Positional Fragility in LLMs: How Offset Effects Reshape Our Understanding of Memorization Risks

Yixuan Xu[1,2] **Antoine Bosselut**[2] **Imanol Schlag**[1]
[1]ETH AI Center, ETH Zürich
[2]School of Computer and Communication Sciences, EPFL
`firstname.lastname@inf.ethz.ch` `antoine.bosselut@epfl.ch`

Alvorecer721/Positional_Fragility_LLM

Abstract

Large language models are known to memorize parts of their training data, posing risk of copyright violations. To systematically examine this risk, we pretrain language models (1B/3B/8B) from scratch on 83B tokens, mixing web-scale data with public domain books used to simulate copyrighted content at controlled frequencies at lengths at least ten times longer than prior work. We thereby identified the offset effect, a phenomenon characterized by two key findings: (1) verbatim memorization is most strongly triggered by short prefixes drawn from the beginning of the context window, with memorization decreasing counterintuitively as prefix length increases; and (2) a sharp decline in verbatim recall when prefix begins offset from the initial tokens of the context window. We attribute this to *positional fragility*: models rely disproportionately on the earliest tokens in their context window as retrieval anchors, making them sensitive to even slight shifts. We further observe that when the model fails to retrieve memorized content, it often produces degenerated text. Leveraging these findings, we show that shifting sensitive data deeper into the context window suppresses both extractable memorization and degeneration. Our results suggest that positional offset is a critical and previously overlooked axis for evaluating memorization risks, since prior work implicitly assumed uniformity by probing only from the beginning of documents or training sequences.

1 Introduction

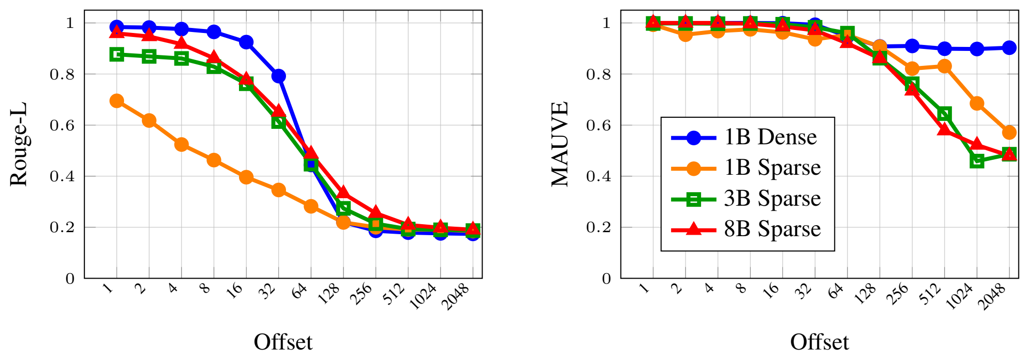
Large language models (LLMs) have demonstrated the capacity to reproduce substantial portions of their training data verbatim, raising serious concerns about copyright violations [Chang et al., 2023, Karamolegkou et al., 2023] and privacy breaches [Huang et al., 2022]. This capability has fueled increasing legal challenges, with high-profile cases, such as the New York Times lawsuit against OpenAI, highlighting the potential misuse of proprietary content [Freeman et al., 2024]. Industry leaders have not denied these risks, and instead called for regulatory exemptions that would allow LLM training on copyright-protected material [Brodkin, 2025]. Against this backdrop, understanding how LLMs memorize and regurgitate legally sensitive content is no longer a niche technical concern, but a central requirement for the responsible and legal development of LLMs [Rosenthal and Veraldi, 2025] and informed policymaking on this subject.

Existing work has begun to study memorization in LLMs, but key gaps remain. First, most mitigation strategies are applied post-training via fine-tuning, decoding constraints, or unlearning methods. However, recent studies show that these approaches can be bypassed [Park et al., 2024, Nasr et al., 2025], indicating that they are better suited as complementary defenses rather than standalone solutions. This highlights the need to address memorization proactively during pretraining, where

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

retention first happens. Second, prior analyses typically focus on short suffixes prompted from the beginning of training sequences, implicitly assuming that memorization behavior is uniform across the context window. This methodological choice overlooks the reality that user inputs may appear at arbitrary offsets in the content window, and that memorization may vary significantly with position.

In this work, we revisit memorization from a legally motivated perspective, focusing on long-form text segments drawn from published books (via Project Gutenberg) to simulate high-risk copyright scenarios. We pretrain LLaMA-style models (1B, 3B, and 8B) from scratch on 83B tokens, combining curated book data with web-scale educational corpora. This setup grants precise control over training dynamics (such as model scale and batch size), data composition (such as exposure frequency and placement of sensitive content), and evaluation configuration (such as prefix length and position within the context window). Compared to the most extensive memorization analysis to date by Hans et al. [2025], which also targets pretraining-time mitigation, our study employs full-context sequences four times longer and evaluates suffixes at least ten times longer. Together, these design choices enable an unprecedented, comprehensive, and fine-grained memorization analysis.



(a) Rouge-L scores by offset

(b) MAUVE scores by offset

Figure 1: Positional fragility in LLMs measured with two complementary metrics: (a) Rouge-L scores quantify verbatim memorization, demonstrating a sharp decline as offset increases; (b) MAUVE scores capture overall language coherence, revealing text degeneration with increasing offset. The dense model results are derived from a 1B model trained exclusively on 10K Gutenberg sequences seen 80 times each, while sparse model results are derived specifically from the frequency-128 bucket of our FM-Probe methodology. All experiments used 500-token prefixes and evaluated on 500-token suffixes. Verbatim memorization and language coherence both degrade sharply as prefix offset increases, except for language coherence in 1B Dense case, revealing positional fragility.

Our main contributions are:

- We identify the **offset effects**: verbatim memorization is most effectively triggered by short prefixes at the beginning of the context window. Memorization recall drops sharply as prefixes are shifted further into the sequence. We attribute offset effects to **positional fragility**: LLMs are highly sensitive to positional shifts as they disproportionately rely on early tokens as retrieval anchors (§ 4).
- We show that memorization failures cause **degeneration**. Under offset or low-frequency conditions, the model produces repetitive and semantically incoherent output, establishing a link between memorization breakdown and text degeneration (§ 5).
- We show that **positional fragility can be leveraged to mitigate memorization:** shifting sensitive sequences deeper into the context window significantly reduces extractability without compromising general performance (§ 7).
- We find that **batch size influences retention:** smaller batches with more frequent updates modestly reduce memorization under identical compute budgets. (§ 8).
- We **reproduce and scale** previous studies in more controlled and legally relevant settings, using longer sequences to better reflect the risks of copyright in the real world (Appendix C).

2 Problem Setup

To facilitate the discussion around verbatim memorization we introduce the following nomenclature:

Text Segment: A sequence of 8192 tokens from the training corpus, beginning with a `BOD` token, that fits exactly the context window of the model.
Prefix x: The input text sequence provided to the model to probe memorization, represented as $x = (x_1, ..., x_n)$ with n tokens extracted consecutively from the training corpus.
True Suffix s: The text sequence that immediately follows the prefix in the original text segment, represented as $s = (s_1, ..., s_m)$ with m consecutive tokens.
Generated Suffix y: The text sequence produced by the model given prefix x, where $y = (y_1, ..., y_m)$, that matches the length of the ground truth suffix m.
Offset o: The number of tokens to skip in the text segment from the beginning-of-document token `BOD` before starting prefix extraction.

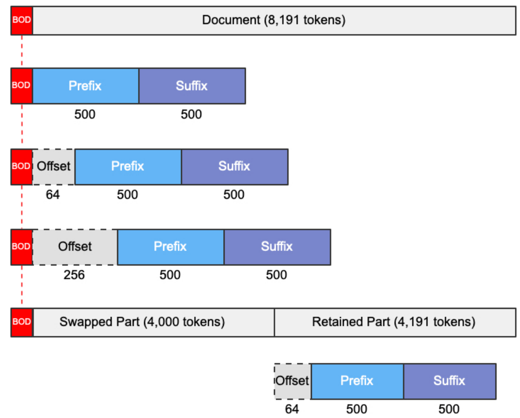


Figure 2: Illustration of prefix-suffix extraction with varying offsets from the document start in Sparse Gutenberg, and from the retained part in Swapped Gutenberg. Not to scale.

Quantifying Verbatim Memorization We follow Nasr et al. [2025]'s definition of verbatim memorization, providing a model with a prefix x from training data and measuring how closely its greedy continuation y matches the true continuation s. Prior work predominantly focuses on short prefix probing, using prefixes of size 20 [Carlini et al., 2021], 32 [Biderman et al., 2023, Huang et al., 2024, Duan et al., 2025], 50 [Karamolegkou et al., 2023, Duan et al., 2025, Ippolito et al., 2023], 50–500 [Carlini et al., 2023], and 1998 [Hans et al., 2025]. Suffix lengths are likewise short, typically ranging from 25 to 50 tokens. Nearly all evaluations extract prefixes from the beginning of the document, with only Carlini et al. [2021] sampling prefixes from arbitrary positions, but they only report the highest-recall case. To address these limitations, our evaluation spans prefix lengths from 50 to 5000 tokens, sampled at offsets ranging from 0 to 2048, with suffixes between 50 and 8000 tokens. We report similarity using three metrics that capture increasing levels of verbatim reproduction: Rouge-L (longest common subsequence relative to reference length) [Lin, 2004], LCCS (longest common contiguous substring) [Freeman et al., 2024], and EM (exact match of the entire suffix).

Measuring Text Degeneration We evaluate output quality using three complementary metrics. *Type-token ratio* (TTR) measures lexical diversity as the fraction of unique tokens in the generated suffix [Kettunen, 2014]. *MAUVE* quantifies distributional similarity between model outputs and human-written text [Pillutla et al., 2023]. As a baseline, unrelated suffix pairs yield a ROUGE-L of approximately 0.18. We also report *perplexity* on the reference suffix to assess how well the model predicts the ground-truth continuation. Given a prefix $x = (x_1, \ldots, x_n)$ and its true suffix $s = (s_1, \ldots, s_m)$, perplexity is defined as $\mathrm{PPL}(s \mid x) = \exp\left(-\frac{1}{m}\sum_{t=1}^{m} \log P(s_t \mid x, s_{<t})\right)$. Intuitively, perplexity reflects how many next-token candidates the model finds plausible at each step [Jelinek et al., 2005]. For downstream performance evaluation, we utilize `lm-eval-harness` [Gao et al., 2024].

3 Design

Data & Model We train a decoder-only transformer model following the Llama architecture [Grattafiori et al., 2024], with complete architectural and training details provided in Appendix A.2. Our training corpus combines two complementary sources: (1) Project Gutenberg[1], used to simulate potential copyright infringement via literary texts; and (2) Fineweb-Edu [Penedo et al., 2024], a curated dataset of high-quality educational web content typical in modern LLM pipelines. A 13-gram contamination analysis confirmed a negligible 0.34% overlap between the two datasets, ensuring observed memorization is attributable to our experimental controls.

[1] `https://huggingface.co/datasets/manu/project_gutenberg`

Dense Gutenberg: Extreme Memorization Study Inspired by the setup of Hans et al. [2025], we establish an extreme memorization scenario, where models are trained exclusively on our curated subset of Project Gutenberg text segments comprising 10,000 sequences of 8,192 tokens each. Each epoch constitutes a complete traversal of the entire corpus, with model checkpoints saved at logarithmically spaced intervals (powers of 2) as well as at the final 80th epoch. Since each sequence appears exactly once per epoch, the checkpoint number directly corresponds to the number of exposures to each training sequence, providing a direct measure of how repetition influences memorization behavior.

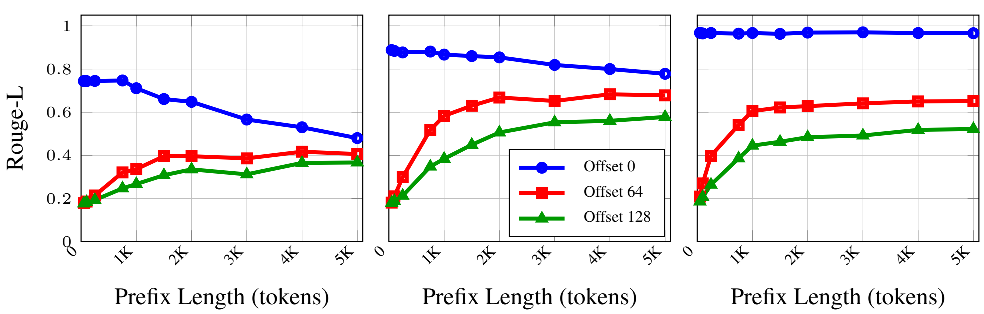
Sparse Gutenberg: Realistic Copyright Memorization Simulation To create a more realistic scenario, we designed a mixed corpus training configuration where Project Gutenberg excerpts represent just 2% (1.74B tokens) of the training data, with the remaining 98% (81.82B tokens) sourced from FineWeb-Edu. For analyzing how exposure frequency affects memorization, we developed Frequency-Varied Memorization Probe Buckets (FM-Probes), with each bucket containing 500 distinct and randomly sampled Gutenberg sequences that appear at precisely controlled frequencies ranging from 1 to 128 repetitions. By performing a single training pass through this carefully structured corpus, our FM-probe methodology enables concurrent evaluation of memorization patterns across multiple exposure levels without requiring separate training runs.

Pretraining Factors Within the Sparse Gutenberg framework, we systematically investigate several key factors that potentially influence memorization dynamics during pretraining. First, we examine batch size variations (0.5M, 1M, 1.5M and 2M tokens) to assess how optimization granularity affects memorization tendencies. Second, we analyze model scale effects by comparing memorization patterns across architecturally consistent models of 1B, 3B, and 8B parameters. Third, we evaluate how exposure frequency impacts memorization propensity through our FM-Probe methodology.

4 Offset: The Missing Dimension in Memorization Evaluation

Despite increasing attention to memorization risks in LLMs, most literature evaluates memorization using prefixes extracted from the beginning of documents [Hans et al., 2025, Carlini et al., 2023, Kiyomaru et al., 2024]. However, memorized content can appear anywhere within a document—not just at the start. At the same time, little attention has been paid to how the absolute position of the prefix within the model's context window—the offset—affects memorization behavior. In this section, we show that both verbatim memorization and language generation quality are highly sensitive to offset, and whether longer prefixes increase verbatim recall is itself offset-dependent.

4.1 Shorter Prefixes Outperform Longer Ones in Triggering Memorization



(a) LLaMA 1B (b) LLaMA 3B (c) LLaMA 8B

Figure 3: Impact of prefix length on positional fragility across LLaMA models (1B, 3B, 8B), measured by Rouge-L scores on frequency-128 bucket with 500-token suffixes under Sparse Gutenberg setup. With zero offset (blue), smaller models show memorization degradation as prefix length increases, while the 8B model maintains high scores. At non-zero offsets (red, green), all models require longer prefixes to trigger substantial memorization, with the 3B model showing slightly higher memorization than the 8B model at offsets 64 and 128, indicating larger models do not worsen positional fragility.

While previous research suggests that longer prefixes can enhance verbatim recall, we find that this observation is incomplete and potentially misleading. Our results reveal a more efficient condition for triggering memorization: using short prefixes that align with the beginning of the context window during training. As shown in Figure 3, at offset 0, **short prefixes yield the highest recall**, particularly in smaller models, whereas **longer prefixes and longer offsets reduce extractable memorization**. Note that this result does not contradict previous work. As we show in Table 1, with a fixed prefix of 50 tokens, we can induce perfect reproduction of suffixes of up to 8,000 tokens long, especially when the target sequence is frequently encountered during training. However, such an evaluation does not take into account larger prefixes and non-zero offsets.

Table 1: Suffix length and sample counts of **perfect matching** by model size and repetition frequency at offset 0, given a fixed 50-token prefix. Frequencies under 8 showed no verbatim memorization over 50 tokens and are therefore omitted.

Freq.	**LLaMA 1B**		**LLaMA 3B**		**LLaMA 8B**	
	Length	Count	Length	Count	Length	Count
8	-	0	-	0	50	2
16	50	1	500	4	5,000	2
32	3,000	1	7,000	1	8,000	46
64	5,000	1	7,000	3	8,000	131
128	7,000	1	8,000	3	8,000	248

With non-zero offsets, longer prefixes substantially improve memorization, suggesting that they could partially compensate for positive offsets. However, as illustrated by the red and green lines in Figure 3, even with extended prefix length, memorization at disadvantaged positions never fully matches the level observed at offset 0. Moreover, the 3B model achieves higher verbatim recall at offsets 64 and 128 compared to the 1B model, but memorization plateaus from 3B to 8B, indicating that larger models do not mitigate positional fragility and suffer similarly at suboptimal positions.

This inversion of optimal prefix length under offset conditions highlights a key property of verbatim memorization in realistic settings: it is strongly tied to positional cues, especially the beginning of the context window. Our findings align with the *"attention sink"* mechanism discovered by Xiao et al. [2024], where initial tokens disproportionately influence attention distribution and serve as computational anchors. Nasr et al. [2025] has also shown that the token that represents the document boundary, such as end-of-document (`EOD`) marker, can trigger memorization (in our case `BOD`). This raises a key question: is the verbatim memorization driven by the positional role of the initial tokens in general or by the special representational status of the `BOD` token itself? To disentangle these two factors, we design controlled ablation experiments that isolate the `BOD` token's contribution to verbatim recall.

4.2 BOD or Initial Tokens: An Ablation Study

We isolate the impact of the `BOD` token by retraining our 1B model with its attention masked, thereby effectively removing it from the input. In this setting, if memorization persists, it suggests that positional cues alone are sufficient; if it degrades, the `BOD` token likely serves as a dedicated retrieval anchor. To explore whether this change interacts with positional fragility, we evaluated memorization performance at three context-window offsets: 0, 50, and 100 tokens.

Figure 4 illustrates the impact of masking the `BOD` token on the model's memorization capability at the 1B scale. At offset 0 (left panel), where the prefix begins at the start of the context window, masking the `BOD` token yields negligible differences in verbatim recall. This suggests that the initial textual content alone can serve as an effective cue for memorization without an explicit `BOD`.

Interestingly, masking the `BOD` token leads to moderately higher verbatim recall in high-frequency buckets under offset conditions. At a 50-token offset, the `BOD`-masked model achieves up to 8% higher Rouge-L scores for prefixes between 750–2000 tokens—equivalent to roughly 40 additional tokens recalled verbatim. This suggests that, without an explicit `BOD`, the model leans more on local context, making it slightly more robust when recalling sequences from non-initial positions. Nonetheless, it continues to exhibit positional fragility (see Appendix D.1).

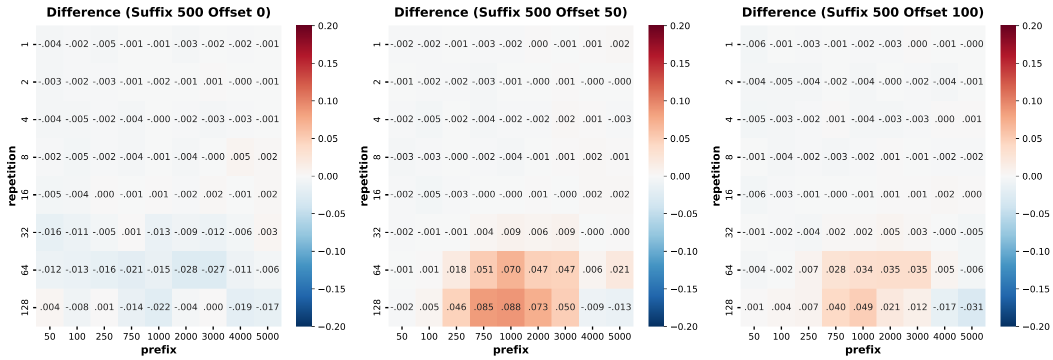


Figure 4: Effect of isolating the BOD token on the memorization capability of the 1B model measured by Rouge-L scores, computed over 500-token suffixes conditioned on varying prefix lengths (50–5000 tokens). Subplots from left to right correspond to context window offsets of 0, 50, and 100.

To ensure that this improvement in recall is not simply a byproduct of degraded generation quality—a symptom of degeneration discussed in §5—we evaluated the model using the MAUVE metric at longer prefix lengths (1000–5000 tokens), corresponding to regions with increased memorization. We also assessed downstream performance (see Table D.1 in the appendix). In both cases, the BOD-masked model performed on par with the baseline, indicating that increased recall does not come at the cost of linguistic ability.

5 Memorization: Under-Explored Cause of Text Degeneration

Our experiments reveal a connection between memorization capabilities and text degeneration in language models. While previous research observed repetitive patterns in training data as the primary cause of degeneration, highlighting a "repetition in, repetition out" phenomenon [Li et al., 2023], our findings suggest that limitations in how models encode and retrieve memorized information represent a more fundamental and previously unexplored mechanism underlying text degeneration. Rather than attributing repetitive outputs solely to repetitive training data, we demonstrate that deficiencies in memorization can also lead to text degeneration.

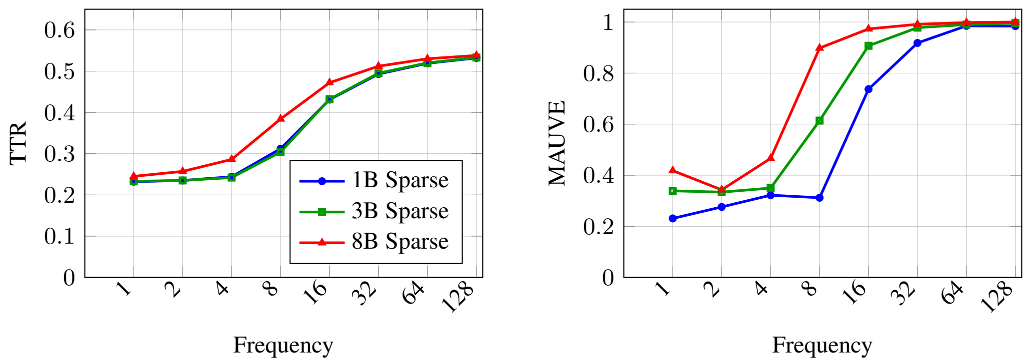
5.1 Context Window Offsets Reduces Output Text Quality

We find that the position of prefixes within the model's context window during pretraining influences the quality of the generated suffixes. As shown in Fig. 1, in the dense Gutenberg setting, where the model is trained on 10K sequences for 80 epochs, we observe that verbatim memorization effectively vanishes, while language modelling capabilities remain stable at larger offsets. In contrast, under the sparse Gutenberg setting with FM-probe bucketing, the model's output **degrades** in both lexical diversity and language coherence. An example is shown in Fig. 10.

This complements the "repetition in, repetition out" hypothesis by Li et al. [2023]. Their work specifically revealed a strong correlation between the repetition rate of 2-gram in training data and the repetitive generation in the model's output. Nevertheless, our results suggest that text degeneration may not be solely driven by repetition in training data.

5.2 Lower Exposure Frequency Reduces Output Text Quality

In addition to context window offsets, the training frequency of a sequence during training independently influences lexical diversity and language coherence. As illustrated in Fig. 5, sequences seen less frequently exhibit pronounced degradation in the model's linguistic capabilities across all model scales (Examples are shown in 11). However, larger models demonstrate greater resilience, achieving higher TTR and MAUVE scores than smaller counterparts, particularly at low frequencies, indicating that model scale modestly enhances linguistic capability.



(a) TTR scores by frequency

(b) MAUVE scores by frequency

Figure 5: Impact of exposure frequency on text quality metrics for Llama models of different sizes, measured with two complementary metrics: (a) TTR scores quantify lexical diversity, showing improvement as training frequency increases; the ground truth TTR lies in the range 0.535–0.541;(b) MAUVE scores capture overall language quality and coherence, reaching near-perfect scores at high frequencies. All experiments used 500-token prefixes and were evaluated on 500-token suffixes with a zero offset.

5.3 Insufficient Memory Retrieval: A contributing Factor to Text Degeneration

Our empirical findings from §5.1 and §5.2 converge on a nuanced understanding of text degeneration in language models. Specifically, they reveal that imbalances in training data frequency can lead to distinct generation issues. High-frequency exposure often results in verbatim memorization, where the model reproduces training data verbatim. Conversely, low-frequency exposure does not necessarily enhance diversity; instead, it can exacerbate degeneration, manifesting as repetitions or thematic loops. This suggests a link between limited memory retrieval capabilities and degenerative text patterns.

The analysis of context window offsets further supports this relationship. As models attempt to continue text from distant positions, they tend to fall into thematic looping. This position-dependent degradation in linguistic ability supports that insufficient memory retrieval can cause text degeneration, particularly when memorisation targets must compete with diverse training data. This connection between retrieval limitations and text degeneration warrants further investigation in future work.

6 Probing the Robustness of Positional Fragility

Our primary analysis revealed positional fragility using greedy decoding on models trained from scratch. To ensure that this phenomenon reflects an inherent model behavior rather than an experimental artifact, we further examine its robustness across decoding strategies and training paradigms.

Robustness Across Decoding Strategies To evaluate the robustness of positional fragility under alternative decoding configurations, we re-ran inference on the repetition-128 FM-Probe bucket using our 8B model under the *Sparse Gutenberg* setting with two additional decoding methods: (1) beam search (k=5), reporting the top-ranked hypothesis, and (2) nucleus sampling (top-p=0.9). Table 2 shows that beam search, a deterministic probability-maximizing strategy, achieves nearly perfect memorization at shallow offsets (e.g., ROUGE-L = 0.999 at offset 0). However, when recall fails at deeper offsets, it exhibits the most severe degeneration, supporting our claim that failed retrieval leads to incoherent outputs. In contrast, nucleus sampling yields lower memorization at shallow offsets but shows a much milder decline in lexical diversity at large offsets (e.g., TTR = 0.460 at offset 2048), suggesting that decoding strategy influences the severity of degeneration while leaving the offset-dependent pattern of positional fragility unchanged.

Persistence Under Continued Pretraining We further examine whether positional fragility persists under a continued pretraining paradigm. Starting from the official *Llama 3.2 1B* model pretrained on 9T tokens, We continued pretraining using otherwise identical settings to the *Sparse Gutenberg* configuration, but with a smaller learning rate 1×10^{-5}. As shown in he continued-pretrained 1B model exhibits slightly stronger memorization (higher ROUGE-L) and more diverse generations (higher TTR) than the from-scratch 1B model, while preserving the same offset-dependent decay pattern, confirming that positional fragility holds across training paradigms.

Table 2: **Robustness of positional fragility across decoding methods and training paradigms.** Top: Decoding strategies on the 8B model. Bottom: 1B continued pretrained model with greedy decoding. All evaluations conducted on the repetition-128 FM-Probe bucket using 500-token prefixes to generate 500-token suffixes. Across all settings, memorization (ROUGE-L) consistently declines with increasing offset, confirming that positional fragility persists regardless of decoding method or training paradigm.

Decoding Strategies (8B Model)	**ROUGE-L ↓**			**TTR ↑**		
Offset	Greedy	Beam	Nucleus	Greedy	Beam	Nucleus
0	0.965	0.999	0.877	0.538	0.541	0.536
8	0.864	0.951	0.701	0.519	0.524	0.523
32	0.652	0.800	0.495	0.474	0.469	0.503
128	0.330	0.507	0.247	0.391	0.361	0.477
512	0.208	0.244	0.186	0.305	0.237	0.461
2048	0.192	0.186	0.180	0.279	0.196	0.460

Training Paradigms (1B Model)	**ROUGE-L ↓**		**TTR ↑**	
Offset	Scratch	Continue	Scratch	Continue
0	0.744	0.669	0.521	0.521
8	0.463	0.569	0.491	0.498
32	0.346	0.426	0.465	0.482
128	0.219	0.249	0.411	0.437
512	0.191	0.202	0.345	0.364
2048	0.185	0.188	0.293	0.340

7 Leveraging Positional Fragility to Mitigate Verbatim Memorization

To demonstrate that positional fragility can be leveraged to mitigate verbatim memorization, we designed a Proof-of-concept (POC) experiment which we refer to as **Swapped Gutenberg**. For each sequence in our FM-Probe buckets, we replace the first 4,000 tokens (the *swapped part*) with tokens from randomly selected Project Gutenberg excerpts within our 10,000-sequence corpus. Thus, the initial segment becomes diverse and contextually unrelated, while the remainder (the *retained part*) maintains controlled repetition frequencies (1 to 256 repetitions). Each sequence includes only one initial `BOS` token, with no additional special tokens between swapped and retained segments (see Fig. 2). We train 1B and 8B models and evaluate verbatim memorization using prefixes from the retained part.

Table 3 shows that the principle demonstrated in this POC works surprisingly well regardless of model scale, keeping verbatim recall consistently low even at the highest exposure frequency. ROUGE-L and LCCS scores remain near the baseline similarity between unrelated texts across all frequencies, indicating **little to no extractable memorization**. As repetition increases, perplexity steadily declines, particularly in the 8B model, yet without corresponding gains in verbatim overlap. Both TTR and MAUVE remain stable, with the 8B model producing more coherent and lexically rich outputs. These results confirm that shifting sensitive content away from the context window's beginning is a simple but effective principle for mitagation at all scales.

Our POC mitigation strategy also neutralizes the offset effects entirely. Recall that this effect manifests in two ways: (1) shorter prefixes are most effective at triggering memorization when taken from the beginning of the context window, and (2) even short prefixes lose their efficacy when shifted to later positions, though longer prefixes can partially compensate. However, both effects disappear when the memorization target is displaced 4,000 tokens deep, as in the Swapped Gutenberg setup. Table 9 shows that increasing prefix length no longer improves memorization, and Table 10 reveals consistent recall across offsets from 0 to 2048 tokens.

Table 3: Comparison of text generation metrics under Swapped and Sparse Gutenberg settings. Each row reports performance on 500-token suffixes generated from 500-token prefixes at offset 0. Selected exposure frequencies (1, 8, 64, 128) are shown to highlight transitions in memorization behavior. Rows 1–4 correspond to Swapped Gutenberg, where prefixes are offset relative to the *retained part* of the sequence; rows 5–8 correspond to standard Sparse Gutenberg. Full results for each setting and frequency range are provided in Table 11 in the Appendix.

Freq.	**Rouge-L↓**		**LCCS↓**		**Perplexity**		**TTR↑**		**MAUVE↑**	
	1B	**8B**	**1B**	**8B**	**1B**	**8B**	**1B**	**8B**	**1B**	**8B**
1	0.178	0.176	0.009	0.009	40.793	82.206	0.348	0.467	0.779	0.970
8	0.180	0.179	0.008	0.009	30.621	12.841	0.350	0.467	0.805	0.934
64	0.182	0.184	0.009	0.011	16.356	3.670	0.377	0.469	0.851	0.905
128	0.181	0.181	0.009	0.011	16.017	3.636	0.377	0.468	0.911	0.958
1	0.181	0.185	0.008	0.009	26.036	16.089	0.225	0.245	0.231	0.418
8	0.183	0.191	0.008	0.016	12.698	3.430	0.231	0.384	0.312	0.898
64	0.522	0.888	0.415	0.858	1.125	1.023	0.497	0.530	0.985	0.998
128	0.744	0.965	0.682	0.951	1.051	1.012	0.521	0.538	0.984	1.000

This disappearance results from two key design choices: the swapped segment is randomly sampled across sequences that break fixed positional associations; and the memorization target is embedded far from the model's typical retrieval anchor near the beginning of the context window.

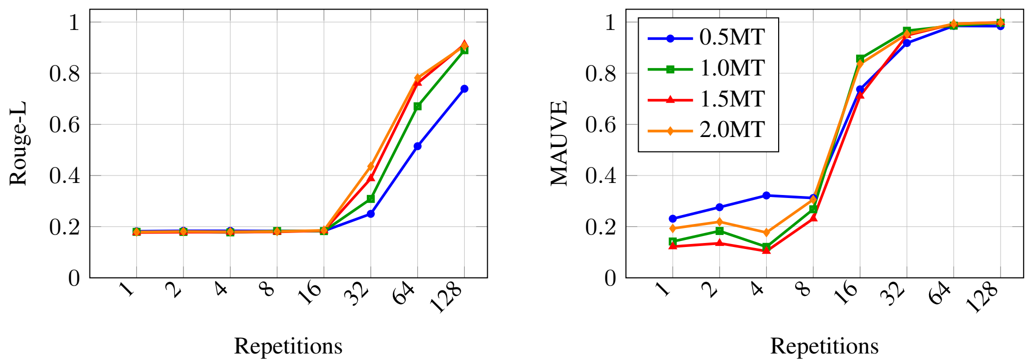
Furthermore, this method not only suppresses memorization but we also find that it improves generation quality. Unlike the sparse Gutenberg setting, where prompting from offset positions leads to lexical degradation and thematic looping, Swapped Gutenberg maintains high coherence and diversity across all conditions. As shown in Table 3, MAUVE and TTR scores remain stable and high, in stark contrast to the quality collapse observed in Sparse Gutenberg setting (also see Figure 1b and 5). These findings highlight a central implication of our study: by decoupling sensitive content from early-context anchors, it is possible to suppress extractable memorization while preserving generation quality and downstream task performance (see Table 12) across model scales and training regimes. This provides a compelling justification for future research into developing more practical, production-ready mitigation techniques based on this principle. Examples are shown in Fig. 12.

8 Batch Size Impact

Our experiments reveal a clear advantage to pretraining language models with smaller batch sizes, which necessitates more iterations given a fixed compute budget. This advantage manifests as reduced verbatim recall (Fig. 6a), improved language coherence (Fig. 6b), and enhanced performance across most downstream benchmarks (Appendix D.2). Notably, these improvements emerge despite maintaining both a constant proportion of Gutenberg sequences in each batch and consistent total training tokens across all experiments. To the best of our knowledge, this has not been shown in prior work.

9 Related Work

Factors Known to Affect Memorization Several factors are known to influence a model's tendency to memorize and reproduce training data verbatim. (1) Model scale: larger models tend to memorize more due to their greater capacity to fit specific examples [Carlini et al., 2021, Lesci et al., 2024]. (2) Repetition frequency: memorization probability increases roughly log-linearly with the number of times a sequence appears during training [Carlini et al., 2023]; we confirm both effects in our setting (Appendix C.1). (3) Complexity: complex sequences seen only once can be latently memorized and retrieved via small parameter perturbations [Duan et al., 2025]. (4) Timing of exposure: sequences seen later in training are more likely to be retained than those encountered earlier [Kiyomaru et al., 2024, Lesci et al., 2024]. (5) Prefix length: while longer prefixes improve recall in prior work [Carlini et al., 2023], we find this effect depends heavily on the prefix's position within the context window (§ 4.1).



(a) Rouge-L scores by repetition frequency

(b) MAUVE scores by repetition frequency

Figure 6: Impact of batch size on verbatim memorization and language coherence for 1B models, as measured by Rouge-L (left) and MAUVE scores (right) across repetition frequencies. Models were trained with batch sizes ranging from 0.5MT to 2.0MT, maintaining constant total training tokens. As batch size increases, we observe enhanced language coherence (MAUVE) at low repetition frequencies, while verbatim memorization (Rouge-L) increases most significantly for sequences seen more than 32 times. The gap between smallest and largest batch sizes reaches up to 0.15 in Rouge-L score at high repetition frequencies, corresponding to approximately 90 fewer tokens being memorized verbatim in the 500-token continuations. This suggests smaller batch sizes with more update steps reduce exact memorization while preserving language quality.

Preventing Verbatim Memorization During Pretraining Pretraining-based strategies have demonstrated promising results in reducing verbatim memorization without sacrificing model performance. One prominent strategy is data deduplication, which removes duplicate or near-duplicate sequences from the training corpus. Lee et al. [2022] showed that deduplication substantially reduces memorization while preserving downstream performance. Kandpal et al. [2022] further demonstrated that it lowers the success rate of extraction attacks. Another promising strategy is the recently proposed Goldfish Loss [Hans et al., 2025], which selectively masks tokens within n-gram windows during training. By consistently omitting these tokens from the loss computation, it prevents the model from learning exact token-to-context mappings, thereby disrupting extractable memorization while retaining language modeling capabilities.

10 Conclusion

We demonstrate that LLMs exhibit a pronounced positional bias in their memorization behavior, with verbatim recall most easily triggered by prefixes near the start of the context window. This offset effect not only distorts standard memorization risk assessments but also offers a new mitigation pathway: simply shifting sensitive content deeper into the context suppresses extractable memorization without degrading generation quality. Our findings further reveal that when retrieval fails—due to offset or insufficient exposure—models often degenerate into repetitive, low-diversity output, linking memorization to generation stability. These insights position offset as a critical and previously underexplored axis in the study of memorization, and suggest simple, scalable mitigation strategies that complement existing techniques. Limitations of our study and directions for future work are discussed in Appendix B.

Acknowledgements

This work was conducted during Yixuan Xu's Master thesis at the ETH AI Center, while he was a Master student at EPFL. Yixuan would like to thank HPC expert Antoni-Joan Solergibert i Llaquet for his technical guidance. This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a06 on Alps.

References

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6481–6491, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.484. URL https://aclanthology.org/2022.findings-emnlp.484/.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle, 2022. URL https://arxiv.org/abs/2207.14255.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Gregory Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Iq0DvhB4Kf.

Jon Brodkin. Openai urges trump: Either settle ai copyright debate or lose ai race to china. https://arstechnica.com/tech-policy/2025/03/openai-urges-trump-either-settle-ai-copyright-debate-or-lose-ai-race-to-china/, March 2025. Ars Technica. Accessed: 2025-05-11.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. URL https://aclanthology.org/2023.emnlp-main.453/.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu

Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila R Fiete. Uncovering latent memories in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KSBx6FBZpE.

Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and Maksym Andriushchenko. Exploring memorization and copyright violation in frontier LLMs: A study of the new york times v. openAI 2023 lawsuit. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=C66DBl9At8.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl

Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Abhimanyu Hans, John Kirchenbauer, Yuxin Wen, Neel Jain, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. Be like a Goldfish, Don't Memorize! Mitigating Memorization in Generative LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 24022–24045, 2025. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/2ad2dffba5079687651226ac8752df97-Abstract-Conference.html`.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.148. URL `https://aclanthology.org/2022.findings-emnlp.148/`.

Jing Huang, Diyi Yang, and Christopher Potts. Demystifying verbatim memorization in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on*

Empirical Methods in Natural Language Processing, pages 10711–10732, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.598. URL `https://aclanthology.org/2024.emnlp-main.598/`.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß, editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.inlg-main.3. URL `https://aclanthology.org/2023.inlg-main.3/`.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, August 2005. ISSN 0001-4966. doi: 10.1121/1.2016299.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/kandpal22a.html`.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright Violations and Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458.

Kimmo Kettunen. Can type-token ratio be used to show morphological complexity of languages? *J. Quant. Linguist.*, 21(3):223–245, July 2014.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. A comprehensive analysis of memorization in large language models. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan, September 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.inlg-main.45/`.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL `https://aclanthology.org/2022.acl-long.577/`.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of memorisation profiles. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15635, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.834. URL `https://aclanthology.org/2024.acl-long.834/`.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. Repetition in repetition out: towards understanding neural text degeneration from the data perspective. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks, 2018. URL `https://arxiv.org/abs/1804.07612`.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=vjel3nWP2a`.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=5G7ve8E1Lu`.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=n6SCkn2QaG`.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92, 2023. URL `http://jmlr.org/papers/v24/23-0023.html`.

David Rosenthal and Livio Veraldi. Training AI language models with third-party content and data from a legal perspective. *Jusletter IT*, (25-Maerz-2025), 2025.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL `https://arxiv.org/abs/1909.08053`.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *ICLR*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract accurately reflects the main contributions, while the Introduction includes additional findings of practical relevance to the study of verbatim memorization.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes we mentioned the limitations of our work in Appendix B.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We use larger batch sizes for efficiency, which may introduce minor numerical differences, but these do not affect the conclusions. For fully deterministic outputs, one can run our scripts with a fixed seed and a batch size of 1 on the same GPU type (e.g., GH200).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our pretrained checkpoint on HuggingFace, and include inference code, evaluation scripts, and FM-Probe buckets after the rebuttal.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We present our experiment settings in § 3 and Appendix A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Observed effects exhibit consistently large effect sizes across over 10 distinct pretraining runs (including 1B/3B/8B sparse, 1B dense, 1B/8B swap, batch size variations, and BOD ablation) and numerous subsequent evaluations. Formal statistical significance testing for all comparisons was therefore computationally prohibitive given this extensive experimental matrix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, Appendix A.1 details information for the computer resources for each experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses negative societal impacts like privacy breaches and copyright violations stemming from LLM memorization throughout the introduction. The work aims for positive impact by proposing mitigation strategies for these risks.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite all datasets and code packages used in the paper. The original sources and papers are referenced in the main text, and the corresponding licenses and terms of use are explicitly detailed in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were only used to assist with language editing and improving clarity.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Implementation Details

A.1 Computing Infrastructure

All experiments were conducted on the Alps supercomputer at the Swiss National Supercomputing Centre (CSCS).[2] Each compute node includes four NVIDIA GH200 Grace Hopper Superchips, with each GH200 integrating a 72-core ARM-based Grace CPU and a Hopper H100 GPU, interconnected via NVLink-C2C. Each GPU provides 96 GB of CUDA memory.

Pretraining For most of our pretrain runs, we utilized a 15-node configuration. Each GPU processed a maximum micro-batch size of 3 sequences, each consisting of 8,192 tokens. Only one specific experiment, involving the 1B model trained with a global batch size of 240 sequences, required a larger setup of 30 nodes. Pretraining of the 1B, 3B, and 8B LLaMA models was orchestrated exclusively using Data Parallelism (DP). Given that the 8B model parameters comfortably fit into the memory of a single GH200 GPU, neither tensor parallelism nor pipeline parallelism was necessary for efficient training. The throughput and compute for each model scale are summarized in Table 4.

Table 4: Aggregate pretraining throughput (tokens per second across all 60 GPUs) and total GPU hours required for pretraining each model scale.

Model Scale	Throughput (tokens/s)	GPU Hours
1B	2,308,935	$\sim$600
3B	936,003	$\sim$1,440
8B	450,017	$\sim$3,060

Inference & Evaluation Inference Time: On a single node, processing one FM-probe bucket (500-token prefix and 500-token suffix) takes $\sim$1.5 minutes for the 1B model and $\sim$3 minutes for the 8B model, including both text generation and evaluation (Perplexity, LCCS, EM, ROUGE-L, TTR). MAUVE is computed separately and takes $\sim$34 minutes per bucket of the same size on a single GH200.

A.2 Model

We pretrain our models using an adapted version of NVIDIA Megatron-LM [Shoeybi et al., 2020], incorporating modifications described in the main paper. Table 5 details the architectural specifications and training hyperparameters of our LLaMA-based models. The tokenizer we utilised is the `OpenMath2-Llama3.1-8B`[3].

A.3 Reproducibility

To enhance reproducibility, we fix the global random seed to 42 across all experiments. However, this alone does not ensure strict determinism in deep learning workflows. In particular, using different batch sizes may trigger different optimized CUDA kernels, leading to minor numerical variations and non-deterministic outputs despite identical seeds and initialization [4]. For guaranteed reproducibility, running inference and downstream evaluation with a batch size of 1 on the same GPU type (NVIDIA GH200) using our provided scripts yields identical results. Due to time constraints and the scale of our experiments, we instead use a batch size of 20 for model generation to maximize throughput, and 4 for downstream evaluation.

[2] `https://www.cscs.ch/computers/alps/`

[3] `https://huggingface.co/nvidia/OpenMath2-Llama3.1-8B`

[4] `https://pytorch.org/docs/stable/notes/numerical_accuracy.html`

Table 5: Architectural specifications and training hyperparameters of the LLaMA-based models. Hyper-parameters are the default value from Megatron-LM.

Parameter	**1B**	**3B**	**8B**
Shared Input-Output Projections	Yes	Yes	No
Hidden Size	2,048	3,072	4,096
Intermediate Size	8,192	8,192	14,336
Number of Layers	16	28	32
Number of Attention Heads	32	24	32
Number of Key-Value Heads	8	8	8
Head Dimension	64	128	128
RoPE scaling factor	32.0	32.0	8.0
RoPE base frequency		500,000	
Max Positional Embeddings		131,072	
Vocabulary Size		128,256	
Warmup Steps	2,000	2,000	200
Initial Learning Rate	3×10^{-4}	3×10^{-4}	2.2×10^{-4}
Weight Decay	0.01	0.01	0.1
Min. Learning Rate	3×10^{-5}	3×10^{-5}	2.2×10^{-5}
Batch Size		60	
Optimizer		Adam	

B Limitations and Future Directions

B.1 Limitations

While our proof-of-concept mitigation validates the principle that positional displacement can suppress extractable memorization, it assumes prior knowledge of which training segments are sensitive and thus does not establish a directly actionable workflow for large-scale pretraining pipelines. Future work may explore how to operationalize this principle—e.g., by combining positional fragility with automated sensitivity detection or data-curation strategies—to mitigate memorization at scale in realistic training environments.

Moreover, our corpus, while comprising approximately 83B tokens, remains several orders of magnitude smaller than real-world foundation model datasets trained on trillions of tokens. As such, our findings may not fully capture the memorization dynamics present at production scale.

Finally, our study lacks formal privacy or generalization guarantees (e.g., differential privacy [Anil et al., 2022]) and is limited to left-to-right autoregressive models trained from scratch. The empirical scope does not extend to fill-in-the-middle or retrieval-augmented architectures, which may exhibit distinct positional behaviors.

B.2 Future Directions

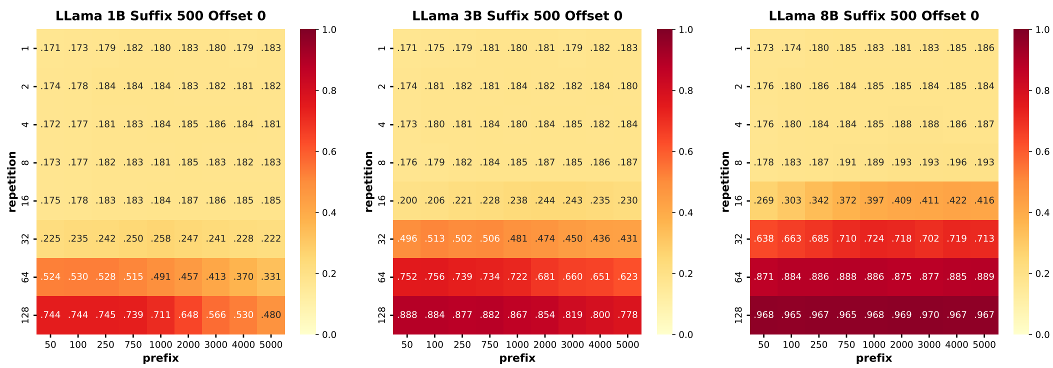
Our analysis is limited to conventional *left-to-right autoregressive models* following the LLaMA architecture. We do not study models trained with *Fill-in-the-Middle (FIM)* objectives, such as OpenAI Codex [Bavarian et al., 2022] and DeepSeek-V3 [DeepSeek-AI et al., 2025], which reconstruct missing spans from both preceding and following context. This bidirectional formulation enables flexible anchoring and contextual recombination, and is conceptually similar to our mitigation strategy of displacing sensitive content deeper in the context window. We leave the study of *positional fragility* and *memorization* in FIM-trained models to future work.

More broadly, it remains an open question whether similar offset-sensitive memorization behaviors emerge in non-text modalities such as code, images, or audio, and whether comparable mitigation strategies are effective. Extending offset-aware analysis to these domains may shed light on the generality of positional fragility and its broader implications for memorization and generation quality.

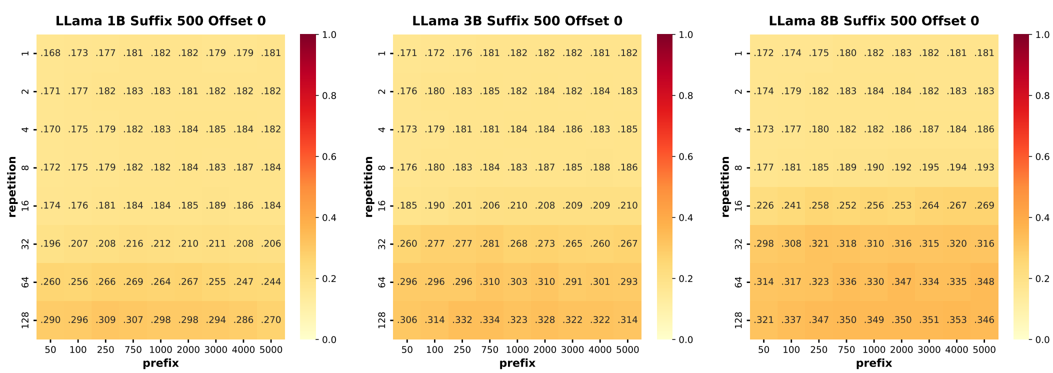
C Reproducing and Extending Prior Work

C.1 Exposure Frequency & Model Size: Interconnected Memorization Dynamics

As shown in Figure 7a, our analysis reveals a significant inverse relationship between model scale and memorization threshold. Following the vertical axis (repetition frequency), the 1B model requires approximately 64 exposures to demonstrate substantial verbatim recall of suffix tokens. This threshold decreases to approximately 32 exposures for the 3B model, and further reduces to only 16 exposures for the 8B model. This systematic pattern suggests that memorization efficiency scales inversely with model size. Extrapolating this trend, a 70B model, without specific memorization mitigation strategies, might memorize sequences after just 2-3 exposures.



(a) Standard Loss



(b) Goldfish Loss

Figure 7: Comparison of verbatim memorization across language models (1B, 3B, and 8B) with varying prefix lengths under Sparse Gutenberg Scenario. The heatmaps display Rouge-L scores computed for 500-token suffixes at offset 0, across different prefix lengths (50–5000 tokens, x-axis) and repetition frequencies (1–128, y-axis).. By comparing Figures 7a and 7b, we observe that Goldfish Loss effectively reduces verbatim recall. Nevertheless, despite employing Goldfish Loss, larger models or higher exposure frequencies still exhibit trends toward memorizing training data.

Complementing this frequency effect, the horizontal axis (prefix length) reveals how different-sized models leverage contextual information. For smaller models (1B and 3B), shorter prefixes of around 50 tokens triggered the strongest verbatim memorization, supporting the hypothesis that initial tokens disproportionately influence memorization patterns. In contrast, the 8B model maintained consistent memorization even with longer prefixes, demonstrating that larger models develop more sophisticated contextual memory mechanisms that operate effectively across varied input lengths.

These dual observations regarding decreased repetition requirements and enhanced contextual processing in larger models together demonstrate that scaling fundamentally transforms memorization capabilities. This explains why larger language models demonstrate both improved generalization and increased memorization risk: their enhanced capacity enables them to memorize more efficiently while simultaneously developing more nuanced representations of linguistic patterns. Our results complement to [Carlini et al., 2023].

C.2 Goldfish Loss

We reproduce the extreme and standard memorization scenarios originally proposed by Hans et al. [2025], adapting them to our dense and sparse Gutenberg experiments, respectively. However, our implementation substantially scales up the original design by pretraining from scratch with sequences that are 4× longer and drawn from 100× more documents. We use the dense Gutenberg setting to tune the optimal Goldfish hyperparameters for the sparse Gutenberg, yielding a dropout ratio of 2% ($k = 50$) and a context width (h) of 50 tokens.

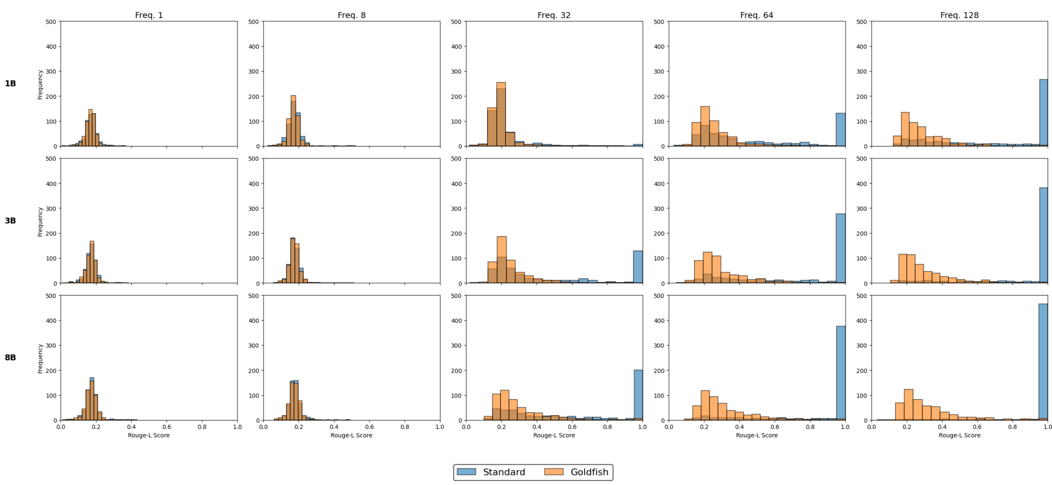


Figure 8: Comparison of Rouge-L distributions between standard and Goldfish Loss LLaMA models of sizes 1B, 3B, and 8B, evaluated with a 50-token prefix and a 500-token suffix at offset 0. The histograms display Rouge-L overlap scores across varying training exposures (1, 8, 32, 64, and 128).

The sparse Gutenberg results confirm that Goldfish Loss effectively mitigates verbatim memorization across all model sizes. As shown in Figure 7b and Figure 8, models trained without memorization mitigation and evaluated with zero positional offset exhibit pronounced verbatim recall abilities that systematically increase with both exposure frequency and model scale. The standard-trained 8B model achieves Rouge-L scores above 0.95 for most test sequences after 128 exposures, while Goldfish Loss models maintain scores consistently below 0.4 even under the most favorable memorization conditions.

Table 6: Downstream Task Performances: Despite training with 2% tokens being dropped, model trained with goldfish loss still shows comparable even superior downstream performance then model trained with standard cross-entropy loss.

Model	Wiki.	Hella.		ARC-c		ARC-e		PIQA	Wino.	CSQA	MMLU
	ppl↓	acc↑	norm↑	acc↑	norm↑	acc↑	norm↑	acc↑	acc↑	acc↑	acc↑
Standard 1B	18.71	40.43	52.31	**33.36**	**35.15**	**68.10**	63.13	71.00	53.91	**21.79**	23.65
Goldfish 1B	18.96	**40.44**	**52.41**	32.08	32.25	67.68	**63.38**	**71.11**	**53.43**	19.00	**25.10**
Standard 3B	15.42	**46.13**	**59.93**	**38.40**	**40.44**	**73.65**	**68.01**	**73.99**	57.06	**21.87**	**25.69**
Goldfish 3B	**15.01**	46.01	59.89	36.52	40.10	71.84	67.76	73.72	**58.41**	20.72	25.42
Standard 8B	13.15	49.74	65.74	42.24	45.99	75.97	72.18	75.52	61.88	**20.56**	24.53
Goldfish 8B	**12.44**	**50.29**	**66.61**	**43.00**	**46.67**	**76.89**	**73.78**	**75.63**	**62.43**	20.39	**26.98**

Beyond verbatim memorization examination, the sparse Gutenberg framework reveals that Goldfish Loss effectively prevents verbatim memorization without compromising downstream performance. In fact, models trained with Goldfish Loss maintain competitive performance across all benchmarks, with this advantage becoming more pronounced as model scale increases. The 8B Goldfish model demonstrates this trend most clearly, outperforming its standard counterpart on several tasks , as shown in Table 6. This exciting result demonstrating that pretraining-based memorization mitigation can simultaneously address copyright concerns and enhance model capabilities.

D Additional Results

D.1 BOD or Initial Tokens Ablation

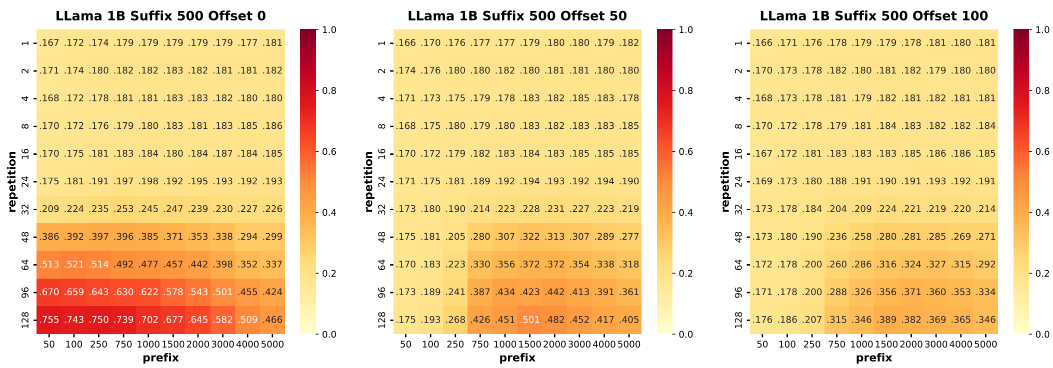


Figure 9: Verbatim memorization of the 1B model trained with `BOD` attention masked, under the Sparse Gutenberg setup. Heatmaps show Rouge-L scores for 500-token suffixes across prefix lengths (50–5000 tokens, x-axis) and repetition frequencies (1–128, y-axis), evaluated at offsets 0, 50, and 100 (left to right). Memorization degrades consistently as the offset increases, highlighting persistent positional fragility despite masking the `BOD` token.

Table 7: Downstream task performance of the baseline 1B model compared to its variant trained with `BOD` attention masked (`-BOD`).

Model	Wiki.	Hella.		ARC-c		ARC-e		PIQA	Wino.	CSQA	MMLU
	ppl↓	acc↑	norm↑	acc↑	norm↑	acc↑	norm↑	acc↑	acc↑	acc↑	acc↑
Baseline	18.71	40.43	**52.31**	**33.36**	**35.15**	**68.10**	63.13	71.00	53.91	**21.79**	23.65
- BOD	**17.89**	**40.98**	52.10	31.40	33.96	68.06	**64.65**	**71.16**	**54.54**	20.88	**24.06**

D.2 Batch Size Impact on Downstream Performance

Larger batch sizes correspond to higher perplexity scores on validation data, with the 240 GBS (global batch size) model showing a perplexity of 22.86 compared to 18.71 for the 60 GBS model — a 22.2% degradation in predictive performance despite seeing identical training content. These same models demonstrate poorer performance as the batch size increases. The 60 GBS model consistently outperforms larger batch variants on reasoning-intensive benchmarks; for example, it achieves 33.36% accuracy on ARC-c compared to just 29.52% for the 240 GBS model, where the baseline random guess is 25%.

Our findings resonate with established patterns documented in deep learning research regarding batch size impact. When models train with smaller batches, the resulting gradient calculations introduce natural variability between updates. This inherent noise effectively serves as an implicit regularization mechanism, as each weight adjustment follows a slightly different trajectory, preventing overspecialization to training examples. This aligns with [Masters and Luschi, 2018] that reduced batch sizes promote more robust generalization through stochastic optimization pathways, ultimately yielding superior performance on evaluation tasks.

Table 8: Downstream task performance of models trained with varying batch sizes. Gradient updates indicate the number of optimization steps taken during pretraining.

Model	Grad. Updates	Wiki. ppl↓	Hella. acc↑	Hella. norm↑	ARC-c acc↑	ARC-c norm↑	ARC-e acc↑	ARC-e norm↑	PIQA acc↑	Wino. acc↑	CSQA acc↑	MMLU acc↑
60 GBS	170,005	**18.71**	**40.43**	**52.31**	**33.36**	**35.15**	**68.10**	**63.13**	**71.00**	53.91	**21.79**	23.65
120 GBS	85,002	18.75	39.85	51.02	30.97	33.36	67.80	61.62	70.13	**55.88**	19.74	23.18
180 GBS	56,668	20.74	39.62	49.83	32.25	34.22	67.76	62.54	69.86	53.20	19.98	**24.78**
240 GBS	42,501	22.86	39.09	49.40	29.52	31.66	67.13	60.94	70.02	52.80	19.49	23.73

Viewed from another angle, our work shows that fewer parameter updates yield reduced model generalization and increased verbatim memorization. This finding aligns with [Hoffer et al., 2017], which attributes large batch performance deficits primarily to reduced update frequency rather than batch size itself. The considerably fewer parameter updates with larger batches (42,501 at 240 GBS compared to 170,005 at 60 GBS) may accelerate verbatim memorization while simultaneously compromising broader generalization capabilities.

D.3 Swapped Gutenberg

Table 9: Impact of prefix length on text generation metrics for 1B and 8B models, evaluated on 500-token suffixes after 256 exposures at offset 0 under Sparse Gutenberg setting. The perplexity remains relatively stable across different prefix lengths, indicating that increasing prefix length does not substantially compensate for our method's effectiveness. For reference suffixes, the TTR ranges between 0.535 and 0.541.

Prefix	Rouge-L↓ 1B	Rouge-L↓ 8B	LCCS↓ 1B	LCCS↓ 8B	Perplexity↓ 1B	Perplexity↓ 8B	TTR↑ 1B	TTR↑ 8B	MAUVE↑ 1B	MAUVE↑ 8B
50	0.172	0.174	0.007	0.009	15.085	3.620	0.356	0.451	0.809	0.914
500	0.181	0.183	0.009	0.011	14.880	3.530	0.380	0.470	0.907	0.948
1000	0.181	0.183	0.009	0.011	15.682	3.464	0.386	0.472	0.914	0.985
2000	0.181	0.183	0.009	0.012	15.613	3.678	0.392	0.480	0.882	0.947
3000	0.184	0.188	0.010	0.014	15.625	3.844	0.398	0.476	0.934	0.945

Table 10: Impact of offset position on text generation metrics for 1B and 8B models after 256 exposures, evaluated on 500-token suffixes generated from corresponding 500-token prefixes. For reference suffixes, the TTR ranges between 0.537 and 0.541. Rouge-L and LCCS scores remain flat across all offsets and hover around the unrelated-text baseline, confirming that our mitigation neutralizes the offset effect. Generation quality (TTR, MAUVE) remains stable, indicating no degradation in fluency or coherence.

Offset	**Rouge-L↓**		**LCCS↓**		**Perplexity↓**		**TTR↑**		**MAUVE↑**	
	1B	**8B**	**1B**	**8B**	**1B**	**8B**	**1B**	**8B**	**1B**	**8B**
0	0.181	0.183	0.009	0.011	14.880	3.530	0.380	0.470	0.907	0.970
1	0.182	0.182	0.009	0.011	14.862	3.525	0.384	0.471	0.862	0.959
2	0.182	0.183	0.009	0.011	14.866	3.518	0.384	0.469	0.939	0.947
4	0.181	0.183	0.009	0.011	14.847	3.526	0.379	0.466	0.833	0.959
8	0.181	0.182	0.009	0.011	14.846	3.512	0.381	0.470	0.900	0.968
16	0.182	0.184	0.009	0.012	14.773	3.534	0.379	0.469	0.876	0.963
32	0.182	0.183	0.009	0.011	14.775	3.573	0.373	0.468	0.876	0.951
64	0.180	0.183	0.009	0.011	14.843	3.528	0.373	0.464	0.899	0.971
128	0.181	0.183	0.009	0.011	15.047	3.621	0.378	0.468	0.900	0.975
256	0.181	0.183	0.009	0.011	15.385	3.763	0.375	0.466	0.870	0.948
512	0.181	0.182	0.009	0.011	15.806	3.917	0.368	0.466	0.934	0.937
1024	0.180	0.182	0.008	0.010	16.247	4.278	0.372	0.467	0.874	0.944
2048	0.181	0.183	0.009	0.011	16.619	4.595	0.373	0.466	0.934	0.962

Table 11: Impact of exposure frequency on text generation metrics for 1B and 8B models evaluated on 500-token suffixes generated from corresponding 500-token prefixes at offset 0. The upper section shows results under Swapped Gutenberg setting, while the lower section shows results under Sparse Gutenberg setting. For reference suffixes, the TTR ranges between 0.535 and 0.541.

Freq.	**Rouge-L↓**		**LCCS↓**		**Perplexity↓**		**TTR↑**		**MAUVE↑**	
	1B	**8B**	**1B**	**8B**	**1B**	**8B**	**1B**	**8B**	**1B**	**8B**
1	0.178	0.176	0.009	0.009	40.793	82.206	0.348	0.467	0.779	0.970
2	0.179	0.176	0.008	0.008	40.710	63.966	0.349	0.467	0.854	0.969
4	0.178	0.178	0.008	0.009	35.502	34.562	0.350	0.462	0.784	0.910
8	0.180	0.179	0.008	0.009	30.621	12.841	0.350	0.467	0.805	0.934
16	0.181	0.182	0.009	0.010	24.063	5.008	0.355	0.466	0.759	0.965
32	0.181	0.181	0.009	0.011	18.585	3.970	0.374	0.470	0.940	0.958
64	0.182	0.184	0.009	0.011	16.356	3.670	0.377	0.469	0.851	0.905
128	0.181	0.181	0.009	0.011	16.017	3.636	0.377	0.468	0.911	0.958
256	0.181	0.183	0.009	0.011	14.880	3.530	0.380	0.470	0.907	0.948
1	0.181	0.185	0.008	0.009	26.036	16.089	0.225	0.245	0.231	0.418
2	0.182	0.184	0.008	0.009	24.474	13.902	0.227	0.257	0.276	0.343
4	0.184	0.184	0.008	0.009	19.065	9.047	0.228	0.286	0.322	0.466
8	0.183	0.191	0.008	0.016	12.698	3.430	0.231	0.384	0.312	0.898
16	0.183	0.372	0.010	0.232	5.741	1.229	0.323	0.472	0.737	0.973
32	0.250	0.710	0.086	0.637	1.565	1.054	0.443	0.512	0.918	0.991
64	0.522	0.888	0.415	0.858	1.125	1.023	0.497	0.530	0.985	0.998
128	0.744	0.965	0.682	0.951	1.051	1.012	0.521	0.538	0.984	1.000

Table 12: Downstream Task Performances for Swapped Gutenberg experiments.

Model	Wiki.	Hella.		ARC-c		ARC-e		PIQA	Wino.	CSQA	MMLU
	ppl↓	acc↑	norm↑	acc↑	norm↑	acc↑	norm↑	acc↑	acc↑	acc↑	acc↑
Swapped 1B	22.31	41.13	52.22	31.83	34.56	68.73	63.51	71.93	55.01	20.56	23.27
Swapped 8B	15.71	49.41	65.48	42.06	44.11	75.88	71.93	75.79	61.09	20.64	25.89

E License for Existing Assets

Project Gutenberg We use texts from the Project Gutenberg collection (via HuggingFace), which consists of works in the public domain in the United States. In accordance with Project Gutenberg's terms[5], we do not redistribute their format or branding. Our use is strictly limited to non-commercial academic research, in compliance with Swiss copyright law.

Fineweb-Edu We use the Fineweb-Edu corpus released on HuggingFace[6], which is distributed under the Open Data Commons Attribution License (ODC-By) v1.0. Our use complies with this license and adheres to CommonCrawl's Terms of Use. The dataset is used solely for non-commercial academic research, and we do not redistribute the data or any derivatives.

F Demonstrations

[5] `https://www.gutenberg.org/policy/license.html`

[6] `https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu`

Rouge-L Score: 1.0000

Line	True Sequence	Line	Model Generation
1	moved they came as conquerors among other races.	1	moved they came as conquerors among other races.
2		2	
3	In their primeval home and probably before the year B.C. 3000, they had	3	In their primeval home and probably before the year B.C. 3000, they had
4	already acquired a fair degree of civilization. They built houses,	4	already acquired a fair degree of civilization. They built houses,
5	ploughed the land, and ground grain into flour for their baking. The	5	ploughed the land, and ground grain into flour for their baking. The
6	family relations were established among them; they had some social	6	family relations were established among them; they had some social
7	organization and simple form of government; they had learned to worship	7	organization and simple form of government; they had learned to worship
8	a god, and to see in him a counterpart of their tribal ruler.	8	a god, and to see in him a counterpart of their tribal ruler.
9		9	
10	From their upland farms they must have looked eastward upon yet higher	10	From their upland farms they must have looked eastward upon yet higher
11	mountains, rising impenetrable above the snowline; but to north and	11	mountains, rising impenetrable above the snowline; but to north and
12	south and west they might turn to lower regions; and by degrees, perhaps	12	south and west they might turn to lower regions; and by degrees, perhaps
13	as they grew too numerous for comfort, a few families wandered off along	13	as they grew too numerous for comfort, a few families wandered off along
14	the more inviting routes. Whichever way they started, their adventurous	14	the more inviting routes. Whichever way they started, their adventurous
15	spirit led them on. We find no trace of a single case where hearts	15	spirit led them on. We find no trace of a single case where hearts
16	failed or strength grew weary and the movement became retrograde, back	16	failed or strength grew weary and the movement became retrograde, back
17	toward the ancient home. Spreading out, radiating in all directions, it	17	toward the ancient home. Spreading out, radiating in all directions, it
18	is they who have explored the earth, who have measured it and marked its	18	is they who have explored the earth, who have measured it and marked its
19	bounds and penetrated almost to its every corner. It is they who still	19	bounds and penetrated almost to its every corner. It is they who still
20	pant to complete the work so long ago begun.	20	pant to complete the work so long ago begun.
21		21	
22	Before B.C. 2000 one of these exuded swarms had penetrated India,	22	Before B.C. 2000 one of these exuded swarms had penetrated India,
23	probably by way of the Indus River. In the course of a thousand years or	23	probably by way of the Indus River. In the course of a thousand years or
24	so, the intruders expanded and fought their way slowly from the Indus to	24	so, the intruders expanded and fought their way slowly from the Indus to
25	the Ganges. The earlier and duskier inhabitants gave way before them or	25	the Ganges. The earlier and duskier inhabitants gave way before them or
26	became incorporated in the stronger race. A mighty Aryan or Hindu empire	26	became incorporated in the stronger race. A mighty Aryan or Hindu empire
27	was formed in India and endured there until well within historic times.	27	was formed in India and endured there until well within historic times.
28		28	
29	Yet its power faded. Life in the hot and languid tropics tends to	29	Yet its power faded. Life in the hot and languid tropics tends to
30	weaken, not invigorate, the sinews of a race. Then, too, a formal	30	weaken, not invigorate, the sinews of a race. Then, too, a formal
31	religion, a system of castes[8] as arbitrary as among the Egyptians,	31	religion, a system of castes[8] as arbitrary as among the Egyptians,
32	laid its paralyzing grip upon the land. About B.C. 600 Buddhism, a new	32	laid its paralyzing grip upon the land. About B.C. 600 Buddhism, a new
33	and beautiful religion, sought to revive the despairing people; but they	33	and beautiful religion, sought to revive the despairing people; but they
34	were beyond its help.[9] Their slothful languor had become too deep.	34	were beyond its help.[9] Their slothful languor had become too deep.
35	From having been perhaps the first and foremost and most civilized of	35	From having been perhaps the first and foremost and most civilized of
36	the Aryan tribes, the Hindus sank to	36	the Aryan tribes, the Hindus sank to

(a) Sequence 0, Offset 0, Frequency 128

Rouge-L Score: 0.1940

Line	True Sequence	Line	Model Generation
1	to north and	1	to north and
2	south and west they might turn to lower regions; and by degrees, perhaps	2	south and west they might turn to lower regions; and so, with the aid of
3	as they grew too numerous for comfort, a few families wandered off along	3	the sun, they might reach new heavens and new earths in their turn. The
4	the more inviting routes. Whichever way they started, their adventurous	4	Chinese, too, after their long journey, have no old traditions to transmit
5	spirit led them on. We find no trace of a single case where hearts	5	to their sons; they have no _faith_, no hopes, no fears, only an
6	failed or strength grew weary and the movement became retrograde, back	6	peasant, or a peasant-farmer, in the heart of the Chinese mountains.
7	toward the ancient home. Spreading out, radiating in all directions, it	7	Chinese, no matter how they may have curiously drawn pictures of their
8	is they who have explored the earth, who have measured it and marked its	8	ancient past, bear in their veins the trace of some degree of intellectual
9	bounds and penetrated almost to its every corner. It is they who still	9	culture. Their homes are Chinese, their customs are Chinese, their
10	pant to complete the work so long ago begun.	10	sorrow is Chinese, and their folk-lore is Chinese. They know how to
11		11	be themselves, and they know how to love for others. In so far, at any
12	Before B.C. 2000 one of these exuded swarms had penetrated India,	12	rate, as the Chinese claim their country, they have certainly made a
13	probably by way of the Indus River. In the course of a thousand years or	13	beginning.
14	so, the intruders expanded and fought their way slowly from the Indus to	14	
15	the Ganges. The earlier and duskier inhabitants gave way before them or	15	
16	became incorporated in the stronger race. A mighty Aryan or Hindu empire	16	THE PERIOD OF TRANSITION
17	was formed in India and endured there until well within historic times.	17	
18		18	From this splendidly successful civilization, the Chinese, like the
19	Yet its power faded. Life in the hot and languid tropics tends to	19	Greeks and Romans, fell back in a period of a hundred years. They
20	weaken, not invigorate, the sinews of a race. Then, too, a formal	20	stretched themselves to the verge of poverty and obscurity, and the
21	religion, a system of castes[8] as arbitrary as among the Egyptians,	21	other European nations, glad to share in their prosperity, came to
22	laid its paralyzing grip upon the land. About B.C. 600 Buddhism, a new	22	their rescue. But, like the Scythians, the Chinese, the Persians, the
23	and beautiful religion, sought to revive the despairing people; but they	23	Macedonians, even the Romans, were not able to hold their own against
24	were beyond its help.[9] Their slothful languor had become too deep.	24	Nature and Time. Their empire was broken up, and the nations that were
25	From having been perhaps the first and foremost and most civilized of	25	the only strong men in the world at the beginning of our era were
26	the Aryan tribes, the Hindus sank to be degenerate members of the race.	26	overthrown. Yet, with the Chinese, the transition was not a hard one;
27	We shall turn to look on them again in a later period; but they will be	27	they showed their willing but their unwilling ignorance. They were
28	seen in no favorable light.	28	naturally strong and vigorous, and self-possessed. They had the gift of
29		29	sensibility, and made the best of their power for the day and night of
30	[Footnote 8: See _The Formation of the Castes_, page 52.]	30	history. They were brave and vigorous--and yet they were always firm
31		31	and dead to what they believed to be right. They had a high ideal of
32	[Footnote 9: See _The Foundation of Buddhism_, page 160.]	32	manhood, and selfhood in itself; they had a noble conception of courage;
33		33	they knew well the value of strength and of its advantages; they knew
34	Meanwhile other wanderers from the Aryan home appear to the north and	34	well the nature of peace, and they found it, as indeed their last and
35	west. Perhaps even the fierce Tartars are an Aryan race, much altered	35	great period of freedom had proved, to be the best of all possible
36	from long dwelling among the yellow peoples. One tribe, the Persians,	36	conditions.
37	moved directly west, and became neighbors of the already noted Semitic	37	
38		38	The Chinese were great manufacturers, and could

(b) Sequence 0, Offset 128, Frequency 128

Rouge-L Score: 0.1840

Line	True Sequence	Line	Model Generation
1	the people quite ignored. We see the heroes	1	the people of the larger cities, with their
2	driving to battle in their chariots, guarded by shield and helmet,	2	generals and their counsels, the _Iliad_ and the _Odyssey_, not at all
3	flourishing sword and spear. We learn what Ulysses did not know of	3	unlike those of the Phoenicians, are more interesting to us than the
4	foreign lands.. We hear Achilles' famed lament amid the dead, and note	4	mere tales of the Trojan war. We are interested in the Greek race as
5	the vague glimmering idea of a future life, which the Greeks had caught	5	a whole. We have come down to the time when the Greek race was the
6	perhaps from the Egyptians, perhaps from the suggestive land of dreams.	6	first conqueror of the world. The Romans, the modern European
7		7	nations, are the direct descendants of the Greek race, and the
8	With the year B.C. 776 we come in contact with a clear marked	8	descendants of the ancient Greeks. The modern nations are the
9	chronology. The Greeks themselves reckoned from that date by means of	9	descendants of the Greeks, and it is in them that we find the real
10	olympiads or intervals between the Olympic games. The story becomes	10	history of the world. We must study them in order to understand the
11	clear. The autocratic little city kings, governing almost as they	11	history of our race. We must study them.
12	pleased, have everywhere been displaced by oligarchies. The few leading	12	
13	nobles may name one of themselves to bear rule, but the real power lies	13	We find only a few scattered remnants of the Greek race in the
14	divided among the class. Then, with the growing prominence of the	14	present-day world. The ancient Greeks themselves were not all Greeks. The
15	Pythian games[15] we come upon a new stage of national development. The	15	Greeks were Greeks, the Etruscans were Greeks, the Spaniards were Greeks,
16	various cities begin to form alliances, to recognize the fact that they	16	the Italians were Greeks, the Spaniards and the French were Greeks. The
17	may be made safer and happier by a larger national life. The sense of	17	Greeks were the first to be Greek, the first to be Roman, the first to
18	brotherhood begins to extend beyond the circle of personal acquaintance.	18	be Christian. The Greeks were the first to be Greek-speaking, the first
19		19	to be Greek-learners. The Greeks were the first to be civilized. The
20	[Footnote 15: See _Pythian Games at Delphi_, page 181.]	20	Greeks were the first to be barbarians. The Greeks were the first
21		21	nations. The Greeks were the first to be the first conquerors. The
22	This period was one of lawmaking, of experimenting. The traditions, the	22	Greeks were the first to be the first conquerors. The Greeks were the
23	simple customs of the old kingly days, were no longer sufficient for the	23	first to have been born as a race. The Greeks were the first to be
24	guidance of the larger cities, the more complicated circles of society,	24	Christian. The Greeks were the first to be Greek-speaking.
25	which were growing up. It was no longer possible for a man who did not	25	
26	like his tribe to abandon it and wander elsewhere with his family and	26	The Greeks were the first to be the first conquerors. The Romans were the
27	herds. The land was too fully peopled for that. The dissatisfied could	27	first to be the first barbarians. The Romans were the first to have
28	only endure and grumble and rebel. One system of law after another was	28	been born as a Greek race. The Romans were the first to have been
29	tried and thrown aside. The class on whom in practice a rule bore most	29	Christian. The Greeks were the first to have been the first conquerors.
30	hard, would refuse longer assent to it. There were uprisings, tumults,	30	The Greeks were the first to have been born as a race. The Greeks were
31	bloody frays.	31	the first to have been born as a race. The Greeks were the first
32		32	conquerors. The Greeks were the first to have been born as a race.
33	Sparta, at this time the most prominent of the Greek cities, evolved a	33	
34	code which made her in some ways the wonder of ancient days. The state	34	The Greeks were the first to have been born as a race. The Romans were
35	was made all-powerful; it took entire possession of the citizen, with	35	the
36	the purpose of making him a fighter, a strong defender of himself and of	36	

(c) Sequence 0, Offset 2048, Frequency 128

Figure 10: Demonstration of verbatim memorization decay and text degeneration as offset increases in the Sparse Gutenberg setting. Each example shows a 500-token suffix generated from a 500-token prefix by 1B model, with the highlighted span indicating the tokens used for ROUGE-L calculation. At offset 0, the model reproduces the training suffix nearly perfectly. At offset 128, memorization is limited to the initial few tokens. By offset 2048, only scattered fragments are recalled, and the output shows thematic looping—e.g., repetitive structures such as “The xxx were the first to...” emerge in place of coherent continuation.

Rouge-L Score: 0.1860

Line	True Sequence	Line	Model Generation
1	guns on their shoulders, and wore high	1	been the men who had been driven from the
2	boots and the high-crowned conventional brigand hats. Ever since we left	2	mountain, and they were the best. The one who was talking to the man
3	Florence we had seen at intervals in the fields and woods a notice with	3	with the hammer was not a man who had seen the country before, but one
4	the words, "_È vietata la bandita_," which we interpreted as a warning	4	who had lived through it. The other man was a little taller, and as tall
5	against the bandits or convicts for whom our Florentine friends had	5	as we were, and he had a face that was as white as a sheet, and his
6	prepared us. And now we seemed to have come face to face with two of	6	hands were as strong as the branches of a tree. He was a man who had
7	these brigands. But it turned out that there was little of the bandit	7	been a soldier, and he had come to the mountains with his family. The
8	about them save their appearance. Their guns were for birds, and later	8	man who was standing there was a young girl, and we were not sure
9	on we learned that the alarming signs were merely to forbid the	9	whether she was a girl or a child, and the voice of the man was a
10	trespassing of these very gentlemen.	10	child, but it was not a child. The man was a man who had been a soldier
11		11	and had been a soldier, and he was now a man who had been a mother. The
12	[Illustration: A SLIGHT OBSTRUCTION.]	12	woman was a woman, and the man was a man, and they were both talking of
13		13	the mountains. The girl was a little afraid, but the man was not. He had
14	A mile or two farther on, the road began to go down again. We were both	14	been a soldier, and he had been a mother, and he had been a father, and
15	glad to be on the machine after our walk. We could see to the bottom of	15	he had been a child. The man had been a man, and the girl was a girl, and
16	the hill, and there was no one in sight. J. let go the brake. None but	16	she was a woman, and the man was a man, and they were both talking of
17	cyclers know the delight of a five-minutes coast after hours of up-hill	17	the mountains. The girl was silent, but the man was very loud, and his
18	toiling. They, however, will sympathize with our pleasure in the	18	hands were very strong. He swung his hammer, and he swung his gun,
19	mountains near Siena. But when it was at its fullest, and the machine	19	and he swung his hand. He threw down his hammer and he threw up his gun
20	was going at the rate of about twenty miles an hour, and neither brake	20	and his gun was ready. He swung his hand and he swung the gun. The
21	nor back-pedalling could bring it to a sudden halt, a man (or the foul	21	hand of the man with the hammer swung the gun, and it was a gun. The
22	fiend himself) drove a flock of sheep out from the woods a few feet in	22	hand of the man with the gun swung the gun, and it was a gun. The man
23	front of us. When we reached them only the first had crossed the road;	23	with the hammer swung the gun, and it was a gun. The man with the hammer
24	of course, all the rest had to follow. They tried to go on right through	24	did not know what to do. He did not know that the hand with the hammer
25	the wheels, but only succeeded in getting under them, setting the	25	could fire a gun. He thought of the mountains, and he said, "I don't
26	machine to pitching like a ship in a heavy sea. But I held on fast; J.	26	know what to do." He had not been in the mountains long enough to know
27	stood on the pedals and screwed the brake down; the little wheel	27	what to do. The woman had been silent for a long time, and then she
28	scattered the sheep like the cow-catcher of an engine, and we brought up	28	said, "I am a man,
29	in the gutter. Before we stopped, J. began a moral lecture to the	29	
30	shepherd, and was showing him how, if the machine had gone over, the	30	
31	consequences would have been worse for us than for his flock. The	31	
32	lecture ended rather _im_morally with _accidente voi_, and _imbecile_,	32	
33	the deadliest of all Italian maledictions	33	

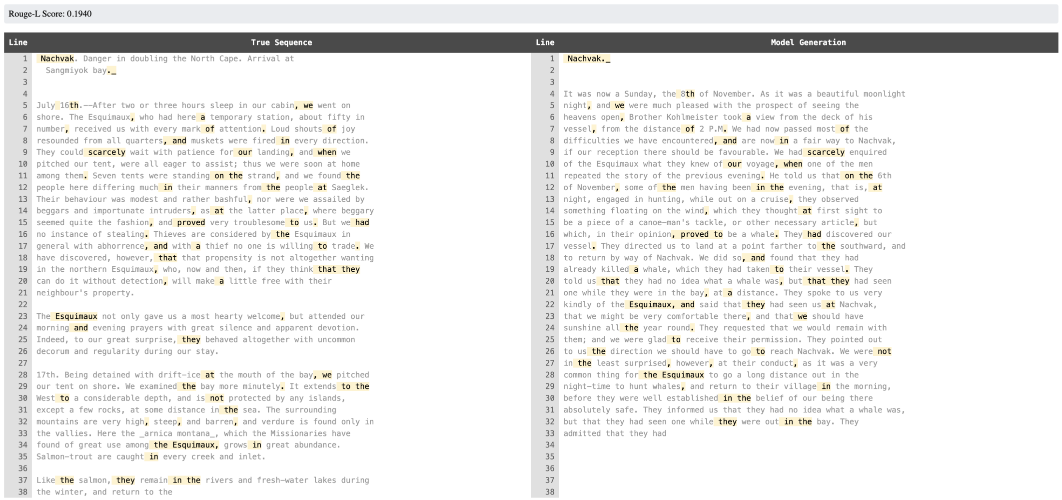
(a) Sequence 289, Offset 0, Frequency 4

Rouge-L Score: 0.2100

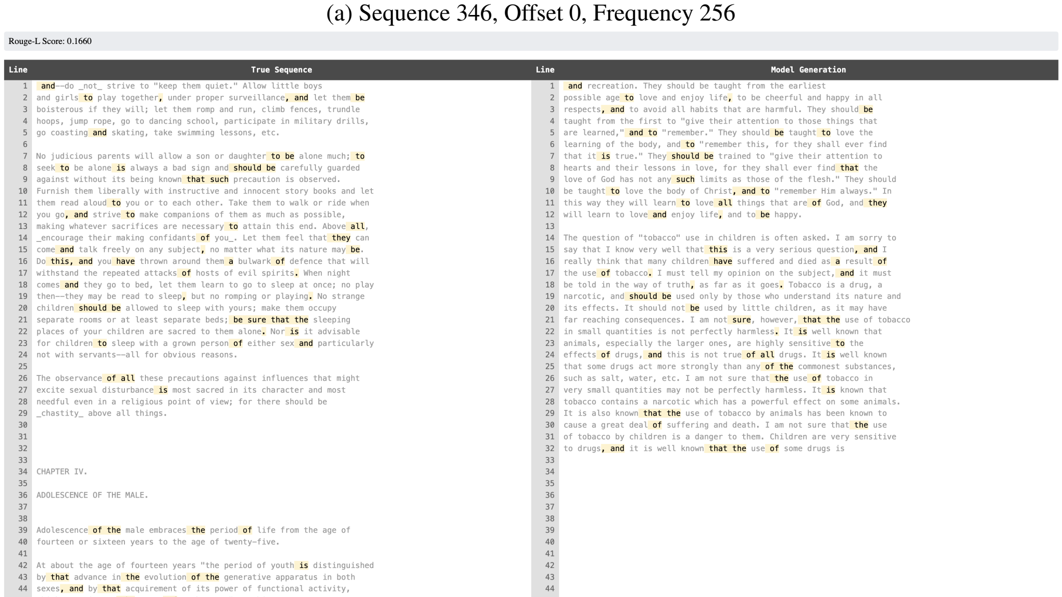
Line	True Sequence	Line	Model Generation
1	battery of 75's stationed near	1	regiment of the German Army
2	Pervyse. His summer home is a little distance out from Liège. His wife,	2	which was in the trenches in France. He has been wounded in the
3	sister-in-law, and his three children were in the house when the Germans	3	thigh, and has been taken to the hospital in Paris. He is wounded and
4	came. Peasants, driven from their village, hid in the cellar. His sister	4	has been brought to the hospital in Paris. He has been wounded and has
5	took one child and hid in a closet. His wife took the two-year-old baby	5	been brought to the hospital in Paris.
6	and the older child and hid in another closet. The troops entered the	6	
7	house, looted it and set it on fire. As they left they fired into the	7	I have been in the habit of visiting the wounded in France, and of
8	cellar. The mother rushed from her hiding place, went to her desk and	8	returning to the same place in the trenches. I have been to the hospital
9	found that her money and the family jewels, one a gift from the	9	in Paris, and to the hospital of the French Army, and to the hospital
10	husband's family and handed down generation by generation, had been	10	in Paris. I have been in the habit of visiting the wounded in France,
11	stolen. With the sister, the baby in arms, the two other children and	11	and of returning to the same place in the trenches. I have been to the
12	the peasants, she ran out of the garden. They were fired on. They hid in	12	hospital in Paris, and to the hospital of the French Army, and to the
13	a wood. Then, for two days, they walked. The raw potatoes which they	13	hospital in Paris, and to the hospital in Paris.
14	gathered by the way were unfit for the little one. Without money, and	14	
15	ill and weakened, they reached Holland. This lady is in a safe place	15	I have been in the habit of visiting the wounded in France, and of
16	now, and her testimony in person is available.	16	returning to the same place in the trenches, and of going to the hospital
17		17	in Paris, and to the hospital of the French Army, and to the hospital
18	[Illustration: THE GREEN PASS, USED ONLY BY SOLDIERS AND OFFICERS OF THE	18	in Paris, and to the hospital of the French Army, and to the hospital
19	BELGIAN ARMY.	19	in Paris, and to the wounded in France, and to the wounded of the
20		20	French Army, and to the wounded of the French Army, and to the wounded
21	It gives passage to the trenches at any hour. The writer, by holding	21	of the French Army.
22	this, and working under the Prime Minister's son, became	22	
23	stretcher-bearer in the Belgian Army.]	23	I have been in the habit of visiting the wounded in France, and of
24		24	returning to the same place in the trenches, and of going to the hospital
25	The apologists of the widespread reign of frightfulness say that war is	25	in Paris, and of going to the hospital of the French Army, and of
26	always "like that," that individual drunken soldiers have always broken	26	returning to the same place in the trenches, and of visiting the wounded
27	loose and committed terrible acts. This defense does not meet the	27	in France, and to the hospital of the French Army, and to the hospital
28	facts. It meets neither the official orders, nor the cold method, nor	28	in France, and to the wounded of the French Army.
29	the immense number of proved murders. The German policy was ordered from	29	
30	the top. It was carried out by officers and men systematically, under	30	I have been to the hospital of the French Army, and to the hospital of
31	discipline. The German War Book, issued by the General Staff, and used	31	the French Army, and to the hospital of the French Army, and to the
32	by officers, cleverly justifies these acts. They are recorded by the	32	hospital of the French Army, and to the wounded of the French Army.
33	German soldiers themselves in their diaries, of which photographic	33	
34	reproductions are obtainable in any large library. The diaries were	34	I have been in the habit of visiting the wounded in France, and of
35	found on the persons of dead and wounded Germans. The name of the man	35	returning to the same place in the trenches, and of going to the hospital
36	and his company are given.	36	in Paris, and of going to the hospital of the French Army, and of
37		37	returning to the same place in the trenches, and of visiting the wounded
38	On Sunday,	38	in France, and to the wounded of the French

(b) Sequence 470, Offset 0, Frequency 4

Figure 11: Demonstration of verbatim memorization decay and text degeneration for low exposure frequency sequences in the Sparse Gutenberg setting. Each example shows a 500-token suffix generated from a 500-token prefix by 1B model, with the highlighted span indicating the tokens used for ROUGE-L calculation. For the generated suffix (right), we show clear evidence of thematic looping and repetitive generations.



(a) Sequence 346, Offset 0, Frequency 256



(b) Sequence 440, Offset 0, Frequency 256

Figure 12: Demonstration of how displacing legally sensitive content deeper into the context window effectively suppresses verbatim memorization while preserving semantic fidelity in the Swapped Gutenberg setting. Each example shows a 500-token suffix generated from a 500-token prefix by 1B model, with the highlighted spans indicating portions used for ROUGE-L calculation. (a) Both the true suffix (left) and the model-generated suffix (right) describe interactions and encounters with the Esquimaux community upon arrival at Nachvak, discussing their behavior, hospitality, and local surroundings. Despite thematic similarity, the generated text diverges notably in specific narrative details—such as dates of arrival (July 16th vs. November 8th) and described observations (presence of drift-ice versus whale sightings)—illustrating effective suppression of verbatim memorization with maintained lexical richness and narrative coherence. (b) The true suffix focuses primarily on parenting advice regarding child behavior, education, and guidelines to prevent inappropriate sexual behavior and maintain chastity, whereas the generated suffix addresses the use of tobacco among children and discusses its potential dangers and harmful effects, highlighting clear thematic divergence and effective memorization suppression.