K	Recall @K	$\hat{\eta}$	Raw EM	Gate EM	ΔEM [95% CI]	Raw F1	Gate F1	Gate EM| h	Raw EM| h	Δ_{use}	McNemar p
4K	29	16	0.94	0.06	0.660	0.640	-0.020 [-0.070, $+0.030$]	0.768	0.751	0.681	0.660	0.021	0.727
8K	58	16	0.94	0.06	0.650	0.620	-0.030 [-0.080, $+0.020$]	0.768	0.738	0.660	0.660	0.000	0.453
16K	115	16	0.94	0.06	0.680	0.630	-0.050 [-0.110, $+0.010$]	0.792	0.723	0.670	0.691	-0.021	0.180
32K	232	16	0.92	0.08	0.650	0.650	$+0.000$ [-0.050, $+0.050$]	0.782	0.757	0.696	0.674	0.022	1.000
128K	930	16	0.86	0.14	0.590	0.640	$+0.050$ [-0.010, $+0.110$]	0.724	0.746	0.709	0.640	0.070	0.227
256K	1857	16	0.82	0.18	0.590	0.580	-0.010 [-0.090, $+0.070$]	0.715	0.694	0.683	0.622	0.061	1.000
512K	3711	16	0.79	0.21	0.570	0.590	$+0.020$ [-0.060, $+0.100$]	0.704	0.705	0.709	0.595	0.114	0.815

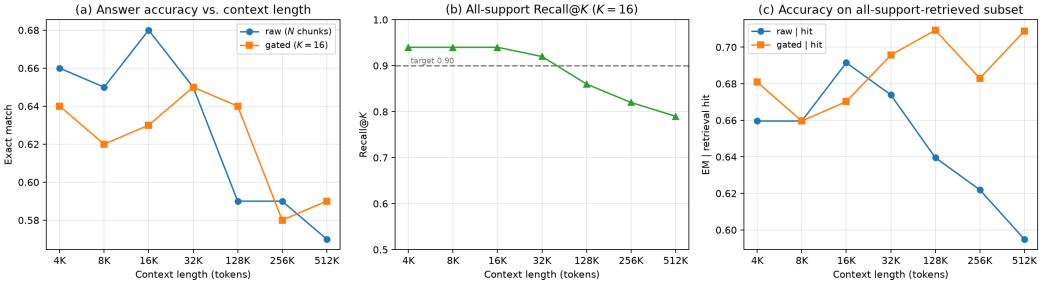


Figure 13: GPT-4.1 with $K = 16$. **(a)** Raw accuracy is approximately stable through 32K and then decreases with context length, while gated accuracy varies less. **(b)** A fixed $K = 16$ meets the 0.90 all-support recall target through 32K, after which recall decreases as the candidate set grows. **(c)** On instances for which the gate retrieves all annotated supporting evidence, raw accuracy decreases with context length while gated accuracy remains comparatively stable, providing evidence of improved evidence use after competing context is removed.

To separate evidence use from retrieval quality, we also report conditional accuracy on the retrieval-hit subset and its difference

$$\Delta_{\text{use}} = A_{\text{gate}|\text{hit}} - A_{\text{raw}|\text{hit}}. \tag{37}$$

On this subset, the gate is known to have retained all annotated supporting evidence. A positive Δ_{use} is therefore consistent with a benefit from reducing competing context rather than merely from successful retrieval.

F.3 Results

Table 14 and Figure 13 support four findings.

(1) No measurable poisoning through 32K. Up to 32K, raw EM remains approximately stable ($0.66 \to 0.65$), and the gate provides no statistically detectable net benefit. The observed ΔEM values remain within $[-0.05, +0.05]$, all reported confidence intervals include zero, and McNemar $p \geq 0.18$.

(2) Degradation appears beyond 32K. From 32K to 512K, raw EM decreases from 0.65 to 0.57 and raw F1 decreases from 0.78 to 0.70. No comparable downward trend is observed among the tested lengths from 4K through 32K. The gated condition exhibits a smaller overall decrease across the longest tested contexts.

(3) The gate improves evidence use on retrieval-hit instances. Figure 13(c) restricts evaluation to questions for which the gate retrieves all annotated supporting evidence. On this subset, raw

Table 15: Scaling K to restore recall at extreme lengths. Only the gated arm is re-run; raw predictions are cached and unchanged. Each length-specific value of K is selected on the development split to achieve Recall@$K \geq 0.90$.

Ctx	K	Recall@K	Raw EM	Gate EM	ΔEM [95% CI]	Δ_{use}	McNemar p
128K	16	0.86	0.590	0.640	$+0.050$ $[-0.010, +0.110]$	0.070	0.227
	64	0.94	0.590	0.620	$+0.030$ $[-0.020, +0.080]$	0.043	0.453
256K	16	0.82	0.590	0.580	-0.010 $[-0.090, +0.070]$	0.061	1.000
	128	0.94	0.590	0.640	$+0.050$ $[-0.010, +0.120]$	0.064	0.227
512K	16	0.79	0.570	0.590	$+0.020$ $[-0.060, +0.100]$	0.114	0.815
	192	0.93	0.570	0.640	$+0.070$ $[+0.000, +0.140]$	0.097	0.092

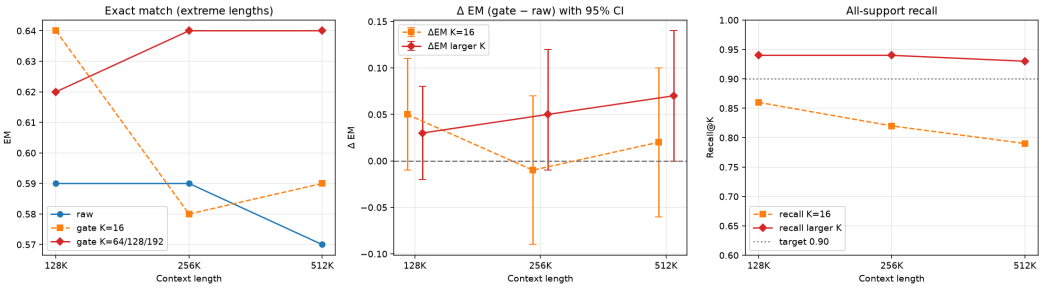


Figure 14: Comparison between fixed $K = 16$ and a retrieval budget scaled to maintain all-support Recall@K of at least 0.90 at extreme context lengths. **Left:** gated accuracy remains comparatively stable under the scaled retrieval budget while raw accuracy decreases. **Middle:** the paired gated-minus-raw difference becomes more positive after the retrieval penalty is reduced. **Right:** the scaled values of K maintain recall above the target where fixed $K = 16$ does not.

accuracy decreases from 0.69 at 16K to 0.595 at 512K, whereas gated accuracy remains near 0.71. The corresponding Δ_{use} values are $+0.070$, $+0.061$, and $+0.114$ at 128K, 256K, and 512K, respectively. Because both conditions contain all annotated gold evidence on this subset, the pattern is consistent with a benefit from removing competing distractor context, rather than merely from successful retrieval.

(4) A fixed K incurs a growing retrieval penalty. With K fixed at 16, all-support recall decreases from 0.94 to 0.79 as the raw candidate count grows to approximately 3,700 chunks. Correspondingly, $\hat{\eta}$ increases from 0.06 to 0.21. The evidence-use benefit is therefore offset by the 14%–21% of instances for which the bounded gate omits required evidence, and the net ΔEM remains near zero at the longest contexts. This pattern matches the qualitative tradeoff in Proposition 1: reducing the candidate set from N to K can improve evidence use, but that benefit is offset when the retrieval penalty $\hat{\eta}$ grows with the candidate set.

F.4 Scaling K to Preserve Evidence Recall

To separate the evidence-use benefit from the retrieval penalty, we re-run only the gated condition at 128K, 256K, and 512K with a length-specific value of K selected on the development split to restore all-support Recall@$K \geq 0.90$. The resulting values are $K = 64$, 128, and 192, respectively. The raw predictions are unchanged and reused from the primary experiment. Even at $K = 192$, the gated prompt remains approximately 25K tokens, compared with approximately 504K tokens in the raw condition.

As shown in Table 15 and Figure 14, restoring recall makes the net gate advantage more positive at the longest tested contexts. At 512K, increasing K from 16 to 192 raises gated EM from 0.59 to 0.64, while raw EM is 0.57. The corresponding ΔEM increases from $+0.02$ to $+0.070$, with a 95% confidence interval of $[+0.000, +0.140]$ and McNemar $p = 0.092$. At 256K, ΔEM increases from -0.01 to $+0.050$.

At 128K, fixed $K = 16$ already retains most annotated support (recall 0.86), and increasing K slightly reduces $\Delta\mathrm{EM}$ from $+0.05$ to $+0.03$, consistent with reintroducing additional distractors. In the logistic model

$$\text{correct} \sim \text{condition} \times \log_2 N, \tag{38}$$

the estimated gate-by-length interaction increases from $+0.035$ with fixed $K = 16$ to $+0.084$ under the scaled retrieval budget. These point estimates suggest that the retrieval budget may need to grow with the candidate set to maintain evidence recall, rather than being held fixed independently of N.

F.5 Discussion and Limitations

The experiment supports the gate-mitigation prediction within the tested regime. For GPT-4.1, no measurable degradation is observed through 32K. At longer contexts, raw accuracy decreases, while the retrieval-hit analysis indicates that the gate can improve the model's use of evidence that remains available. The net accuracy benefit becomes more positive when K is increased to keep the retrieval penalty $\hat{\eta}$ small. Together, these results are consistent with the $N \to K$ versus $\hat{\eta}$ tradeoff described by Proposition 1.

Several limitations bound these claims. First, with $M = 100$ questions, the strongest net effect at 512K remains borderline: the confidence interval has a lower bound of zero and McNemar $p \approx 0.09$. A larger evaluation set would improve precision; under standard square-root scaling, doubling the sample size would reduce standard errors by approximately a factor of $1/\sqrt{2}$, not by one half.

Second, we study a single hosted model. Other models may enter the degradation regime at different context lengths, so replication across model families and deployment settings is needed to assess the generality of finding (2).

Third, contexts are constructed and measured using a tokenizer proxy rather than the proprietary serving tokenizer. The reported lengths are therefore internally consistent but may not exactly equal server-side token counts. In addition, temperature-0 hosted decoding is not necessarily perfectly deterministic.

Finally, as stated at the outset, these results test the proposition's qualitative tradeoff rather than the theorem's distributional assumptions or asymptotic rate.