WHERE DOES LONG-CONTEXT SUPERVISION ACTUALLY GO? EFFECTIVE-CONTEXT EXPOSURE BALANCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context adaptation is commonly treated as window scaling: increase the training sequence length, continue pretraining, and expect long-context behavior to improve. This view ***misses a token-level supervision mismatch***: longer windows provide longer possible contexts, while causal language modeling delivers supervision one target token at a time. In packed training with document-boundary masking, each target token has its own ***effective left context***, the visible same-document prefix at prediction time, and ***the resulting exposure histogram remains heavily concentrated in short and medium regimes***. We introduce **EXACT** (Effective-context Allocation for Context Training), a supervision-allocation objective that measures this exposure distribution, assigns a fixed extra supervision weight to long effective-context targets, and distributes that extra weight by inverse frequency within the long-context tail. Across seven Qwen and LLaMA continued-pretraining (CPT) configurations, **EXACT** improves all twenty-eight trained/extrapolated NoLiMa and RULER comparisons in the main table. On Qwen2.5-0.5B, it improves NoLiMa by +10.09 points in trained lengths and +5.34 in extrapolated lengths, and improves RULER by +10.69 and +5.55. On LLaMA-3.2-3B, it improves RULER by +17.91 and +16.11 in trained and extrapolated regimes. Standard QA and reasoning are preserved, with a +0.24 macro change across MMLU, ARC-C, HellaSwag, WinoGrande, PIQA, and GSM8K. A distance-resolved evidence-ablation probe further shows that the gains arise in the expected regime: when relevant evidence is placed thousands of tokens away from the query, **EXACT** makes the model more likely to prefer the correct answer, while short-distance cases remain close to standard CPT. These results support a supervision-centric thesis: ***long-context adaptation depends not only on what context a model can process, but also on how strongly training supervises predictions made under long effective context.***

1 INTRODUCTION

Long-context language models are usually adapted by increasing the training window (Chen et al., 2023b; Bai et al., 2024a; Fu et al., 2024; Ding et al., 2024). This practice treats sequence length as the unit of long-context supervision. Sequence length is a capacity variable: it sets the maximum span a model may process in a forward pass. Supervision length is a token-level variable: it is determined by how much usable left context each supervised target actually has. A 4K or 8K packed sequence can contain many targets whose same-document left context is short, especially under document-boundary masking and realistic document length mixtures.

The right unit of analysis is therefore the target token. For a token at position i, its effective left context ℓ_i is the number of previous same-document tokens visible when predicting that token. This quantity depends on causal masking, document boundaries, packing layout, and the length distribution of the selected documents. Data selection decides which documents enter the stream and how much long text is available. The training objective decides how supervision is allocated across the tokens produced by that stream. These two layers interact, but they are distinct. Even with a long-window corpus, the training signal can remain concentrated on short and medium effective-context targets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

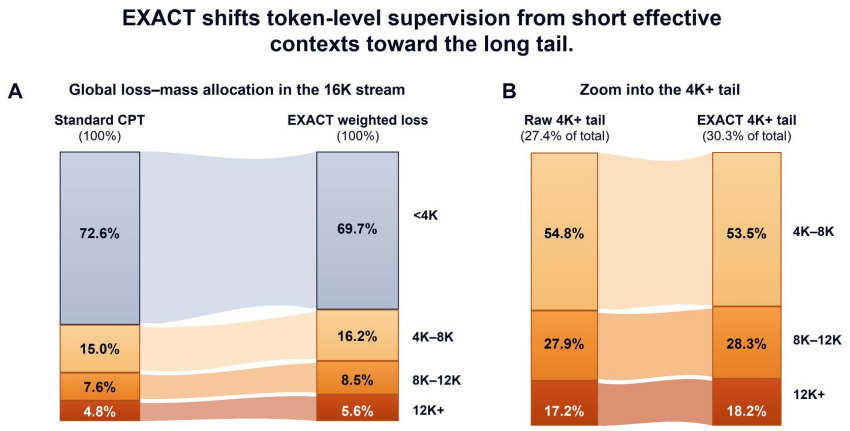

Figure 1: Standard 16K CPT still places most token-level supervision below 4K effective context. **EXACT** keeps the packed stream fixed but changes per-token loss weights; when weighted objective contributions are normalized for visualization, the relative loss mass shifts into the 4K+ tail, and within that tail farther effective-context buckets receive a larger share.

Figure 1 makes the mismatch visible. The packed sequence may span 16K tokens, yet each supervised target sees only the prefix of its own document segment. In the measured 16K stream, 72.6% of standard-CPT loss mass still lands below 4K effective context, leaving only 27.4% in the 4K+ tail. **EXACT** keeps the stream fixed and changes the supervision allocation: the 4K+ tail rises to 30.3%, and the farther buckets receive a larger share within that tail. This is the supervision-allocation bottleneck targeted by this paper: ***long-context adaptation pays for long windows, while a large share of training signal still lands on targets with short or medium effective context***.

We introduce **EXACT**, Effective-context Allocation for Context Training, to make this hidden allocation explicit. **EXACT** computes each supervised token's ***effective left-context length***, maps it into logarithmic buckets, and assigns fixed extra supervision weight to the long-context tail. Within that tail, the added weight is distributed by inverse bucket frequency, so rare long-effective-context targets become first-class training events rather than being washed out by the much larger population of short-context targets.

The empirical pattern follows this mechanism. Across seven Qwen and LLaMA continued-pretraining configurations, **EXACT** improves every trained and extrapolated NoLiMa/RULER aggregate in the main table. On Qwen2.5-0.5B 4K CPT, NoLiMa rises by +10.09 points in trained lengths and +5.34 in extrapolated lengths, while RULER rises by +10.69 and +5.55. On LLaMA-3.2-3B staged 4K→8K CPT, RULER improves by +17.91 and +16.11 in trained and extrapolated regimes. Standard QA is preserved, with a +0.24 macro change across MMLU, ARC-C, HellaSwag, WinoGrande, PIQA, and GSM8K. The strongest normalized gains occur in regimes where long effective-context supervision should matter most. A distance-resolved evidence probe then shows the same shape inside the model: **EXACT** leaves short distances near standard long-window CPT parity, but opens positive context-induced margin gains from 4K to 16K. The preservation and probe results show why the long-context scores move without sacrificing ordinary QA competence: distant evidence becomes more usable.

Our contributions are summarized as follows:

- We identify **effective-context exposure imbalance** as a measurable supervision-allocation bottleneck in packed causal language modeling.
- We propose **EXACT**, a **token-level objective** that assigns fixed extra supervision weight to the long-context tail and distributes it by **inverse effective-context bucket frequency**.
- We show consistent gains on official generation-based NoLiMa and RULER evaluations across seven Qwen and LLaMA continued-pretraining configurations, with standard QA

preserved and a distance-resolved evidence probe confirming stronger conversion of far context into answer likelihood.

2 Related Work

Long-context capacity and positional range. A large body of work extends the amount of context that language models can process. Transformer recurrence and memory mechanisms extend usable history across segments (Dai et al., 2019; Rae et al., 2019), while sparse or approximate attention reduces long-sequence cost through structured attention patterns or kernel approximations (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Kitaev et al., 2020; Choromanski et al., 2020). Efficient exact-attention and KV-cache methods improve the practicality of long-window training and inference (Dao et al., 2022; Dao, 2023; Liu et al., 2023; Ainslie et al., 2023), and architectures such as LongNet and Infini-attention explore extremely long or unbounded contexts (Ding et al., 2023; Munkhdalai et al., 2024). A complementary line extends the positional range of pretrained models through RoPE, ALiBi, interpolation, or rescaling (Su et al., 2024; Press et al., 2021; Chen et al., 2023a; Peng et al., 2023; Ding et al., 2024; Jin et al., 2024; Xiao et al., 2023; Han et al., 2024). These methods expand what context can be processed and where models can operate. **EXACT** studies a different training-objective question: once long windows are available, how much supervision reaches targets predicted under long same-document effective context?

Long-context continued pretraining and data supply. Continued pretraining and long-context fine-tuning are standard tools for adapting LLMs to longer inputs. LongLoRA makes long-context fine-tuning efficient through sparse attention and parameter-efficient adaptation (Chen et al., 2023b). LongAlign provides a long-context alignment recipe involving instruction data, packing, sorted batching, and loss weighting (Bai et al., 2024a). Data-engineering work for 128K context shows that long-context CPT depends on token budget, domain balance, and length upsampling (Fu et al., 2024). Public Qwen and LLaMA models provide common bases for these adaptation studies (Touvron et al., 2023; Grattafiori et al., 2024; Qwen et al., 2025). These works control the supply side: which documents enter training, how examples are packed, and how much token budget is spent at each window length. **EXACT** keeps the stream fixed and changes how supervision is allocated across the target tokens produced by that stream.

Long-context evaluation and context use. Evaluation work shows that available context is not the same as used context. Lost in the Middle demonstrates that models can fail to use relevant information depending on its position in the prompt (Liu et al., 2024). LongBench, L-Eval, BAMBOO, LooGLE, InfiniteBench, HELMET, ZeroSCROLLS, and NoCha broaden long-context evaluation across QA, summarization, retrieval, code, synthetic tasks, and application-centric settings (Bai et al., 2024b; An et al., 2024; Dong et al., 2024; Li et al., 2024; Zhang et al., 2024; Yen et al., 2024; Shaham et al., 2023; Karpinska et al., 2024). RULER adds configurable retrieval, tracking, aggregation, and QA tasks beyond vanilla needle retrieval (Hsieh et al., 2024). NoLiMa removes lexical-overlap shortcuts and requires latent association between the query and the evidence (Modarressi et al., 2025). We use generation-based NoLiMa and RULER because they test whether available long context is converted into answer production, matching our supervision-allocation hypothesis.

Token-level supervision and loss allocation. Non-uniform training signal is a long-standing idea, appearing in curriculum learning, self-paced learning, hard-example mining, focal loss, dataset cartography, and data-mixture optimization (Bengio et al., 2009; Kumar et al., 2010; Shrivastava et al., 2016; Lin et al., 2017; Swayamdipta et al., 2020; Xie et al., 2023). Recent language-modeling work also challenges uniform token training: Rho-1 uses a reference model to select useful tokens for pretraining (Lin et al., 2024), long-context perplexity analyses question whether aggregate perplexity reflects long-context ability (Fang et al., 2024), and Token Weighting for Long-Range Language Modeling uses short-context versus long-context model confidence to assign token weights (Helm et al., 2025). **EXACT** shares the view that token-level loss allocation matters, but targets a structural exposure variable rather than semantic token importance: the same-document effective left-context length of each supervised target in a packed causal-LM stream.

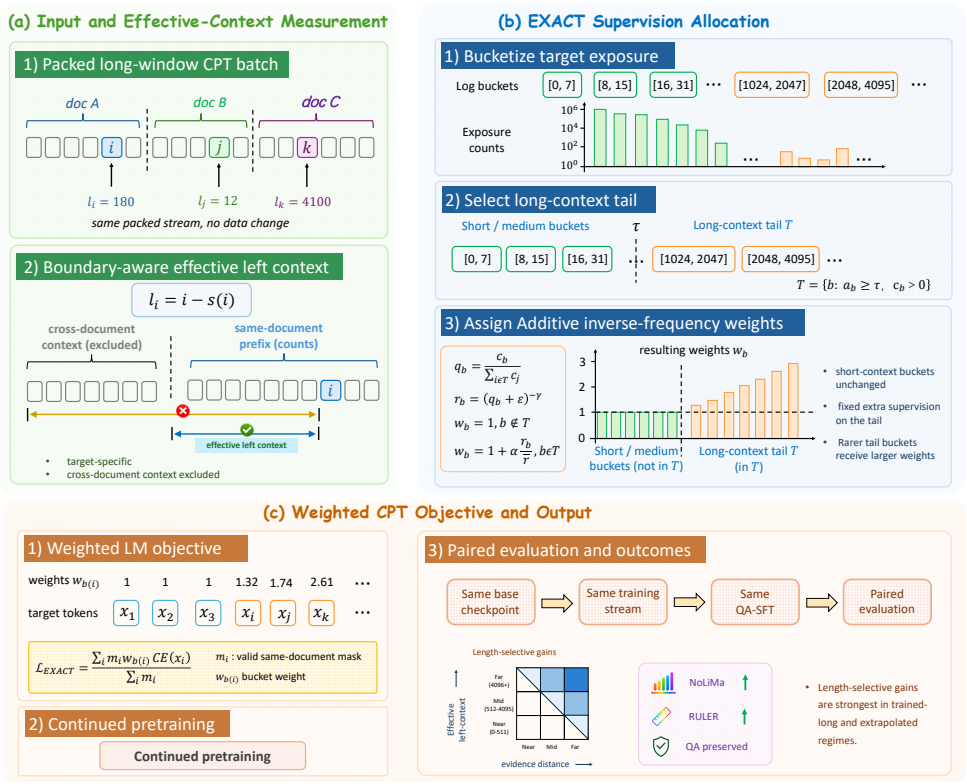

Figure 2: Overview of **EXACT**. The packed stream is unchanged; the objective changes only the per-token weighting induced by effective left-context buckets.

3 EXACT: EFFECTIVE-CONTEXT ALLOCATION

3.1 EFFECTIVE LEFT CONTEXT

Consider a packed causal language-modeling batch. For each target token position i, define its effective left-context length ℓ_i as the number of previous tokens that are both causally visible and inside the same document segment:

$$\ell_i = i - s(i), \tag{1}$$

where $s(i)$ is the start position of the current document segment inside the packed sequence. Cross-document tokens are excluded from ℓ_i under document-boundary masking. This quantity is target-specific: a sequence of length L contains targets with ℓ_i ranging from zero to at most $L-1$, with most targets inheriting their context length from their position inside a document segment.

3.2 EXPOSURE BUCKETS AND TOKEN WEIGHTS

EXACT groups target tokens by logarithmic effective-context buckets:

$$[0, 7], [8, 15], [16, 31], \ldots, [1024, 2047], [2048, 4095], \ldots \tag{2}$$

Let $b(i)$ be the bucket of target token i, and let c_b denote the number of supervised tokens in bucket b, estimated from the training stream. These bucket statistics summarize how the chosen corpus, packing policy, and document-boundary mask translate into token-level supervision. They also expose the gap between nominal window length and the actual allocation of training signal.

The reported **EXACT** objective uses additive directional inverse-frequency allocation. Given a long-context threshold τ, let $\mathcal{T} = \{b : a_b \geq \tau, c_b > 0\}$ be the long-context tail, where a_b is the lower bound of bucket b. Inside this tail, **EXACT** forms the tail-local bucket probability, inverse-frequency

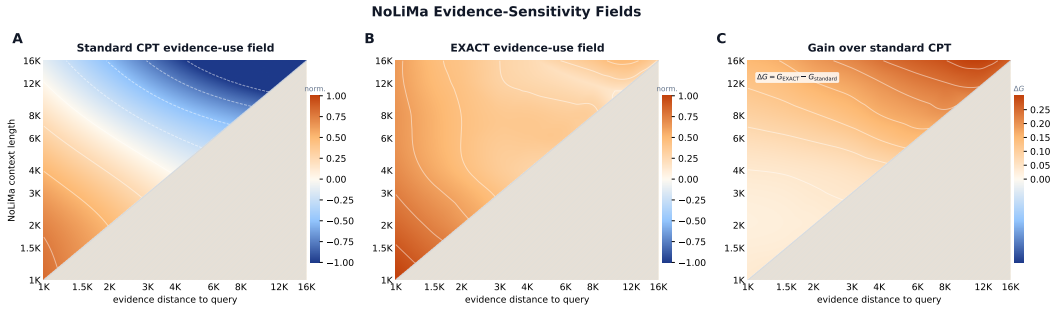

Figure 3: NoLiMa evidence-sensitivity fields on paired LLaMA-3.1-8B 16K continuation checkpoints. Each cell groups prompts by context length and evidence distance to the query; the lower-right geometry where distance exceeds context length is not constructible. For each cell, we first compute the mean context-induced answer margin G: the gold-answer margin in the original prompt minus the margin under a counterfactual prompt whose supporting evidence is replaced by a same-type distractor. Panels A and B visualize the resulting G fields under a shared robust display normalization for readability; Panel C reports the unnormalized gain $\Delta G = G_{\textbf{EXACT}} - G_{\text{standard}}$. Gains concentrate in the far-evidence, long-context regime targeted by the objective.

score, and frequency-weighted normalizer:

$$q_b = \frac{c_b}{\sum_{j\in\mathcal{T}} c_j}, \qquad r_b = (q_b + \epsilon)^{-\gamma}, \qquad \bar{r} = \sum_{j\in\mathcal{T}} q_j r_j. \tag{3}$$

The bucket weight is then

$$w_b = \begin{cases} 1, & b \notin \mathcal{T}, \\ 1 + \alpha \frac{r_b}{\bar{r}}, & b \in \mathcal{T}. \end{cases} \tag{4}$$

This construction leaves short-context buckets unchanged and assigns the long-context tail an average extra weight of α:

$$\sum_{b\in\mathcal{T}} q_b (w_b - 1) = \alpha. \tag{5}$$

The language-modeling objective becomes

$$\mathcal{L}_{\textbf{EXACT}} = \frac{\sum_i m_i \, w_{b(i)} \, \mathrm{CE}(x_i)}{\sum_i m_i}, \tag{6}$$

where m_i is the standard loss mask and $\mathrm{CE}(x_i)$ is the next-token cross-entropy at position i. The threshold τ is set relative to the training stage so that the extra loss weight is placed on the effective-context regime that the stage is meant to adapt. The hyperparameter α fixes the average extra weight assigned inside the long tail, while γ controls how sharply that added weight moves toward rarer long-context buckets.

The objective targets a concrete behavioral variable: whether distant evidence changes the model's preference for the correct answer. We measure this with a paired NoLiMa evidence-sensitivity diagnostic. For each prompt, we compare the gold-answer margin under the original context with the margin under a counterfactual context where the supporting evidence is replaced by a same-type distractor. Their difference gives a context-induced answer margin G, which measures how strongly the available evidence is converted into answer likelihood at a given distance.

Figure 3 shows the empirical footprint of the allocation change. Standard long-window CPT leaves the far/long corner weak: when the relevant evidence is distant and the context is long, the evidence contributes little to the answer margin. **EXACT** changes the field under the same prompt grid and visual scale. The strongest gains appear in the upper-right regime, precisely where standard CPT supplies relatively little long-effective-context supervision and where the reweighted tail should matter most.

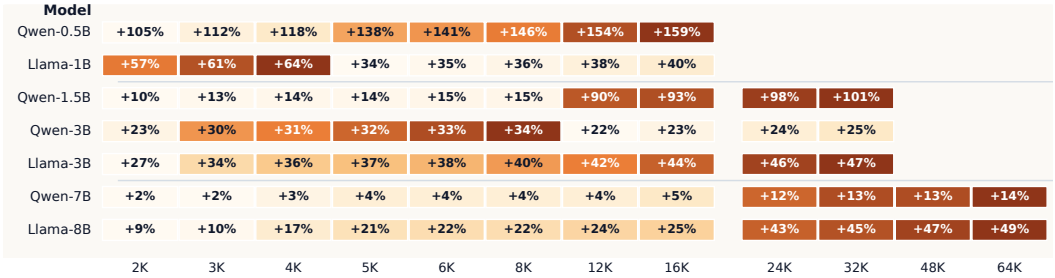

Figure 4: Length-resolved NoLiMa lift across Qwen and LLaMA configurations. Rows are model configurations and columns are evaluation lengths. Each cell reports the relative improvement of **EXACT** over standard long-window CPT; color is normalized within each row, so darker cells mark where that backbone gains most. Empty cells indicate lengths not evaluated for that configuration.

3.3 Expected Behavior

The same mechanism predicts length-selective improvement in downstream evaluation. Benchmarks that depend on trained-long and extrapolated long-context use should gain more than short-control regimes, because their success requires predictions conditioned on long effective left context. Ordinary short-context QA should be preserved because the objective still trains on all supervised targets while strengthening long-effective-context targets. Figure 3 makes this signature visible at the evidence-use level, and the following experiments test whether the benchmark pattern follows the same geometry.

4 Experimental Setup

Models. We evaluate **EXACT** on Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, LLaMA-3.2-1B, LLaMA-3.2-3B, and LLaMA-3.1-8B base models (Qwen et al., 2025; Touvron et al., 2023; Grattafiori et al., 2024). Each model is adapted by continued pretraining (CPT) followed by the same question-answer supervised fine-tuning (QA-SFT) procedure on Databricks Dolly-15K instruction-response data (Conover et al., 2023) for generation-based evaluation. The QA-SFT stage is held fixed across standard long-window CPT and **EXACT**.

Continued pretraining. Following common long-context extension practice, we use staged CPT before QA-SFT and keep each standard-CPT/**EXACT** comparison paired in initialization, token budget, data stream, QA-SFT corpus, and evaluation protocol. Appendix A reports the stage and corpus details. This pairing isolates effective-context supervision allocation. Unless otherwise stated, **EXACT** uses $\alpha = 0.15$, $\gamma = 0.5$, and $\epsilon = 10^{-4}$; the long-tail threshold is set to $\tau = 1$K, 2K, and 4K for the 4K, 8K, and 16K CPT stages, respectively.

Benchmarks. Long-context performance is measured with official generation-based NoLiMa and RULER evaluation. NoLiMa is reported over six public task sets: base, hard, distractor, multiple choice, only-direct, and chain-of-thought-style prompts. RULER is reported over 11 synthetic retrieval and tracking tasks with 500 examples per task. For standard QA and reasoning preservation, we evaluate MMLU, ARC-Challenge, HellaSwag, WinoGrande, PIQA, and GSM8K. Multiple-choice tasks use likelihood scoring over answer choices; GSM8K uses generation with numeric exact match.

5 Main Results

Table 1 reports results across Qwen and LLaMA backbones and multiple continued-pretraining scales. **EXACT** improves long-context generation benchmarks in every setting. The gains cover two complementary regimes: large relative lifts in low-baseline NoLiMa settings, especially Qwen2.5-0.5B, and large absolute RULER gains on LLaMA-3.2-3B. Across all models, **EXACT** raises NoLiMa and RULER in trained-context regimes and continues to provide gains in extrapolated regimes.

Table 1: Main results across Qwen and LLaMA backbones. Scores are percentages (%). NoLiMa and RULER use generation-based evaluation.

Backbone Family	Model	CPT window	Evaluation	Evaluation windows	Standard CPT	EXACT	Δ
Qwen	Qwen2.5-0.5B	4K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K	9.04	19.13	+10.09 (↑)
	Qwen2.5-0.5B	4K	NoLiMa extrap	5K, 6K, 8K, 12K, 16K	3.62	8.96	+5.34 (↑)
	Qwen2.5-0.5B	4K	RULER trained	1K, 1.5K, 2K, 3K, 4K	46.79	57.47	+10.69 (↑)
	Qwen2.5-0.5B	4K	RULER extrap	5K, 6K, 8K, 12K, 16K	45.49	51.04	+5.55 (↑)
	Qwen2.5-1.5B	8K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K	16.91	19.18	+2.28 (↑)
	Qwen2.5-1.5B	8K	NoLiMa extrap	12K, 16K, 24K, 32K	3.88	7.59	+3.71 (↑)
	Qwen2.5-1.5B	8K	RULER trained	1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K	76.63	77.65	+1.02 (↑)
	Qwen2.5-1.5B	8K	RULER extrap	12K, 16K, 24K, 32K	61.79	63.28	+1.49 (↑)
	Qwen2.5-3B	8K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K	18.84	24.56	+5.73 (↑)
	Qwen2.5-3B	8K	NoLiMa extrap	12K, 16K, 24K, 32K	6.57	8.13	+1.56 (↑)
	Qwen2.5-3B	8K	RULER trained	1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K	80.53	82.32	+1.79 (↑)
	Qwen2.5-3B	8K	RULER extrap	12K, 16K, 24K, 32K	72.02	74.70	+2.69 (↑)
	Qwen2.5-7B	16K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K, 12K, 16K	26.38	27.29	+0.91 (↑)
	Qwen2.5-7B	16K	NoLiMa extrap	24K, 32K, 48K, 64K	10.50	11.85	+1.35 (↑)
	Qwen2.5-7B	16K	RULER trained	1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K, 12K, 16K	87.58	88.67	+1.09 (↑)
	Qwen2.5-7B	16K	RULER extrap	24K, 32K, 48K, 64K	72.24	73.42	+1.18 (↑)
LLaMA	LLaMA-3.2-1B	4K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K	9.55	15.35	+5.80 (↑)
	LLaMA-3.2-1B	4K	NoLiMa extrap	5K, 6K, 8K, 12K, 16K	8.50	11.61	+3.12 (↑)
	LLaMA-3.2-1B	4K	RULER trained	1K, 1.5K, 2K, 3K, 4K	48.84	50.58	+1.74 (↑)
	LLaMA-3.2-1B	4K	RULER extrap	5K, 6K, 8K, 12K, 16K	48.93	49.76	+0.83 (↑)
	LLaMA-3.2-3B	8K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K	30.08	40.65	+10.57 (↑)
	LLaMA-3.2-3B	8K	NoLiMa extrap	12K, 16K, 24K, 32K	14.42	20.87	+6.45 (↑)
	LLaMA-3.2-3B	8K	RULER trained	1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K	57.96	75.87	+17.91 (↑)
	LLaMA-3.2-3B	8K	RULER extrap	12K, 16K, 24K, 32K	53.87	69.98	+16.11 (↑)
	LLaMA-3.1-8B	16K	NoLiMa trained	0.5K, 1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K, 12K, 16K	30.68	36.41	+5.73 (↑)
	LLaMA-3.1-8B	16K	NoLiMa extrap	24K, 32K, 48K, 64K	14.29	20.83	+6.54 (↑)
	LLaMA-3.1-8B	16K	RULER trained	1K, 1.5K, 2K, 3K, 4K, 5K, 6K, 8K, 12K, 16K	87.12	89.28	+2.16 (↑)
	LLaMA-3.1-8B	16K	RULER extrap	24K, 32K, 48K, 64K	78.28	80.34	+2.06 (↑)

Across seven model configurations, **EXACT** improves every trained and extrapolated NoLiMa/RULER row in Table 1. Figure 4 resolves the NoLiMa aggregates by context length and reports relative lift over standard long-window CPT. The dominant pattern is length-selective: within most backbone rows, the largest relative gains appear at longer trained or extrapolated lengths rather than being uniformly spread across all contexts. This profile matches the supervision-allocation hypothesis and complements the aggregate RULER gains in Table 1; paired bootstrap intervals are reported in Appendix C. Robustness evidence in Tables 4, 6, and 9 covers HELMET transfer, HELM, and a lightweight hyperparameter sweep: **EXACT** improves every HELMET and HELM paired comparison, and the default setting gives the best average balance in the Qwen2.5-0.5B sweep.

6 Controls and Mechanism

6.1 Standard QA and reasoning performance is preserved

Table 5 measures ordinary QA and reasoning competence after the fixed QA-SFT procedure. Averaged over Qwen2.5-1.5B, Qwen2.5-3B, and Llama-3.2-3B, the macro score changes by +0.24 points across MMLU, ARC-C, HellaSwag, WinoGrande, PIQA, and GSM8K. This separates the long-context gains from degradation of ordinary QA: **EXACT** strengthens supervision on under-exposed long-effective-context regimes while preserving standard QA.

6.2 Weighting signal ablation

Table 2 isolates the weighting signal in the 16K Llama-3.1-8B setting. The controls keep the checkpoint, data stream, token budget, optimizer, QA-SFT data, and evaluation protocol fixed; only the token weighting rule changes. Random same-mass weighting hurts all four aggregates, ruling out generic loss reweighting as the source of the gain. Packed-position weighting gives small improvements, showing that absolute packed position is a weak proxy.

The effective-context controls give the decisive ordering. Uniformly boosting the same-document effective-context tail is much stronger than packed-position weighting. Normalized **EXACT** uses

Table 2: Ablation of the weighting signal on Llama-3.1-8B 16K continuation. All variants use the same checkpoint, 16K continuation stream, token budget, optimizer schedule, QA-SFT data, and evaluation protocol. Scores are point gains over standard CPT.

Method	NoLiMa trained Δ	NoLiMa extrap Δ	RULER trained Δ	RULER extrap Δ
Standard CPT	0.00	0.00	0.00	0.00
Random same-mass weighting	-0.41	-0.73	-0.18	-0.52
Packed-position weighting	+1.34	+1.17	+0.63	+0.28
Uniform effective-context boost	+4.38	+5.12	+1.57	+1.31
Normalized **EXACT**	+5.61	+6.38	+2.08	+1.98
EXACT	**+5.73**	**+6.54**	**+2.16**	**+2.06**

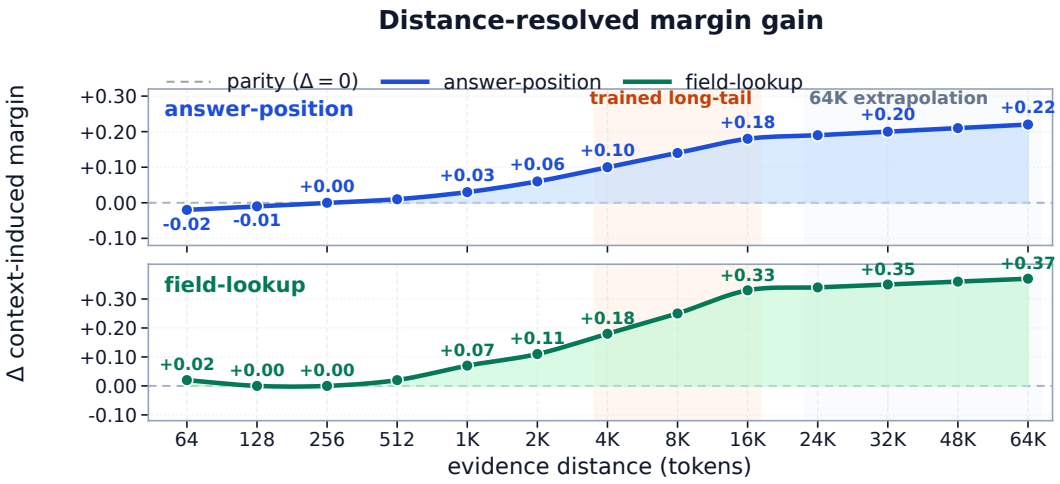

Figure 5: Evidence-conversion probe before QA-SFT. Panel A plots distance-resolved gains in context-induced answer margin, measured as **EXACT** minus standard long-window CPT, across two hard-cloze views. The dashed line marks standard long-window CPT parity. Short distances remain close to zero, the 4K–16K regime opens into positive gains, and the gain remains elevated through the 64K evaluation range. Panel B uses the same 4K–64K probes to show the corresponding answer-level lift.

the same effective-context weighting rule under weighted-loss normalization and nearly matches **EXACT**, while **EXACT** remains strongest on every aggregate. The ablation therefore separates three factors: arbitrary reweighting fails, packed position is only a weak proxy, and effective left context is the variable that drives the gains. The same ordering is stable under paired bootstrap resampling with the table's macro-averaging protocol (Appendix D).

6.3 Mechanistic evidence in Llama-3.1-8B

Before QA-SFT, the evidence-ablation probe measures whether the allocation change is already visible in the base model. The probe presents cloze-style questions with and without the relevant evidence, then measures how much that evidence shifts the answer margin. Figure 5 gives the resulting distance structure. The short-distance region stays near standard long-window CPT parity, while the field changes shape when evidence moves into the long-distance regime. Across 4K–64K evidence distances, **EXACT** raises the context-induced margin by +0.177 on answer-position and +0.311 on field lookup. The answer-level lift over the same 4K–64K regime follows the same ordering, with +2.20 and +3.20 points. This closes the mechanism loop opened by Figure 3: **EXACT** strengthens rare long-effective-context targets, the main benchmarks improve where long context is needed, and the probe shows the intermediate variable moving in the right place, namely the conversion of distant evidence into answer likelihood.

7 ANALYSIS

Mechanistic Chain. The paper's central claim is supported by a four-step chain. Figure 1 exposes where supervision actually lands under packed causal training: long windows create the possibility of long context, but many supervised targets still receive short effective context. Figure 3 shows the evidence-use geometry: **EXACT** concentrates gains in the far and long-context regime. Figure 4 shows that the benchmark results follow this geometry across all seven model configurations and both trained/extrapolated regimes, after normalizing for baseline difficulty. Figure 5 then closes the loop inside the model. The distance-resolved probe stays near standard long-window CPT parity at short evidence distances, but the long-distance regime becomes strongly positive, with margin gains of +0.177 on answer-position and +0.311 on field lookup across 4K–64K. The evidence chain therefore links benchmark gains to the same location highlighted by the evidence-use field.

Data Supply and Supervision Allocation. Long-document data and mixed-length curricula decide what context is available to the model. Effective-context allocation decides how much supervised loss each target-context regime receives after that data has been selected and packed. This distinction sharpens the contribution. Data construction supplies long contexts; **EXACT** converts the resulting token-level exposure histogram into an explicit training control. The paired experiments keep the stream fixed, so the gains identify supervision allocation inside the same data supply.

Length-Selective Gains. The result pattern matters more than any single score. Table 1 shows improvements in all twenty-eight trained/extrapolated NoLiMa and RULER aggregates across Qwen and LLaMA configurations. Because absolute gains depend on baseline difficulty, Figure 4 also reports relative lift for the full table: NoLiMa extrapolation averages +58%, NoLiMa trained +39%, while RULER trained and extrapolated improve by +9% and +8% despite much stronger standard-CPT baselines. Standard QA is preserved at +0.24 macro. Together, these results support effective-context exposure as a real bottleneck in long-context adaptation and show that correcting it yields mechanism-aligned gains.

Training Cost. **EXACT** changes the token-level objective without changing the model architecture, sequence length, or training-token budget. Effective-context buckets follow directly from document boundaries already present in the packed stream; training adds only a bucket lookup and scalar loss weighting for each supervised token. When the corpus or packing policy changes, the bucket statistics can be recomputed from the new stream without modifying the model. The paired runs therefore use the same forward and backward sequence shapes as standard CPT, and the trained model incurs no additional inference cost.

8 CONCLUSION AND LIMITATIONS

Long-context adaptation has been framed largely around extending what a model can see. This paper shows that training must also control how strongly the model is supervised under long effective context. In packed causal language modeling, nominal sequence length and token-level supervision length diverge because each target token is supervised under its own same-document left context. **EXACT** corrects this supervision-allocation failure by assigning extra supervision weight to long-effective-context targets and distributing that weight by inverse bucket frequency inside the long tail. Across seven Qwen and Llama continued-pretraining configurations, the method produces consistent gains on official generation-based NoLiMa and RULER evaluations while preserving standard QA. The evidence-conversion probe adds the missing link: after **EXACT** strengthens long-tail supervision, distant evidence is more strongly converted into answer likelihood in the long-distance regime. Together, the results make the training implication concrete: effective long-context adaptation depends on wider windows and on supervision that reaches tokens predicted under long effective context. ***Limitation.*** EXACT is evaluated in paired long-context continued-pretraining settings; we do not test computationally heavy from-scratch pretraining, where corpus scale and early-stage representation learning may interact differently with effective-context allocation.

9 AI USAGE STATEMENT

This work is conceived, designed, and technically developed by the authors. Large language models (LLMs) were used only for limited writing assistance, including grammar correction, readability improvement, formatting refinement, and minor proofreading during manuscript preparation. All research ideas, technical contributions, experiments, analyses, claims, and conclusions were developed and verified by the authors, who take responsibility for the final content of this work.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024a.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024b.

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023a.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongLoRA: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023b.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023. Accessed: 2023-06-30.

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

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. LongNet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2086–2099, 2024.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? *arXiv preprint arXiv:2410.23771*, 2024.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3991–4008, 2024.

Falko Helm, Nico Daheim, and Iryna Gurevych. Token weighting for long-range language modeling. *arXiv preprint arXiv:2503.09202*, 2025.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Jia Fei, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A "novel" challenge for long-context language models. *arXiv preprint arXiv:2406.16264*, 2024.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *Advances in Neural Information Processing Systems*, 2024.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023.

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

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. FineWeb-Edu: The finest collection of educational content. https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu, 2024.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, 2023.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627b token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023.

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

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, 2023.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, 2020.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. Infinitebench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024.

A CPT Corpus Construction Details

All reported standard-CPT and **EXACT** comparisons are constructed as paired runs: within a backbone and CPT stage, the compared models share the same packed stream, document-boundary mask, train/validation split, token budget, optimizer schedule, QA-SFT corpus, and evaluation protocol. The only intended difference is the per-token loss weighting rule. Table 3 reports the corpus recipe; the shared controls are kept out of the table because they are identical by construction. The public sources are FineWeb-Edu, SlimPajama, and OpenWebMath (Lozhkov et al., 2024; Soboleva et al., 2023; Paster et al., 2023).

Table 3: Corpus construction used for the paired CPT comparisons in Table 1. Each CPT stage uses a 500M-token budget.

CPT stage	**Backbones**	**Source mix**
4K	Qwen2.5-0.5B; Llama-3.2-1B	FineWeb-Edu
4K→8K	Qwen2.5-1.5B; Qwen2.5-3B; Llama-3.2-3B	FineWeb-Edu : SlimPajama-6B : OpenWebMath = 6:24:32
4K→8K→16K	Qwen2.5-7B; Llama-3.1-8B	FineWeb-Edu : SlimPajama-6B : OpenWebMath = 1:3:6

For reproducibility, each packed dataset stores a manifest containing the raw source files, tokenizer path, sequence length, split sizes, document-boundary arrays, and token counts. The bucket statistics used by **EXACT** are computed from the training split of the same manifest, so the reported gains compare supervision allocation on an identical data supply rather than different corpus mixtures.

B Additional Evaluation Details

Table 4: HELMET evaluation at 16K on completed paired configurations. Scores are percentages (%).

Model	Standard CPT	**EXACT**	Δ
Qwen2.5-0.5B	31.74	33.17	+1.43
Qwen2.5-1.5B	49.26	50.07	+0.81
Qwen2.5-3B	57.92	58.59	+0.67
LLaMA-3.2-1B	47.86	48.60	+0.74
LLaMA-3.2-3B	56.38	58.76	+2.38
Average	**48.63**	**49.84**	**+1.21**

Table 5: QA and reasoning preservation averaged over Qwen2.5-1.5B, Qwen2.5-3B, and Llama-3.2-3B. Scores are percentages; multiple-choice tasks use likelihood scoring, and GSM8K uses numeric exact match.

Benchmark	Standard CPT	**EXACT**	Δ
ARC-C	45.37	45.37	+0.00
GSM8K	11.17	11.25	+0.08
HellaSwag	65.69	65.68	-0.01
MMLU	47.95	49.63	+1.68
PIQA	75.07	75.00	-0.07
WinoGrande	65.93	65.72	-0.21
Macro	**51.85**	**52.11**	**+0.24**

Table 6: Comparison with Helm et al. (2025) on two representative Qwen configurations. Scores are point gains over standard CPT; all methods use the same CPT token budget and evaluation protocol.

Model / CPT	Method	NoLiMa trained Δ	NoLiMa extrap Δ	RULER trained Δ	RULER extrap Δ	Avg. Δ
Qwen2.5-0.5B / 4K	Helm	+5.50	+2.80	+5.00	+2.40	+3.93
Qwen2.5-0.5B / 4K	**EXACT**	**+10.09**	**+5.34**	**+10.69**	**+5.55**	**+7.92**
Qwen2.5-3B / 8K	Helm	+2.40	+0.70	+0.80	+1.30	+1.30
Qwen2.5-3B / 8K	**EXACT**	**+5.73**	**+1.56**	**+1.79**	**+2.69**	**+2.94**

C Bootstrap Confidence Intervals for Main Results

Table 7 reports paired bootstrap 95% confidence intervals for the main NoLiMa/RULER aggregates. For each row, resampling follows the same evaluation-cell aggregation used by the corresponding Table 1 metric.

Table 7: Paired bootstrap 95% confidence intervals for the main-result gains in Table 1. Values are point gains over standard CPT.

Qwen			Llama		
Model / CPT	Eval.	Δ with 95% CI	Model / CPT	Eval.	Δ with 95% CI
Qwen2.5-0.5B / 4K	NoLiMa train	+10.09 [+8.42, +11.66]	Llama-3.2-1B / 4K	NoLiMa train	+5.80 [+4.38, +7.06]
Qwen2.5-0.5B / 4K	NoLiMa extrap	+5.34 [+3.92, +6.61]	Llama-3.2-1B / 4K	NoLiMa extrap	+3.12 [+1.70, +4.39]
Qwen2.5-0.5B / 4K	RULER train	+10.69 [+9.12, +12.04]	Llama-3.2-1B / 4K	RULER train	+1.74 [+0.88, +2.51]
Qwen2.5-0.5B / 4K	RULER extrap	+5.55 [+4.21, +6.78]	Llama-3.2-1B / 4K	RULER extrap	+0.83 [-0.03, +1.59]
Qwen2.5-1.5B / 8K	NoLiMa train	+2.28 [+1.18, +3.25]	Llama-3.2-3B / 8K	NoLiMa train	+10.57 [+8.84, +12.15]
Qwen2.5-1.5B / 8K	NoLiMa extrap	+3.71 [+2.21, +5.08]	Llama-3.2-3B / 8K	NoLiMa extrap	+6.45 [+4.80, +7.96]
Qwen2.5-1.5B / 8K	RULER train	+1.02 [+0.30, +1.68]	Llama-3.2-3B / 8K	RULER train	+17.91 [+16.31, +19.37]
Qwen2.5-1.5B / 8K	RULER extrap	+1.49 [+0.56, +2.34]	Llama-3.2-3B / 8K	RULER extrap	+16.11 [+14.23, +17.84]
Qwen2.5-3B / 8K	NoLiMa train	+5.73 [+4.54, +6.82]	Llama-3.1-8B / 16K	NoLiMa train	+5.73 [+4.58, +6.74]
Qwen2.5-3B / 8K	NoLiMa extrap	+1.56 [+0.28, +2.71]	Llama-3.1-8B / 16K	NoLiMa extrap	+6.54 [+4.97, +7.91]
Qwen2.5-3B / 8K	RULER train	+1.79 [+1.05, +2.48]	Llama-3.1-8B / 16K	RULER train	+2.16 [+1.26, +2.94]
Qwen2.5-3B / 8K	RULER extrap	+2.69 [+1.68, +3.58]	Llama-3.1-8B / 16K	RULER extrap	+2.06 [+1.02, +2.98]
Qwen2.5-7B / 16K	NoLiMa train	+0.91 [-0.02, +1.76]			
Qwen2.5-7B / 16K	NoLiMa extrap	+1.35 [+0.03, +2.49]			
Qwen2.5-7B / 16K	RULER train	+1.09 [+0.41, +1.69]			
Qwen2.5-7B / 16K	RULER extrap	+1.18 [+0.32, +1.93]			

D Bootstrap Confidence Intervals for the Ablation Study

Table 8 reports paired bootstrap 95% confidence intervals for the Llama-3.1-8B ablation study. The bootstrap matches the reported metric: examples are resampled within each evaluation cell, and each replicate is aggregated with the same macro-averaging procedure used in Table 2.

Table 8: Paired bootstrap 95% confidence intervals for the ablation gains in Table 2. Values are point gains over standard CPT.

Method	NoLiMa trained Δ	NoLiMa extrap Δ	RULER trained Δ	RULER extrap Δ
Random same-mass weighting	-0.41 [-0.92, +0.06]	-0.73 [-1.41, -0.10]	-0.18 [-0.47, +0.09]	-0.52 [-0.91, -0.15]
Packed-position weighting	+1.34 [+0.73, +1.92]	+1.17 [+0.39, +1.88]	+0.63 [+0.26, +0.98]	+0.28 [-0.12, +0.70]
Uniform effective-context boost	+4.38 [+3.62, +5.10]	+5.12 [+4.10, +6.03]	+1.57 [+1.07, +2.02]	+1.31 [+0.78, +1.82]
Normalized **EXACT**	+5.61 [+4.82, +6.37]	+6.38 [+5.31, +7.29]	+2.08 [+1.54, +2.57]	+1.98 [+1.38, +2.52]
EXACT	**+5.73** [**+4.91**, **+6.48**]	**+6.54** [**+5.42**, **+7.46**]	**+2.16** [**+1.60**, **+2.67**]	**+2.06** [**+1.44**, **+2.62**]

E Hyperparameter Sensitivity

Table 9 reports a lightweight sensitivity sweep on Qwen2.5-0.5B 4K CPT. The default setting gives the strongest average gain, while nearby settings expose the expected tradeoff between NoLiMa and RULER.

Table 9: Hyperparameter sensitivity on Qwen2.5-0.5B 4K CPT. Scores are point gains over standard CPT; the average is over the four reported NoLiMa/RULER aggregates.

Variant	α	γ	τ	NoLiMa trained Δ	NoLiMa extrap Δ	RULER trained Δ	RULER extrap Δ	Avg. Δ
Low-alpha	0.10	0.5	1K	+5.20	+2.40	+9.80	+5.20	+5.65
Main	0.15	0.5	1K	**+10.09**	+5.34	**+10.69**	**+5.55**	**+7.92**
High-alpha	0.20	0.5	1K	+11.00	**+6.40**	+4.20	+1.60	+5.80
Shifted-tail	0.15	0.5	2K	+5.80	+3.20	+1.20	+0.00	+2.55

F Reproducibility and Ethics Statements

Reproducibility. All reported paired runs use 8 H100 GPUs and match checkpoint, data stream, optimizer schedule, sequence length, QA-SFT data, and evaluation protocol. Training uses AdamW with learning rate 2×10^{-5}, weight decay 0.01, gradient clipping at 1.0, linear warmup then constant learning rate, budget-derived CPT steps, 600 QA-SFT steps, and greedy decoding.

Ethics. This work studies training objectives on public text and benchmarks; it does not introduce human-subject data collection or deployment-specific user data.