LONGFOCUS: FOCUSING ON KEY TOKENS FOR EFFICIENT LONG-CONTEXT MID-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite extensive progress in data selection for long-context mid-training, the objective typically remains a uniform token-averaged cross-entropy. Since only a small fraction of tokens substantially benefit from distant context, uniform averaging can dilute long-range supervision even within dependency-rich documents. Existing token-reweighting approaches address this mismatch, but their fine-grained token dependency scoring method introduces substantial training overhead. We therefore investigate whether a coarse context contrast can suffice when coupled with appropriate preprocessing. We propose LONGFOCUS, an online objective that combines an inexpensive non-overlapping segmented reference with eligibility filtering and bounded rank-percentile upweighting for long-context mid-training. Specifically, LONGFOCUS computes an inexpensive online dependency score, SEGRATIO, by measuring each token's relative loss increase from the full-context view to a coarse segmented view with no cross-segment attention. *PreSelect* restricts additional weighting by document length and full-view loss, while *Reweight* assigns bounded weights according to rank percentiles within the eligible set. We evaluate LONGFOCUS on Llama-3-8B-Instruct, MoE-16B-A3B, and MoE-125B-A10B using a two-stage 64K-to-512K mid-training recipe, with evaluations before and after SFT. LONGFOCUS improves Long-Avg over Vanilla CE in all matched model-stage comparisons. On MoE-125B-A10B, the post-SFT gains are 7.28 points (13.80% relative improvement) after 64K mid-training and 5.48 points (10.24% relative improvement) after the subsequent 512K stage. On Llama-3-8B-Instruct, it outperforms LongCE in Long-Avg at both post-SFT endpoints, with gains of up to 2.0 points (3.29% relative improvement), while reducing per-step training time and peak memory usage by up to 20.2% and 31.0%, respectively. Code is available at `https://anonymous.4open.science/r/LongFocus-vsfnhyta`.

1 INTRODUCTION

Long-context capability has become a fundamental requirement for modern LLMs, enabling repository-level code understanding, multi-document reasoning, and long-horizon agentic workflows. A prevailing approach to acquiring this capability is *long-context mid-training*: a pretrained model with a moderate context window is further trained on long sequences, often followed by supervised fine-tuning (Xiong et al., 2024; Fu et al., 2024; Gao et al., 2024). A substantial body of work has therefore focused on improving the *data* used in this stage, through better domain and length mixtures, dependency-aware filtering, and synthetic examples that explicitly require distant information (Chen et al., 2024a; Wu et al., 2025; Deng et al., 2025; Yang et al., 2025).

A long sequence, however, does not necessarily provide dense supervision for learning long-range dependencies. Most tokens remain predictable from nearby context, while only a relatively small subset benefits substantially from long-range context (Fang et al., 2025). Existing data-centric methods primarily address this sparsity *across documents* by selecting documents that exhibit stronger long-range dependencies. (Wu et al., 2025; Deng et al., 2025; Chen et al., 2024a). However, even within dependency-rich documents, locally predictable tokens can still dominate the average standard cross-entropy objective, thereby diluting the true long-range supervision(see Figure 3). We refer to this objective-level mismatch as *intra-document signal dilution*.

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.

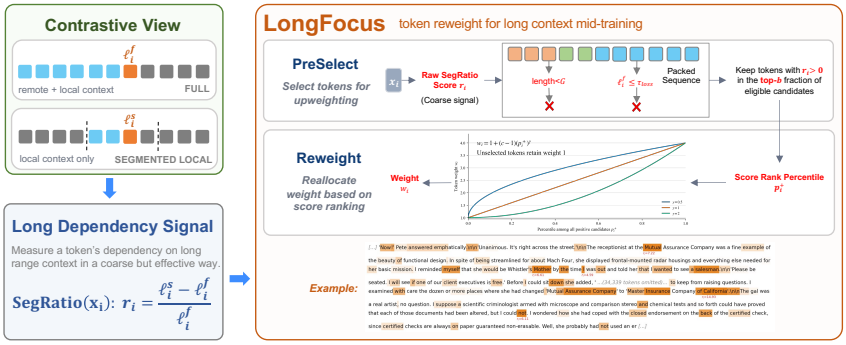


Figure 1: Overview of LONGFOCUS

A natural remedy is to identify tokens whose predictions benefit from distant information and assign them greater optimization emphasis. Prior work has shown that contrasting long- and short-context predictions can guide token weighting for long-context training (Fang et al., 2025; Helm et al., 2025). Building on these results, we study whether effective online reweighting can be obtained from a low-cost, non-overlapping segmented reference. Such a reference avoids repeated evaluation of overlapping context, but provides different amounts of local history at different token positions. Our focus is therefore the joint design of the reference, candidate eligibility, and weight allocation, rather than token reweighting itself or precise estimation of each token's long-range dependence.

We propose LONGFOCUS, an online mid-training objective that turns the coarse full-versus-segmented contrast SEGRATIO into bounded additional training emphasis through candidate filtering and rank-based weight allocation. LONGFOCUS features 2 stages: PreSelect restricts additional weighting to tokens in sufficiently long documents whose full-view loss meets a minimum threshold, retaining positive-score candidates within a top-ranked budget; Reweight uses rank percentiles rather than raw score magnitudes to determine additional emphasis, with bounded increments over a unit-weight cross-entropy background. Filtering removes noisy contrasts, while the rank-based map reduces reliance on its numerical calibration. The full-view losses are reused from ordinary training, and the segmented reference requires one additional forward pass without backpropagation. We evaluate this specific combination through downstream training results, component ablations, and comparisons of training overhead.

Our main contributions are as follows:

- **Robust long-context training from coarse dependency signals.** We introduce LONGFOCUS, which couples an inexpensive segmented reference with candidate filtering and bounded rank-percentile upweighting over a unit-weight CE background. The design addresses how coarse scores are used to allocate training emphasis, reducing reliance on their numerical calibration rather than seeking precise token-level dependency estimates.
- **Large-scale effectiveness and practical scalability.** We evaluate LONGFOCUS on dense and MoE models with up to 125B total parameters using training contexts up to 512K. LONGFOCUS improves Long-Avg over Vanilla CE in all 12 matched model-stage comparisons, including both post-SFT endpoints on all three models. On MoE-125B-A10B, the post-SFT gains reach 7.28 points (13.80% relative) after the 64K stage and 5.48 points (10.24% relative) after the 512K continuation. On Llama-3-8B-Instruct, it also exceeds LongCE at both post-SFT endpoints, while decreasing per-step training time and peak memory by up to 20.2% and 31.0%, respectively. In the tested MoE configurations, LONGFOCUS remains feasible under the same GPU budgets for which the evaluated LongCE implementation encounters out-of-memory errors.

2 RELATED WORK

To improve long-context capabilities, one line of work focus on architectural and systemic advances (Dai et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Gu et al., 2022; Gu & Dao, 2024; Su et al., 2024; Press et al., 2022; Chen et al., 2023; Peng et al., 2024; Ding et al., 2024; Chen et al., 2024b; Dao et al., 2022; Dao, 2024; Liu et al., 2024; Jacobs et al., 2023; Xiao et al., 2024). On the data side, most studies emphasize on which training data best supports long-context learning, including long-document pretraining (Xiong et al., 2024), data quantity and mixtures (Fu et al., 2024; Gao et al., 2024), and data augmentation (Tian et al., 2024). Beyond these recipes, various

dependency-aware data selection methods such as LongFilter (Deng et al., 2025) and GATEAU (Si et al., 2025) prioritize examples with informative long-range dependencies, motivating scoring methods that quantify the contribution of distant context. Likelihood-based methods include ProLong-LDS (Chen et al., 2024a), which combines perplexity reduction with dependency distance and specificity, and LongPPL (Fang et al., 2025), which identifies context-sensitive tokens through long-versus-short likelihood contrasts. Attention-based methods instead use distant attention as a dependency proxy, as in LongAttn (Wu et al., 2025) and CDS/LADM (Chen et al., 2025). However, selecting dependency-rich examples does not by itself resolve the dilution of long-range supervision within each sequence. Related work therefore seeks token-level supervision by identifying informative tokens. LongCE (Fang et al., 2025) computes online token weights from capped long-to-short likelihood ratios using the model being trained, with overlapping windows for short-context scoring. (Helm et al., 2025) study token weighting through a scoring-and-postprocessing framework, including sparse quantile selection, dense normalized weighting with uniform interpolation, and different choices of short-context models. These works establish important ingredients for context-aware token weighting. Rho-1 (Lin et al., 2024) uses excess loss relative to a reference model to select tokens for language-model training. Building on this line of work, LONGFOCUS combines a non-overlapping segmented reference with document-length and full-view-loss eligibility filters and bounded rank-dependent upweighting. Compared with LongCE, it changes both the reference construction and the score-to-weight mapping. Within its selected subset, rank percentiles determine the magnitudes of the additional weights, rather than serving only as a binary selection criterion. All valid tokens retain a unit-weight CE contribution. Our contribution is this specific online scoring-and-weighting recipe and its empirical evaluation for long-context mid-training.

3 LONGFOCUS: GUIDING LONG CONTEXT MID-TRAINING WITH COARSE DEPENDENCY SIGNAL

To alleviate *intra-document dilution*, long-context mid-training should place greater emphasis on tokens that benefit from distant context. Doing so requires a dependency signal that is informative yet inexpensive to compute online.

Our design combines low-cost segmented reference scoring with candidate filtering and bounded rank-based upweighting. We first define SEGRATIO, a full-versus-segmented relative-loss contrast (Section 3.1). We then specify candidate eligibility through *PreSelect* and allocate bounded additional weights through *Reweight* (Section 3.2), while retaining a unit-weight CE contribution for every valid token.

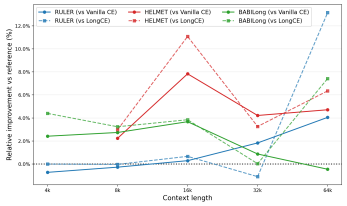
3.1 SEGRATIO: A LIGHTWEIGHT COARSE CONTRASTIVE SIGNAL FOR LONG-RANGE DEPENDENCY

A token's prediction benefits from additional context when removing that context increases its prediction loss. To estimate this benefit efficiently, we construct a *segmented local view*: the input is partitioned into non-overlapping segments of length S and all cross-segment attention is blocked (see Figure 1). The segments can be processed together without repeatedly evaluating overlapping reference windows, as used in LongCE (Fang et al., 2025), which trades uniform local-context coverage for lower computation cost. We therefore use the resulting contrast as a coarse proxy for long-range dependency. For an input $\mathbf{x} = (x_1, \ldots, x_N)$, we compute two next-token losses using the same model parameters θ: $\ell_i^{\mathrm{f}} = -\log p_\theta^{\mathrm{f}}(x_i \mid x_{<i}), \ell_i^{\mathrm{s}} = -\log p_\theta^{\mathrm{s}}(x_i \mid x_{<i})$ where p_θ^{f} uses the standard full-view attention mask and p_θ^{s} additionally blocks attention across segment boundaries.

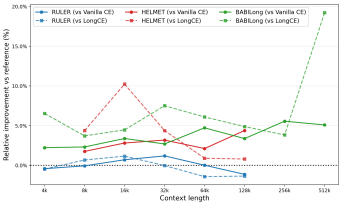
The difference $\Delta_i = \ell_i^{\mathrm{s}} - \ell_i^{\mathrm{f}}$ quantifies the change in prediction loss under this context restriction, with $\Delta_i > 0$ indicating a predictive benefit from context available only in the full view. Normalizing this contrast by the full-view loss yields the token-level SEGRATIO score:

$$r_i = \frac{\ell_i^{\mathrm{s}} - \ell_i^{\mathrm{f}}}{\ell_i^{\mathrm{f}}}. \tag{1}$$

A large positive r_i indicates a large relative benefit from the additional context, whereas $r_i \approx 0$ indicates similar prediction losses under the two views.



(a) 64K+SFT stage



(b) 512K+SFT stage

Figure 2: Length-wise relative gains on Llama-3-8B-Instruct.

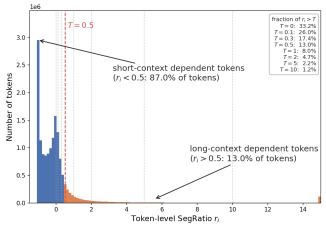


Figure 3: Token-level SEGRATIO distribution.

We compute token-level SEGRATIO scores for 360 ProLong-64K Gao et al. (2024) training samples with length≥16K using Llama-3-8B-Instruct and plot their distribution in Figure 3. The scores concentrate at low values, indicating that *tokens with a strong long-range dependency are sparse even in long and dependency-rich training documents*.

As an auxiliary diagnostic of whether r_i carries usable information, Section B aggregates it into a sequence-level score and probes its sensitivity to controlled document perturbations.

3.2 ROBUST TOKEN REWEIGHTING

Rather than treating SEGRATIO as a calibrated estimate of dependency strength, we use its rank within the eligible set to assign bounded additional weights. We therefore build LONGFOCUS with 2 stages: PreSelect controls which tokens are eligible, while Reweight controls the coverage and magnitude of the additional emphasis. This design reduces dependence on precise score calibration while controlling both the coverage of the correction and the maximum influence of spuriously high scores. The procedure is illustrated in Figure 1 and summarized in Algorithm 1.

PreSelect. Packed training inputs can contain documents of different lengths. To focus additional optimization emphasis on sufficiently long documents, we require the source document of a candidate token to contain at least G tokens. We also require $\ell_i^{\mathrm{f}} \geq \tau_{\mathrm{loss}}$: tokens below this floor are already well predicted under the full view, and their ratios can be dominated by small-denominator effects. Let $L_{\mathrm{doc}(x_i)}$ denote the length of the document containing token x_i. We can define $\mathcal{C} = \left\{ i : L_{\mathrm{doc}(i)} \geq G, \quad \ell_i^{\mathrm{f}} \geq \tau_{\mathrm{loss}} \right\}$. For each $i \in \mathcal{C}$, we compute r_i using Equation (1). Let k_i be the one-based descending rank of r_i within $\mathcal{C}$, with ties resolved by the sorting implementation. We retain only positive-score candidates within the top-ranked selection budget:

$$\mathcal{S}_b = \{ i \in \mathcal{C} : r_i > 0,\ k_i \leq \lfloor b|\mathcal{C}| \rfloor \}, \qquad b \in (0, 1]. \tag{2}$$

The budget allow at most $\lfloor b|\mathcal{C}| \rfloor$ tokens receive additional weights, and fewer are selected when positive-score candidates are insufficient. All tokens outside $\mathcal{S}_b$ retain unit weight.

Reweight. For each retained token $i \in \mathcal{S}_b$, define its rank percentile among the positive-score candidates as

$$p_i^+ = \frac{1}{|\mathcal{C}_+|} \sum_{j \in \mathcal{C}_+} \mathbf{1}[r_j \leq r_i] \in (0, 1]. \tag{3}$$

With maximum weight $c > 1$ and exponent $\gamma > 0$, we map this percentile to its loss weight:

$$w_i = 1 + (c - 1)(p_i^+)^\gamma \qquad i \in \mathcal{S}_b. \tag{4}$$

The percentile is normalized over $\mathcal{C}_+$, including candidates outside $\mathcal{S}_b$. Larger γ concentrates additional weight toward higher percentiles. If $\mathcal{S}_b$ is empty, all weights remain one.

Here, we use $\mathcal{C}_+$ as the ranking reference since it keeps each token's percentile relative to the positive-score population. For a fixed candidate pool, changing b therefore leaves the weights of common selected tokens unchanged, therefore recomputing percentiles within $\mathcal{S}_b$ would rescale them.

Training objective and computation. The final objective is

$$\mathcal{L}_{\mathrm{LONGFOCUS}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{sg}(w_i)\, \ell_i^{\mathrm{f}} = \mathcal{L}_{\mathrm{CE}} + \frac{1}{N} \sum_{i \in \mathcal{S}_b} \mathrm{sg}(w_i - 1)\, \ell_i^{\mathrm{f}},$$

Algorithm 1 LONGFOCUS loss computation for one training step

Require: Packed input; $\theta, S, G, \tau_{\text{loss}}, \epsilon, b, c, \gamma$
Compute aligned full-view and segmented-view loss entries
Initialize $w_i \leftarrow 1$ at all full-view objective positions
Form $\mathcal{C}$ using aligned finite entries, $L_{\text{doc}(i)} > G$, and $\ell_i^{\text{f}} \geq \tau_{\text{loss}}$
Compute r_i from detached losses using Equation (1)
Form $\mathcal{C}_+$ and $\mathcal{S}_b$ using Equation (2) ▷ PreSelect
if $|\mathcal{S}_b| > 0$ **then**
    Compute p_i^+ for $i \in \mathcal{S}_b$ using Equation (3)
    Assign w_i using Equation (4) ▷ Reweight
end if
$\mathcal{L} \leftarrow N_{\text{f}}^{-1} \sum_{i \in \mathcal{F}} \text{sg}(w_i) \ell_i^{\text{f}}$
Update θ using $\nabla_\theta \mathcal{L}$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\mathcal{L}_{\text{CE}} = N^{-1} \sum_i \ell_i^{\text{f}}$. On the implementation side, the full-view losses reuse the ordinary training computation, while the segmented scores require one additional reference forward, and gradients flow only through the full-view losses. This formulation separates weight allocation from the numerical scale of the coarse SEGRATIO score, reducing reliance on precise score calibration, while PreSelect improves the suitability of the candidates being ranked. By combining inexpensive SEGRATIO computation with filtered, rank-based reweighting, LONGFOCUS serve as a robust long context training objective with lower training time and peak memory than LongCE(see Table 2 and Section 4.2).

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Base models. We evaluate on three base architectures of increasing scale: (1) a dense 8B model, Llama-3-8B-Instruct (Grattafiori et al., 2024); (2) a 16B-parameter Mixture-of-Experts model with roughly 3B active parameters per token (MoE-16B-A3B); and (3) a scaled MoE-125B-A10B model.

Training pipeline. We use a two-stage 64K-to-512K mid-training curriculum, with subsets of the publicly released ProLong-64k and ProLong-512k datasets for the respective stages (Gao et al., 2024). At the end of each stage, we evaluate the mid-training checkpoint both before and after a lightweight SFT pass on UltraChat. This yields two mid-training endpoints, each with pre-SFT and post-SFT evaluations: 64K, 64K+SFT, 64K+512K and 64K+512K+SFT. We use document mask and sequence packing throughout the training. We hold the data mixture and training schedule fixed across all compared methods and vary only the training objective. Please refer to Section A for further details.

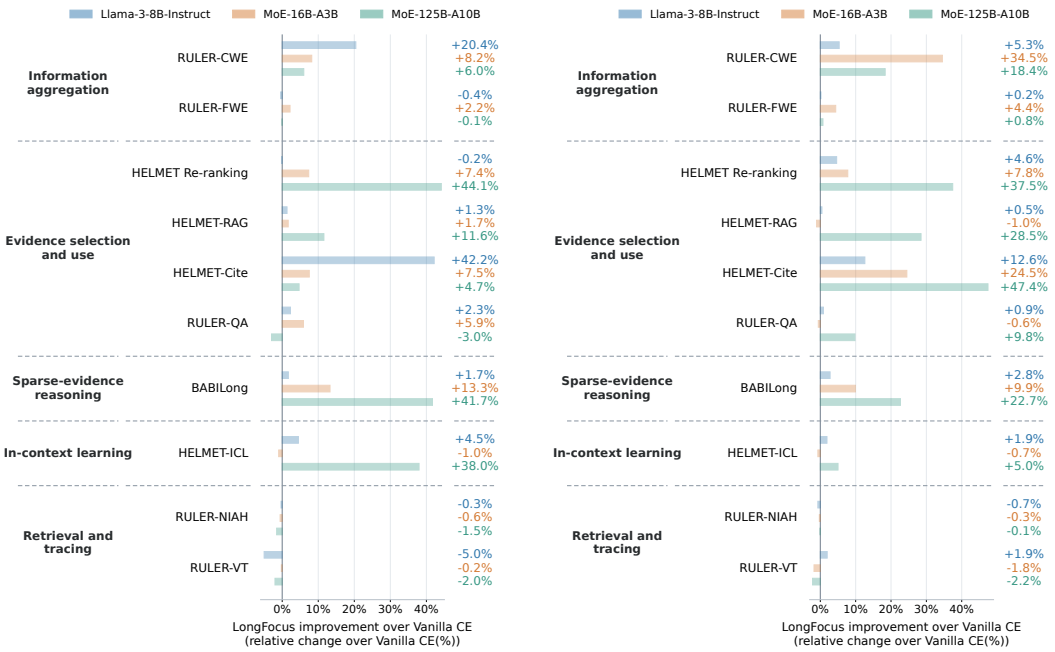
Compared methods. We compare three training objectives, applied on top of identical data and schedule within each stage: **Vanilla CE**: standard token-uniform cross-entropy loss; **LongCE** (Fang et al., 2025): configured with short-window length K=512, stride d=256, maximum weight cap 4.0; **LONGFOCUS (ours)**: using segment length $S = 512$, upweighting the top 50% of eligible tokens ranked by the SEGRATIO score (b=50%), with maximum weight cap c=4 and reweighting exponent γ=2. Preselect is configured with document-level length gate $G = 16K$, loss filter $\tau_{loss} = 0.1$. We vary the training objective only during mid-training; all SFT stages use standard Vanilla CE. We allocate the same GPU budget to all methods within each model and mid-training stage. We directly compare all three objectives on Llama-3-8B-Instruct. In the tested MoE configurations, LONGFOCUS remains feasible under the same GPU budgets for which the evaluated LongCE implementation runs out of memory. We therefore report MoE performance comparisons only against Vanilla CE.

Evaluation. We evaluate on three long-context suites: HELMET (Yen et al., 2025), RULER (Hsieh et al., 2024), BABILong (Kuratov et al., 2024), together with ten short-context benchmarks spanning knowledge and commonsense QA (ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al.,

Table 1: Main evaluation results on Llama-3-8B-Instruct, MoE-16B-A3B, and MoE-125B-A10B. Long-Avg averages HELMET and RULER before SFT, and additionally includes BABILong after SFT. Detailed benchmark results are in Section D. Parenthesized relative changes are measured against Vanilla CE.

Stage	Method	HELMET	RULER	BABILong	**Long-Avg**	**Short-Avg**
Llama-3-8B-Instruct						
64K	Vanilla CE	45.40	86.07	–	65.73	48.66
	LongCE	**47.80**	85.68	–	66.74 (+1.53%)	**49.03** (+0.76%)
	LONGFOCUS (ours)	47.34	**87.52**	–	**67.43** (+2.58%)	48.71 (+0.10%)
64K+SFT	Vanilla CE	47.96	87.76	48.79	61.50	53.14
	LongCE	47.45	86.91	47.98	60.78 (-1.18%)	53.41 (+0.51%)
	LONGFOCUS (ours)	**50.20**	**88.50**	**49.62**	**62.78** (+2.07%)	**53.62** (+0.90%)
64K+512K	Vanilla CE	45.24	88.75	–	67.00	48.11
	LongCE	**47.43**	88.44	–	**67.94** (+1.40%)	**48.73** (+1.28%)
	LONGFOCUS (ours)	46.43	**89.26**	–	67.85 (+1.27%)	48.41 (+0.62%)
64K+512K+SFT	Vanilla CE	48.43	89.31	45.47	61.07	52.36
	LongCE	47.71	**89.56**	44.23	60.50 (-0.93%)	**52.72** (+0.70%)
	LONGFOCUS (ours)	**49.77**	89.37	**46.73**	**61.96** (+1.45%)	52.40 (+0.08%)
MoE-16B-A3B						
64K	Vanilla CE	**35.41**	85.74	–	60.57	52.71
	LONGFOCUS (ours)	34.60	**87.02**	–	**60.81** (+0.39%)	**54.26** (+2.95%)
64K+SFT	Vanilla CE	43.41	87.40	29.53	53.45	**56.34**
	LONGFOCUS (ours)	**44.21**	**88.07**	**33.44**	**55.24** (+3.36%)	56.26 (-0.14%)
64K+512K	Vanilla CE	30.80	82.46	–	56.63	53.24
	LONGFOCUS (ours)	**34.46**	**85.58**	–	**60.02** (+5.98%)	**53.70** (+0.86%)
64K+512K+SFT	Vanilla CE	43.67	86.22	30.16	53.35	**55.97**
	LONGFOCUS (ours)	**44.23**	**87.25**	**33.15**	**54.88** (+2.86%)	55.73 (-0.42%)
MoE-125B-A10B						
64K	Vanilla CE	29.96	90.54	–	60.25	**55.92**
	LONGFOCUS (ours)	**31.71**	**90.90**	–	**61.31** (+1.75%)	55.69 (-0.42%)
64K+SFT	Vanilla CE	36.25	**91.20**	30.80	52.75	53.96
	LONGFOCUS (ours)	**46.32**	90.14	**43.65**	**60.03** (+13.80%)	**55.07** (+2.05%)
64K+512K	Vanilla CE	30.02	87.89	–	58.95	**55.44**
	LONGFOCUS (ours)	**31.31**	**89.18**	–	**60.25** (+2.20%)	55.29 (-0.26%)
64K+512K+SFT	Vanilla CE	39.97	88.28	32.12	53.46	52.38
	LONGFOCUS (ours)	**47.50**	**89.90**	**39.40**	**58.93** (+10.24%)	**55.85** (+6.63%)

2024), and WinoGrande (Sakaguchi et al., 2020)), math (GSM8K (Cobbe et al., 2021) and MATH-hard (Hendrycks et al., 2021b)), reasoning (BBH (Suzgun et al., 2023) and MuSR (Sprague et al., 2024)), and instruction-following (IFEval (Zhou et al., 2023)). At non-SFT stages (64K and 64K+512K), we report only HELMET and RULER, since checkpoints without an SFT pass often fail to follow the required answer format on BABILong; at SFT stages (64K+SFT and 64K+512K+SFT) we report all 3 suites. Correspondingly, *Long-Avg* is the unweighted mean of HELMET and RULER at the non-SFT stages, and the unweighted mean of all 3 suites at the SFT stages; *Short-Avg* is the



(a) 64K+SFT stage

(b) 512K+SFT stage

Figure 4: Long context capability improvements introduced by LONGFOCUS across models and training stages. Relative gains are computed as $100(s_{\text{ours}} - s_{\text{vanillaCE}})/s_{\text{vanillaCE}}$, where s denotes the corresponding task score.

mean of all ten short-context tasks. Specifically for HELMET, we report the average score of four of its sub task categories(RAG, ICL, Cite, and Re-rank). We exclude Recall since it substantially overlaps with RULER, and omit LongQA/Summ because their official category scores depend proprietary LLM judge, which is less reproducible. For the 64K/64K+SFT models, evaluation includes length buckets up to and including 64K. For the 64K+512K endpoints, RULER and HELMET extend to 128K, and post-SFT BABILong extends to 512K. We evaluate the last checkpoint of each stage and compare objectives at matched endpoints; Long-Avg values across stages are not directly comparable because the evaluated lengths, and before versus after SFT the included suites, differ.

4.2 MAIN RESULTS

Table 1 reports all 4 endpoints for each model. Following the post-SFT evaluation practice of Gao et al. (2024), we focus the analysis on the two post-SFT endpoints, with objective comparisons made separately at each endpoint. For detailed results on each benchmark please refer to Section D.

Improved downstream performance for both post-SFT endpoints over existing baselines. At 64K+SFT, LONGFOCUS improves Long-Avg over Vanilla CE on all three models. On MoE-125B-A10B, the score increases from 52.75 to 60.03, achieving a remarkable **7.28-point gain (13.8% relative)**. MoE-16B-A3B and Llama-3-8B-Instruct improve by 1.80 and 1.27 points, respectively. On Llama, where all three objectives can be compared directly, LONGFOCUS also exceeds LongCE by **2.00 points**. At 64K+512K+SFT, LONGFOCUS improves Long-Avg over Vanilla CE on all three models. MoE-125B-A10B increases from 53.46 to 58.93, a **5.48-point gain (10.24% relative)**; its HELMET, RULER, and BABILong scores improve by 7.53, 1.62, and 7.28 points, respectively. MoE-16B-A3B and Llama-3-8B-Instruct improve Long-Avg by 1.53 and 0.89 points, respectively. On Llama, it also exceeds LongCE by **1.46 points**, achieving higher Long-Avg than both baselines at each post-SFT endpoint. Before SFT, LONGFOCUS also improves Long-Avg over Vanilla CE in all six model-stage comparisons. The corresponding MoE-125B-A10B gains are 1.05 points at 64K and 1.30 points at 64K+512K. The larger gains after SFT are consistent with prior observations that instruction tuning can better reveal long-context capabilities acquired during continued training

Table 2: Training efficiency of different methods. Savings rows report reductions of LONGFOCUS relative to LongCE. See Section A for evaluation details.

Model	Context	Method	Time/step (s)	Peak Mem/GPU (MB)
Llama-3-8B-Instruct	64K	Vanilla CE	12.660 ± 0.010	22506 ± 0
		LongCE	21.840 ± 0.032 (1.73×)	58939 ± 1 (+161.9%)
		LONGFOCUS	17.424 ± 0.014 (1.38×)	40686 ± 0 (+80.8%)
		Savings vs. LongCE	**20.2%**	**31.0%**
	512K	Vanilla CE	35.460 ± 0.015	17258 ± 0
		LongCE	43.399 ± 0.004 (1.22×)	43123 ± 0 (+149.9%)
		LONGFOCUS	39.544 ± 0.020 (1.12×)	30191 ± 0 (+74.9%)
		Savings vs. LongCE	**8.9%**	**30.0%**

that are less apparent in pre-SFT evaluation (Gao et al., 2024). Against LongCE, Llama gains 0.69 points at 64K but is 0.09 points lower at 64K+512K.

Task-wise gains concentrate in evidence use and long range reasoning. The aggregate improvements vary across task types (Figure 4). At 64K+SFT, MoE-125B-A10B gains 10.06 points on HELMET and 12.84 points on BABILong, while RULER decreases by 1.06 points. Its largest relative gains include HELMET Re-ranking (44.1%), BABILong (41.7%), and HELMET-ICL (38.0%). At 64K+512K+SFT, the complete MoE-125B-A10B results show remarkable gains on HELMET Re-ranking (37.5%), RAG (28.5%), Cite (47.4%), BABILong (22.7%), and RULER-CWE (18.4%). All three models improve on RULER-CWE, HELMET-Cite, and BABILong at both post-SFT endpoints. These shared gains coexist with slight trade-offs: RULER-NIAH decreases slightly in all six model-endpoint pairs, and MoE-125B-A10B's RULER-VT decreases by 2.2% at the 512K endpoint. The results support that improvements concentrate in selected aggregation, evidence-use, and reasoning tasks, while retrieval tasks remain mixed and often saturated.

Length-wise benefits extend to 512K, especially on BABILong. Figure 2 shows Llama-3-8B-Instruct at matched post-SFT endpoints. At 64K+SFT, HELMET exceeds both Vanilla CE and LongCE at every plotted length from 8K to 64K. RULER varies with length, with its largest gains at 64K: approximately 4.1% over Vanilla CE and 13.2% over LongCE. BABILong improves over Vanilla CE from 4K to 32K but is slightly lower at 64K (43.4 versus 43.6). At 64K+512K+SFT, HELMET exceeds both baselines throughout 8K–128K, and BABILong exceeds both throughout the plotted 4K–512K range. Specifically at 512K, BABILong scores 24.8 versus 23.6 for Vanilla CE and 20.8 for LongCE, corresponding to relative gains of 5.1% and 19.2%. RULER remains mixed at this endpoint.

Lower online weighting overhead and practical scalability. An online weighting rule must be affordable throughout mid-training, so its scoring cost is part of the design objective rather than a separate implementation concern. Table 2 measures training overhead separately at each mid-training length. Relative to LongCE on Llama-3-8B-Instruct, LONGFOCUS reduces per-step training time and peak memory by 20.2% and 31.0% at 64K, and by 8.9% and 30.0% at 512K, respectively. In the tested MoE-16B-A3B and MoE-125B-A10B configurations, the evaluated LongCE implementation encounters CUDA out-of-memory errors, whereas LONGFOCUS remain feasible under the same GPU budgets (Section 4.1), demonstrating its practicality.

Short-context performance relative to matched baselines. On Llama-3-8B-Instruct and MoE-16B-A3B, post-SFT Short-Avg remains close to the corresponding Vanilla CE score. At 64K+SFT, the differences are $+0.48$ and -0.08 points, respectively; at 512K+SFT, they are $+0.04$ and -0.23 points. Llama's 512K+SFT score is also 0.32 points below LongCE. On MoE-125B-A10B, post-SFT Short-Avg increases from 53.96 to 55.07 at 64K and from 52.38 to 55.85 after the 512K continuation, gains of 1.11 and 3.47 points. GSM8K scores use the corrected stopping and answer-extraction protocol described in Table 26. Detailed task scores are provided in Section D.

We also provide further analysis on the reweight geometry and a case study in Sections C and E.

Table 3: Hyperparameter ablations on Llama-3-8B-Instruct. Δ_{CE} denotes the absolute Long-Avg gain over Vanilla CE. Detailed benchmark-level results are provided in Table 17.

	64K		64K+SFT	
Variant	Long-Avg ↑	Δ_{CE}	Long-Avg ↑	Δ_{CE}
Vanilla CE	65.73	–	61.50	–
LONGFOCUS (default)	67.43	+1.70	62.78	+1.27
$b = 10\%$	66.98	+1.25	62.70	+1.20
$b = 70\%$	67.45	+1.71	62.77	+1.27
$S = 128$	67.34	+1.61	62.41	+0.91
$S = 2048$	67.50	+1.77	**62.78**	**+1.28**
$S = 8192$	66.53	+0.80	61.71	+0.20
$\gamma = 0.5$	**67.66**	**+1.93**	62.17	+0.66
$\gamma = 1$	66.67	+0.94	62.36	+0.85
w/o PreSelect Document Gate G	66.90	+1.17	62.38	+0.88

4.3 HYPER-PARAMETER ABLATION

Table 3 evaluates the reweighting fraction b, segment-view length S, reweighting exponent γ, and the PreSelect gate. These comparisons assess how reference construction, candidate restriction, and weight allocation affect the training utility of the coarse contrast. All tested LONGFOCUS variants outperform Vanilla CE in Long-Avg at both stages, showing that the observed gains are not confined to the default configuration.

For $b \in \{10\%, 50\%, 70\%\}$, Long-Avg remains rather steady. In particular, assigning additional weights to only the top 10% of eligible tokens achieves 62.70 after SFT, just 0.07 points below the default, indicating that *competitive performance does not require upweighting most eligible tokens*. For $S \in \{128, 512, 2048\}$, Long-Avg also remains steady, ranging from 67.34 to 67.50 at 64K and from 62.41 to 62.78 after SFT. However, increasing S to 8192 lowers Long-Avg by 0.90 and 1.07 points relative to the default. These results show that *excessive long local references may cause performance degradation*. As for weight concentration and candidate filtering, the preferred exponent is stage-dependent: $\gamma = 0.5$ achieves the highest 64K score of 67.66, whereas the default $\gamma = 2$ performs best among the tested exponents after SFT, exceeding $\gamma = 0.5$ and $\gamma = 1$ by 0.61 and 0.42 points, respectively. Removing the PreSelect document-length gate reduces Long-Avg by 0.53 points at 64K and 0.40 points after SFT. These results support the default exponent for the evaluated post-SFT setting and show an additional benefit from candidate filtering before rank-based reweighting.

5 CONCLUSIONS

In this work, we investigate how inexpensive, coarse context contrasts can support effective token reweighting for online long-context mid-training. LONGFOCUS combines an inexpensive segmented reference with PreSelect filtering and bounded rank-based reweighting to *turn the coarse* SEGRATIO *score into useful additional training emphasis*. Across dense and MoE architectures and training contexts from 64K to 512K, LONGFOCUS improves Long-Avg over Vanilla CE in all matched comparisons. MoE-125B-A10B gains 13.80% and 10.24% at the two post-SFT endpoints, respectively. On Llama-3-8B-Instruct, it exceeds LongCE at both post-SFT endpoints while reducing per-step training time and peak memory. Taken together, these results support co-designing inexpensive context contrasts and their weight-allocation rules as a practical route to effective online mid-training in the evaluated settings.

6 AI USE STATEMENT

Generative AI tools were used solely to improve the readability of the manuscript and to review portions of the code and algorithm implementations for potential errors; they were not involved in the design of the proposed method, and all AI-assisted outputs were independently verified by the authors.

REFERENCES

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

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. LADM: Long-context training data selection with attention-based dependency measurement for llms. *arXiv preprint arXiv:2503.02502*, 2025.

Longze Chen, Ziqiang Liu, Wanwei He, Yunshui Li, Run Luo, and Min Yang. Long context is not long at all: A prospector of long-dependency data for large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.

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

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. In *International Conference on Learning Representations (ICLR)*, 2024b.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Haoran Deng, Yingyu Lin, Zhenghao Lin, Xiao Liu, Yizhou Sun, Yi-An Ma, and Yeyun Gong. Beyond length: Quantifying long-range information for long-context LLM pretraining data. *arXiv preprint arXiv:2510.25804*, 2025.

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. In *International Conference on Machine Learning (ICML)*, 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? In *International Conference on Learning Representations (ICLR)*, 2025.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. In *International Conference on Machine Learning (ICML)*, 2024.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2024.

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

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Falko Helm, Nico Daheim, and Iryna Gurevych. Token weighting for long-range language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1440–1459. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.79. URL https://aclanthology.org/2025.findings-naacl.79/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021b.

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

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. DeepSpeed Ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 106519–106554. Curran Associates, Inc., 2024. doi: 10.52202/079017-3381. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c0d62e70dbc659cc9bd44cbcf1cb652f-Paper-Datasets_and_Benchmarks_Track.pdf.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 29029–29063. Curran Associates, Inc., 2024. doi: 10.52202/079017-0914. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3322a9a72a1707de14badd5e552ff466-Paper-Conference.pdf.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. In *International Conference on Learning Representations (ICLR)*, 2024.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Shuzheng Si, Haozhe Zhao, Gang Chen, et al. GATEAU: Selecting influential samples for long context alignment. *arXiv preprint arXiv:2410.15633*, 2025.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2023.

Junfeng Tian, Da Zheng, Yang Cheng, Rui Wang, Colin Zhang, and Debing Zhang. Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models. *arXiv preprint arXiv:2409.04774*, 2024. URL https://arxiv.org/abs/2409.04774.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Longyun Wu, Dawei Zhu, Guangxiang Zhao, et al. LongAttn: Selecting long-context training data via token-level attention. *arXiv preprint arXiv:2502.16860*, 2025.

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

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

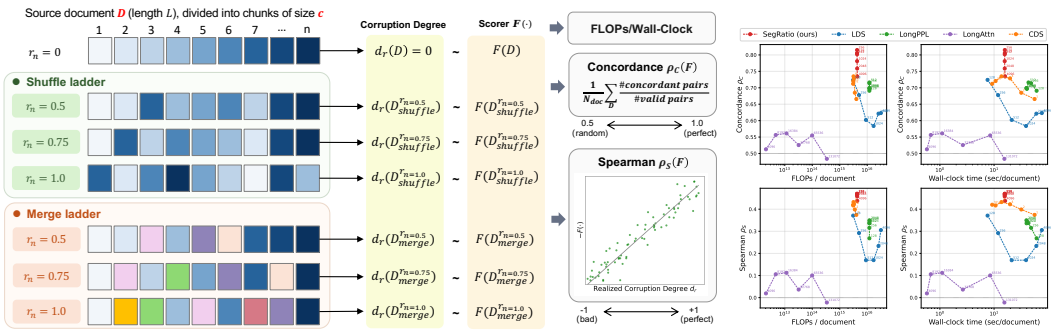
An Yang, Bowen Yu, Chengyuan Li, et al. Qwen2.5-1M technical report. *arXiv preprint arXiv:2501.15383*, 2025. URL https://arxiv.org/abs/2501.15383.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Manzil Zaheer, Guru Guruganesh, 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 (NeurIPS)*, 2020.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.



(a) (b)

Figure 5: (a) Overview of the corruption-ladder-based ranking task suite. (b) Overall cost-accuracy curves for all five evaluated scorers, averaged over L and c.

A Detailed Experiment Settings

Detailed Training Configurations. All Llama runs use `adamw_torch` (β_1=0.9, β_2=0.95, weight decay 0.1, max grad norm 1.0), bf16 precision and a cosine learning-rate schedule. Mid-train stages use lr=10^{-5} (min-lr ratio 0.1, warmup ratio 0.1); SFT stages use lr=2×10^{-5} (warmup ratio 0.05). RoPE base is 8×10^6 at 64K and 1.28×10^8 at 512K (inherited unchanged through the following SFT stage in each branch). Detailed reproduction details are listed in our codebase. Our MoE model is a DeepSeek-V2-style Mixture-of-Experts model (the first layer is dense; the remaining layers are MoE). For propriety reasons, we cannot disclose detailed training configurations. All experiments are run on the same midtraining token budget. We also apply global gradient-norm clipping (max norm 1.0) after gradient accumulation, limiting increases in gradient magnitude induced by token reweighting before the AdamW update.

Efficiency Evaluation. In Table 2, we conduct a controlled efficiency probe that isolates the per-step computational cost of each training. We run methods for 20 steps from the same base checkpoint on an identical GPU budget, matching sequence-parallel degree to each stage's production configuration. We report the median sec/step over three repetitions (steps 5-20, discarding warmup) and the maximum of a per-step, all-reduced exact peak-memory.

B Corruption Ladder Based Ranking Task Suite for Long Context Dependency Scoring Evaluation

B.1 Validating SegRatio on the sequence-level.

Aggregating this coarse signal with noise filtering. To examine whether the coarse construction of LongFocus retains useful dependency information, we aggregate token scores after excluding tokens with negligible full-view loss, since their small denominators can amplify minor loss differences. We use a positive loss floor τ_{loss} to determine which scores participate in subsequent signal aggregation or reweighting.

Let $\mathcal{E}(\mathbf{x}) = \{i : \ell_i^{\text{f}} \geq \tau_{\text{loss}}\}$. For the document-ranking evaluation, let $\mathcal{E}_{\text{score}}(\mathbf{x})$ contain positions with $\ell_i^{\text{f}} > \tau_{\text{loss}}$ and aligned full- and segmented-view prediction losses. The scoring implementation excludes the first token of each segment and any incomplete trailing segment. For nonempty $\mathcal{E}_{\text{score}}$, it aggregates the contrast as a ratio of summed losses, equivalently a full-view-loss-weighted average of token scores:

$$R(\mathbf{x}) = \frac{\sum_{i\in\mathcal{E}_{\text{score}}(\mathbf{x})}(\ell_i^{\text{s}} - \ell_i^{\text{f}})}{\sum_{i\in\mathcal{E}_{\text{score}}(\mathbf{x})}\ell_i^{\text{f}}} = \frac{\sum_{i\in\mathcal{E}_{\text{score}}(\mathbf{x})}\ell_i^{\text{f}} r_i}{\sum_{i\in\mathcal{E}_{\text{score}}(\mathbf{x})}\ell_i^{\text{f}}}. \tag{5}$$

The ranking suite uses $\tau_{\text{loss}} = 0.1$. Under a fixed scoring model and segmentation setting, this filtered aggregate achieves strong ranking performance on documents with different degrees of controlled document corruption(Section B). The result provides evidence that the coarse contrast retains useful information for dependency discrimination after filtering and aggregation. This observation

motivates using SEGRATIO for training while explicitly controlling which tokens are reweighted and how their scores determine the additional emphasis.

B.2 TASK CONSTRUCTION

As detailed in Section 2, a broad family of off-line scoring functions has been proposed to quantify the long-range dependency of a sequence, and these scores are routinely used in document-selection pipelines to prioritize data for long-context training. However, the *accuracy* of these scorers is typically validated only indirectly, through the downstream performance of models trained on the selected data, which makes it difficult to compare competing methods on a fair and direct basis. This issue is compounded by *computational cost*: since even a single forward pass over a long document is expensive, a scorer that offers only marginal accuracy gains over a significantly cheaper alternative is rarely worth the overhead, particularly at pretraining or mid-training scales. Thus, *a comprehensive evaluation that jointly accounts for scoring accuracy and computational efficiency is needed.*

To fill this gap, we construct a **corruption-ladder based ranking task suite** (see Figure 5a) to assess long-context scorers along both dimensions. Specifically, we perturb open-source long documents by reordering chunks or replacing them with unrelated content. The resulting corruption levels provide a controlled proxy for disrupted document structure, rather than direct ground-truth measurements of long-range dependency strength. For scoring accuracy, we measure **concordance**: how reliably a candidate scorer ranks these documents by their actual degree of corruption. For efficiency, we measure the **compute cost** required to produce each score, including both theoretical FLOPs and actual wall-clock time.

Using this ladder, we evaluate the sequence-level score $R(\mathbf{x})$ defined in Equation (5) against baseline scorers. This evaluation serves as an auxiliary diagnostic of the aggregated score: it measures sensitivity to disrupted document structure, and does not directly establish token-level ranking accuracy or explain the training gains. What it does test is whether the aggregate score assigns higher values to less corrupted variants of the same document. This enables us to compare our aggregated score with existing baselines on strict accuracy-per-compute grounds, rather than relying solely on expensive end-to-end training results.

Source corpus. We use PG19 (Rae et al., 2020) as our source corpus, sampling $N_{\text{doc}} = 100$ eligible books and taking a prefix at each nominal length tier $L \in \{32\text{K}, 64\text{K}, 128\text{K}\}$, yielding 300 document-tier samples; samples across tiers need not come from distinct books. Each document is partitioned into non-overlapping chunks of size c tokens. We test $c \in \{128, 512, 4096\}$ in ablation studies to examine sensitivity to chunk granularity. The construction tokenizer is OPT-350M's tokenizer. Scorers re-tokenize the resulting text using the scoring model's tokenizer and apply their configured length caps; thus L and c denote construction lengths, not exact token counts under the scoring model.

Corruption ladders. For each document D, we construct two independent corruption ladders under nominal corruption ratios $r_n \in \{0, 0.5, 0.75, 1.0\}$.

- **Shuffle ladder (reorder chunks).** Given a ratio r_n, we randomly select $k = \text{round}(r_n \cdot L/c)$ chunk positions and apply a *derangement* (a cyclic shift with non-zero offset) exclusively to these selected positions, ensuring every selected chunk moves to a different position while unselected chunks remain fixed. Since the actual displacement varies due to random selection, we quantify the realized corruption degree d_r by the Kendall-τ distance($\tau \in [0, 1]$) between the shuffled and original chunk sequences, rather than using the nominal ratio r_n directly.

- **Merge ladder (replace chunk contents).** Given the same ratio r_n, we select k chunk positions and replace the content of each selected chunk with an equal-length excerpt randomly drawn from a different document. In this mode, the realized corruption degree is $d_r = k/(L/c)$, which equals r_n for the length, chunk-size, and ratio combinations evaluated here.

Table 4: Resolution parameters and approximate analytic FLOPs for each evaluated method. $n_c = \lfloor L/s \rfloor$, H is hidden size, P_ℓ is the parameter count per transformer layer, and N_ℓ is model depth. $\mathrm{FLOPs}_{\mathrm{body}}$ excludes the vocabulary projection. CDS uses fixed attention chunk size $a = 128$, distinct from corruption chunk size c, with $d_{\mathrm{afs}} = 4$.

Method	Evaluated resolution parameters	**FLOPs**(L)
SEGRATIO	segment size $s = \{128, 256, 512, 1024, 2048, 4096\}$	$\mathrm{FLOPs}(L) + n_c\,\mathrm{FLOPs}(s)$
LDS	chunk size $s = \{128, 256, 512, 1024, 2048, 4096\}$	$n_c\,\mathrm{FLOPs}(s) + n_{\text{pair}}\,\mathrm{FLOPs}(2s)$ $n_{\text{pair}} = \min(500, \binom{n_c}{2})$
LongPPL	truncation length $s = \{128, 256, 512, 1024, 2048, 4096\}$, $w = \max(1, \lfloor s/4 \rfloor)$	$2\,\mathrm{FLOPs}(L) + n_{\text{step}}\,\mathrm{FLOPs}(s+w)$ $n_{\text{step}} = \lvert\{0, w, 2w, \dots\} \cap [0, L-s)\rvert$
LongAttn	max length $M = \{4096, 8192, 16384, 32768, 65536, 131072\}$, filtered to $M \le L$	$2P_\ell M + 4HM^2$
CDS	stride multiplier $d_{\mathrm{cds}} = \{1, 2, 4, 8, 16, 32\}$, $\mathrm{st} = a \cdot d_{\mathrm{cds}}$, $a = 128$	$\mathrm{FLOPs}_{\mathrm{body}}(L) + N_\ell \cdot \dfrac{H\,a\,L^2}{a \cdot d_{\mathrm{cds}}}$

For each base document D, each ladder contains four variants corresponding to the four r_n values. The ladders share the clean variant, yielding seven distinct conditions per document and corruption chunk size. For every variant, we record its realized corruption degree $d_r \in [0, 1]$.

Metrics. For each fixed method, resolution parameter, length tier, corruption chunk size, and ladder mode, we evaluate the four variants of each of 100 documents. We calibrate the score direction within this group: $q = +1$ if the mean score at nominal corruption $r_n = 1$ exceeds the clean mean, and $q = -1$ otherwise. **Concordance accuracy** ρ_C pools the valid within-document pairs across the group. Pairs with equal realized corruption or an undefined score are excluded. For each remaining pair with $d_a < d_b$, we assign credit 1, $1/2$, or 0 according to whether $q(F_b - F_a)$ is positive, zero, or negative. The mean credit is ρ_C; 0.5 is the random-order reference for a fixed direction. Because the direction is calibrated on the evaluated group itself, these results describe direction-calibrated discrimination, not performance with a direction calibrated on held-out data. For SEGRATIO, all evaluated groups select $q = -1$, consistently associating higher scores with less corruption. We also report **Spearman correlation** ρ_S between realized corruption and the direction-adjusted scores, pooling valid variants across documents within each group. LongPPL produces no key tokens for 23 of its 37,800 scored variants; their stored zero values are failure sentinels, so we exclude those scores and their associated comparisons rather than treating them as valid perplexities. All 522 book-domain evaluation groups remain represented.

Alongside scoring accuracy, we report **analytic FLOPs**, estimated at the nominal scoring length using the call schedules in Table 4, and **wall-clock time** per document. Each recorded duration covers one complete 100-document scoring process, including model loading, scoring, and I/O. We divide that duration by 100 before averaging over conditions and configurations. These times depend on the implementation and hardware; they are not isolated model-forward latencies. Analytic FLOPs use nominal length caps rather than the varying re-tokenized lengths.

Evaluated Methods. We evaluate the following long-context scoring methods: SEGRATIO (ours), LDS (Chen et al., 2024a), LongPPL (Fang et al., 2025), LongAttn (Wu et al., 2025), and CDS (Chen et al., 2025) as summarized in Table 4. For a vocabulary size V, we approximate a causal transformer forward by $\mathrm{FLOPs}(T) = 2(N_\ell P_\ell + HV)T + 2N_\ell HT^2$, and define $\mathrm{FLOPs}_{\mathrm{body}}(T) = \mathrm{FLOPs}(T) - 2HVT$. CDS and LongAttn do not execute a vocabulary projection. LongAttn explicitly computes dense query–key and attention–value products, giving the $4HM^2$ term for its single layer. These estimates count matrix multiplications and omit elementwise operations. For each method, we sweep its **resolution parameter**—a hyperparameter that controls the granularity of the score thereby trading off evaluation accuracy against cost as detailed in Table 4. Sweeping this parameter traces out the cost-accuracy curves for each method. We use the same backbone checkpoint, Llama-3-8B-ProLong-512k-InstructGao et al. (2024)(`https://huggingface.co/princeton-nlp/Llama-3-8B-ProLong-512k-Instruct`), for all scoring methods. LongAttn uses only its first layer. The evaluated SEGRATIO score is the loss-weighted sequence aggregate in Equation (5), with a loss floor of 0.1 nats.

Evaluation Setup. To systematically assess the robustness of each scoring method, we conduct experiments across all combinations of document length tiers $L \in \{32\text{K}, 64\text{K}, 128\text{K}\}$ and chunk sizes $c \in \{128, 512, 4096\}$. For each (L, c) configuration, we compute ρ_C, ρ_S, analytic FLOPs, and wall-clock time over both ladder modes and report the average value. For each evaluated method, we sweep its resolution parameter over the value sets listed in Table 4, as this parameter directly governs the trade-off between accuracy and computational cost, enabling the construction of cost-accuracy curves for a comprehensive view. We use unweighted means at each aggregation level. Each detailed-table tier cell averages the two ladder modes, and its Overall cell averages the feasible tier means. The summary table first averages over modes, corruption granularities, and feasible resolution settings within each tier, then averages the three tier means equally. All numbers are computed from unrounded group statistics and rounded only for presentation.

B.3 Evaluation Results

Table 5: Direction-calibrated concordance ρ_C and Spearman correlation ρ_S. We average over ladder mode, corruption chunk size c, and feasible resolution settings within each length tier, then average the three tier means equally for Overall. Tied scores receive half credit; undefined scores are excluded. LongAttn has four, five, and six feasible settings at 32K, 64K, and 128K, respectively.

	Overall		$L = 32\mathbf{K}$		$L = 64\mathbf{K}$		$L = 128\mathbf{K}$	
Method	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S
SegRatio (ours)	**0.783**	**0.461**	**0.772**	**0.453**	**0.785**	**0.453**	**0.792**	**0.477**
CDS	0.708	0.411	0.707	0.416	0.676	0.368	0.741	0.448
LongPPL	0.703	0.326	0.680	0.298	0.703	0.329	0.725	0.350
LDS	0.639	0.257	0.619	0.227	0.626	0.233	0.671	0.310
LongAttn	0.537	0.065	0.543	0.070	0.522	0.038	0.547	0.089

Table 5 reports, for each method, both the concordance accuracy ρ_C and the Spearman rank correlation ρ_S averaged over $c \in \{128, 512, 4096\}$ and the evaluated resolution parameter values, separately at each length tier $L \in \{32\text{K}, 64\text{K}, 128\text{K}\}$. Figure 5b reports both the scoring accuracy and the compute cost on a 2-D graph to demonstrate the accuracy-cost trade-off. Tables 8 to 12 report the complete results across the evaluated configurations. Table 13 reports mean wall-clock time per document and estimated analytic FLOPs.

Overall Results. Across the evaluated configurations, SegRatio achieves the highest ranking accuracy, exceeding the strongest baseline by 7.5 percentage points (Table 5), with a favorable accuracy-cost trade-off (Figure 5b). The strong document-level discrimination supports our design of using filtering and aggregation to obtain a reliable signal from the coarse full-versus-segmented contrast, providing an empirical basis for its use in LongFocus. Suite construction, the evaluation protocol, and detailed results are provided in Section B.

SegRatio shows strong ranking accuracy across corruption scales and document lengths with favorable accuracy-cost trade-off. SegRatio achieves the highest grid-averaged concordance (0.783) and Spearman correlation (0.461), exceeding the strongest baseline by 7.5 percentage points in concordance (Table 5). With segment size $s = 256$, it also outperforms every baseline's best setting at each corruption granularity after averaging over length tiers and ladder modes (Tables 7 and 8). Across length tiers, its grid-averaged concordance remains stable; LongPPL and LDS improve more with length, whereas CDS and LongAttn show non-monotonic changes (Table 5). The two-dimensional plots in Figure 5b further show that SegRatio combines the highest ranking accuracy with moderate computation, highlighting its favorable balance between ranking quality and scoring cost. At their respective best-concordance settings, LongPPL requires $2.81\times$ the FLOPs and $2.74\times$ the wall-clock time of SegRatio (Tables 10 and 13).

Corruption granularity affects scorers differently. At fixed document length, increasing c produces fewer, larger chunks: each perturbation preserves longer contiguous spans, so it changes the structure of the ranking task rather than simply increasing its difficulty. As c grows from 128 to 4096, LongPPL's average concordance drops from 0.765 to 0.614, whereas LDS performs best at the intermediate granularity $c = 512$ (Tables 9 and 10). CDS also weakens at the coarsest granularity, while LongAttn remains a weak discriminator across granularities; SegRatio retains the highest

Table 6: Concordance ρ_C separated by corruption mode. At each c, we average over feasible resolution settings within each length tier, then average the three tiers equally. Overall additionally averages the three corruption granularities; modes remain separate. Before rounding, the mean of the two Overall columns equals the concordance in Table 5. Bold denotes the highest value in each column.

	Overall		$c = 128$		$c = 512$		$c = 4096$	
Method	Shuffle	Merge	Shuffle	Merge	Shuffle	Merge	Shuffle	Merge
SEGRATIO (ours)	**0.604**	**0.962**	0.547	**0.995**	**0.596**	**0.993**	**0.669**	**0.897**
CDS	0.477	0.939	0.468	0.963	0.489	0.969	0.473	0.885
LongPPL	0.573	0.832	**0.599**	0.930	0.563	0.896	0.558	0.670
LDS	0.530	0.747	0.507	0.670	0.554	0.874	0.529	0.697
LongAttn	0.536	0.539	0.544	0.549	0.531	0.541	0.533	0.526

Table 7: For each corruption chunk size c, the resolution setting maximizing concordance ρ_C, averaged equally over ladder modes and feasible length tiers; concordance is given in parentheses. LongAttn settings with $M > 32768$ cover fewer tiers.

c	**SEGRATIO (ours)**	**CDS**	**LongPPL**	**LDS**	**LongAttn**
128	s=128 (0.828)	d_{cds}=8 (0.755)	s=128 (0.799)	s=2048 (0.639)	M=16384 (0.576)
512	s=256 (0.841)	d_{cds}=4 (0.763)	s=256 (0.764)	s=128 (0.784)	M=65536 (0.570)
4096	s=1024 (0.807)	d_{cds}=4 (0.706)	s=4096 (0.677)	s=128 (0.764)	M=8192 (0.551)

average concordance at all three, with a narrower range of 0.771–0.794 (Figure 6 and Tables 8, 11 and 12).

Discrimination is stronger for content replacement than for reordering. Table 6 separates the two corruption modes: SEGRATIO achieves the highest overall concordance in both, but performs substantially better on merge (0.962) than on shuffle (0.604). Its advantage is not uniform across all mode–granularity combinations: LongPPL ranks higher on shuffle at $c = 128$. As c increases, SEGRATIO's shuffle concordance rises from 0.547 to 0.669, while its merge concordance falls from 0.995 to 0.897. Thus, the relatively stable combined average masks different responses to the two perturbation types.

Table 8: Full results for SEGRATIO (ours): direction-calibrated concordance ρ_C and Spearman correlation ρ_S for every (corruption chunk size c, segment size s) cell. Each tier averages the shuffle and merge ladders; Overall averages the tier means. See Section B.3 for analysis.

c	s	**Overall**		$L = 32$**K**		$L = 64$**K**		$L = 128$**K**	
		ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S
128	128	0.828	0.541	0.833	0.559	0.831	0.546	0.820	0.517
128	256	0.824	0.514	0.818	0.528	0.818	0.517	0.837	0.498
128	512	0.777	0.489	0.765	0.501	0.782	0.490	0.784	0.476
128	1024	0.747	0.467	0.731	0.473	0.748	0.467	0.763	0.460
128	2048	0.732	0.448	0.724	0.448	0.736	0.449	0.737	0.448
128	4096	0.718	0.420	0.693	0.403	0.723	0.421	0.738	0.436
512	128	0.831	0.520	0.834	0.501	0.833	0.521	0.826	0.538
512	256	0.841	0.516	0.839	0.495	0.848	0.519	0.836	0.533
512	512	0.838	0.504	0.830	0.481	0.844	0.508	0.838	0.521
512	1024	0.789	0.482	0.763	0.457	0.798	0.490	0.808	0.500
512	2048	0.747	0.463	0.731	0.440	0.754	0.470	0.755	0.480
512	4096	0.721	0.437	0.709	0.407	0.723	0.443	0.733	0.460
4096	128	0.759	0.346	0.757	0.346	0.751	0.303	0.768	0.390
4096	256	0.780	0.380	0.758	0.370	0.784	0.346	0.797	0.424
4096	512	0.791	0.414	0.789	0.408	0.788	0.385	0.796	0.450
4096	1024	0.807	0.442	0.793	0.433	0.809	0.416	0.819	0.475
4096	2048	0.797	0.461	0.778	0.457	0.796	0.437	0.817	0.491
4096	4096	0.764	0.456	0.754	0.452	0.760	0.432	0.778	0.485

Table 9: Full results for LDS: concordance accuracy ρ_C and Spearman correlation ρ_S for every (corruption chunk size c, resolution parameter s) cell, averaged over the shuffle and merge corruption ladders, overall and broken down by length tier L. see Section B.3 for summary statistics derived from this grid.

c	s	**Overall**		$L = 32$**K**		$L = 64$**K**		$L = 128$**K**	
		ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S
128	128	0.625	0.219	0.643	0.241	0.623	0.205	0.610	0.210
128	256	0.563	0.122	0.555	0.108	0.531	0.050	0.603	0.208
128	512	0.508	0.025	0.481	-0.024	0.439	-0.089	0.603	0.188
128	1024	0.578	0.180	0.561	0.153	0.550	0.143	0.623	0.243
128	2048	0.639	0.292	0.629	0.288	0.643	0.300	0.645	0.288
128	4096	0.618	0.312	0.594	0.279	0.620	0.317	0.639	0.342
512	128	0.784	0.457	0.780	0.481	0.780	0.431	0.791	0.457
512	256	0.773	0.424	0.766	0.427	0.767	0.423	0.788	0.423
512	512	0.716	0.350	0.629	0.274	0.749	0.373	0.770	0.404
512	1024	0.685	0.337	0.674	0.322	0.742	0.369	0.640	0.322
512	2048	0.698	0.376	0.711	0.401	0.673	0.353	0.710	0.375
512	4096	0.630	0.330	0.600	0.279	0.635	0.332	0.655	0.379
4096	128	0.764	0.438	0.755	0.411	0.771	0.448	0.765	0.455
4096	256	0.697	0.330	0.652	0.275	0.648	0.289	0.791	0.425
4096	512	0.583	0.132	0.519	0.030	0.546	0.078	0.683	0.288
4096	1024	0.488	-0.008	0.457	-0.093	0.475	-0.048	0.533	0.115
4096	2048	0.526	0.033	0.522	-0.028	0.498	0.003	0.559	0.124
4096	4096	0.621	0.274	0.613	0.266	0.586	0.220	0.665	0.337

Table 10: Full results for LongPPL: concordance accuracy ρ_C and Spearman correlation ρ_S for every (corruption chunk size c, resolution parameter s) cell, averaged over the shuffle and merge corruption ladders, overall and broken down by length tier L. See Section B.3 for summary statistics derived from this grid.

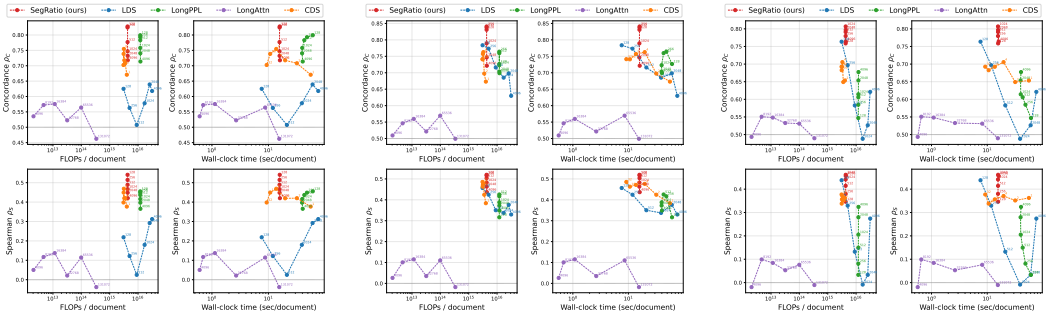
c	s	**Overall**		$L = 32$**K**		$L = 64$**K**		$L = 128$**K**	
		ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S
128	128	0.799	0.455	0.787	0.444	0.806	0.474	0.805	0.448
128	256	0.793	0.449	0.778	0.429	0.797	0.470	0.804	0.446
128	512	0.783	0.437	0.771	0.416	0.780	0.462	0.798	0.432
128	1024	0.758	0.416	0.728	0.386	0.766	0.449	0.780	0.412
128	2048	0.741	0.397	0.722	0.354	0.753	0.436	0.747	0.401
128	4096	0.714	0.366	0.703	0.329	0.723	0.400	0.715	0.368
512	128	0.727	0.316	0.688	0.276	0.728	0.312	0.765	0.361
512	256	0.764	0.417	0.747	0.398	0.757	0.419	0.789	0.434
512	512	0.760	0.426	0.756	0.420	0.742	0.425	0.782	0.432
512	1024	0.723	0.388	0.667	0.355	0.743	0.394	0.759	0.414
512	2048	0.704	0.372	0.703	0.356	0.673	0.369	0.735	0.392
512	4096	0.699	0.349	0.680	0.340	0.698	0.330	0.718	0.375
4096	128	0.547	0.033	0.507	0.027	0.565	-0.001	0.569	0.074
4096	256	0.585	0.081	0.580	0.069	0.573	0.043	0.601	0.132
4096	512	0.606	0.150	0.595	0.122	0.598	0.125	0.625	0.202
4096	1024	0.614	0.205	0.563	0.147	0.627	0.203	0.653	0.266
4096	2048	0.654	0.280	0.601	0.205	0.658	0.292	0.702	0.342
4096	4096	0.677	0.324	0.659	0.289	0.669	0.319	0.704	0.364

Table 11: Full results for CDS: concordance accuracy ρ_C and Spearman correlation ρ_S for every (corruption chunk size c, resolution parameter d_{cds}) cell, averaged over the shuffle and merge corruption ladders, overall and broken down by length tier L. See Section B.3 for summary statistics derived from this grid.

c	d_{cds}	**Overall**		$L = 32$**K**		$L = 64$**K**		$L = 128$**K**	
		ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S
128	1	0.671	0.377	0.681	0.379	0.662	0.374	0.670	0.377
128	2	0.708	0.420	0.790	0.533	0.672	0.371	0.663	0.356
128	4	0.718	0.420	0.680	0.365	0.665	0.360	0.808	0.533
128	8	0.755	0.469	0.778	0.519	0.675	0.360	0.811	0.528
128	16	0.739	0.449	0.747	0.475	0.681	0.361	0.788	0.511
128	32	0.703	0.398	0.663	0.329	0.659	0.353	0.786	0.512
512	1	0.673	0.384	0.710	0.396	0.658	0.376	0.651	0.379
512	2	0.698	0.424	0.764	0.529	0.678	0.383	0.650	0.362
512	4	0.763	0.476	0.763	0.521	0.705	0.381	0.821	0.528
512	8	0.757	0.471	0.758	0.521	0.693	0.371	0.822	0.522
512	16	0.741	0.463	0.740	0.493	0.698	0.377	0.784	0.519
512	32	0.741	0.486	0.706	0.451	0.762	0.494	0.757	0.512
4096	1	0.653	0.362	0.672	0.354	0.650	0.345	0.638	0.387
4096	2	0.649	0.352	0.665	0.356	0.640	0.335	0.642	0.364
4096	4	0.706	0.371	0.667	0.343	0.660	0.334	0.790	0.436
4096	8	0.692	0.356	0.652	0.323	0.656	0.323	0.769	0.423
4096	16	0.683	0.338	0.649	0.311	0.646	0.299	0.753	0.405
4096	32	0.693	0.377	0.635	0.290	0.710	0.432	0.734	0.409

Table 12: Full results for LongAttn: concordance accuracy ρ_C and Spearman correlation ρ_S for every (corruption chunk size c, resolution parameter M) cell, averaged over the shuffle and merge corruption ladders, overall and broken down by length tier L. See Section B.3 for summary statistics derived from this grid.

c	M	**Overall**		$L = 32$**K**		$L = 64$**K**		$L = 128$**K**	
		ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S	ρ_C	ρ_S
128	4096	0.536	0.050	0.559	0.076	0.463	-0.065	0.584	0.140
128	8192	0.572	0.117	0.588	0.134	0.535	0.067	0.593	0.150
128	16384	0.576	0.137	0.557	0.110	0.574	0.150	0.596	0.152
128	32768	0.523	0.021	0.493	-0.020	0.505	-0.005	0.569	0.089
128	65536	0.564	0.114	–	–	0.552	0.096	0.577	0.133
128	131072	0.463	-0.038	–	–	–	–	0.463	-0.038
512	4096	0.509	0.026	0.541	0.090	0.477	-0.041	0.510	0.029
512	8192	0.546	0.101	0.555	0.112	0.535	0.071	0.548	0.119
512	16384	0.560	0.115	0.545	0.072	0.568	0.118	0.567	0.155
512	32768	0.521	0.035	0.508	-0.031	0.493	0.008	0.563	0.127
512	65536	0.570	0.110	–	–	0.548	0.079	0.592	0.142
512	131072	0.499	-0.018	–	–	–	–	0.499	-0.018
4096	4096	0.493	-0.019	0.548	0.082	0.463	-0.108	0.469	-0.032
4096	8192	0.551	0.099	0.565	0.136	0.531	0.057	0.557	0.105
4096	16384	0.548	0.084	0.535	0.053	0.542	0.071	0.568	0.130
4096	32768	0.533	0.053	0.516	0.022	0.528	0.026	0.554	0.111
4096	65536	0.531	0.076	–	–	0.520	0.039	0.542	0.113
4096	131072	0.490	-0.010	–	–	–	–	0.490	-0.010



(a) $c = 128$ (mean over tier) (b) $c = 512$ (mean over tier) (c) $c = 4096$ (mean over tier)

Figure 6: Cost-accuracy curves for all five scorers, averaged over length tier L and split by corruption chunk size c.

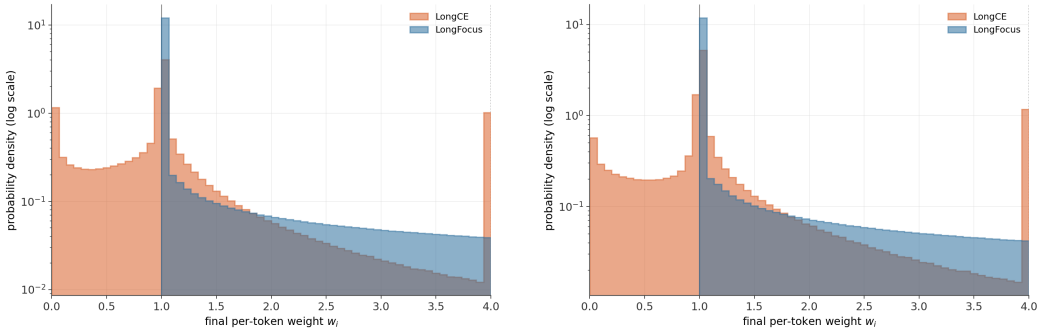
Table 13: Mean wall-clock seconds per document and approximate analytic FLOPs per document, averaged over ladder mode, corruption chunk size, and feasible length tiers at each resolution setting. Wall-clock time is the end-to-end duration of a 100-document process divided by 100, including model loading and I/O. LongAttn averages only tiers satisfying $M \leq L$.

Method	Resolution parameter	Wall-clock (s/doc)	FLOPs/doc
SegRatio (ours)	$s = 128$	15.79	4.268e15
SegRatio (ours)	$s = 256$	15.76	4.271e15
SegRatio (ours)	$s = 512$	15.74	4.276e15
SegRatio (ours)	$s = 1024$	15.71	4.286e15
SegRatio (ours)	$s = 2048$	15.72	4.307e15
SegRatio (ours)	$s = 4096$	16.00	4.348e15
LDS	$s = 128$	7.58	3.080e15
LDS	$s = 256$	11.87	5.029e15
LDS	$s = 512$	21.14	8.980e15
LDS	$s = 1024$	38.91	1.704e16
LDS	$s = 2048$	60.33	2.569e16
LDS	$s = 4096$	76.15	3.140e16
LongPPL	$s = 128$	60.97	1.198e16
LongPPL	$s = 256$	48.28	1.199e16
LongPPL	$s = 512$	43.17	1.200e16
LongPPL	$s = 1024$	39.77	1.202e16
LongPPL	$s = 2048$	39.59	1.207e16
LongPPL	$s = 4096$	40.37	1.215e16
CDS	$d_{\text{cds}} = 1$	55.95	4.023e15
CDS	$d_{\text{cds}} = 2$	31.81	3.530e15
CDS	$d_{\text{cds}} = 4$	19.71	3.284e15
CDS	$d_{\text{cds}} = 8$	13.70	3.161e15
CDS	$d_{\text{cds}} = 16$	10.67	3.099e15
CDS	$d_{\text{cds}} = 32$	9.23	3.068e15
LongAttn	$M = 4096$	0.58	2.062e12
LongAttn	$M = 8192$	0.68	4.673e12
LongAttn	$M = 16384$	1.10	1.154e13
LongAttn	$M = 32768$	2.65	3.189e13
LongAttn	$M = 65536$	8.48	9.896e13
LongAttn	$M = 131072$	15.52	3.386e14

C Analysis of Token Weighting

C.1 Overall Weight Distribution

Setup and purpose. We reuse the archived Llama-3-8B-Instruct weight-distribution probes on the 64K and 512K training corpora. The 64K runs start from the base model; the 512K runs start from the respective methods' trained checkpoints. We summarize their recorded weights to examine how broadly each method changes token weights. These pooled archives describe the recorded runs, rather than a paired comparison at a common model state. The quantitative 64K statistics use 120 retained batches for LongFocus and 670 for LongCE, including positions whose weights remain one.



(a) 64K stage

(b) 512K stage

Figure 7: Archived raw token-weight distributions at the 64K and 512K stages, on a logarithmic density scale.

Metric	LongFocus	LongCE
Gini coefficient $G(w)$	0.1665	0.3881
$A^{\text{mult}}_{0.05}$	20.04%	62.51%
$A^{\text{mult}}_{0.10}$	18.94%	56.96%
$A^{\text{mult}}_{0.20}$	17.23%	50.10%
$A^{\text{mult}}_{0.50}$	13.74%	39.09%
$A^{\text{mult}}_{1.00}$	9.82%	30.48%

Table 14: Descriptive weight statistics from the original 64K archives, pooled separately for each method and including unchanged weights.

Measurements. The *active-token ratio* counts weights that differ sufficiently from the standard CE weight of one:

$$A^{\text{mult}}_{\epsilon} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[w_i < \frac{1}{1+\epsilon} \text{ or } w_i > 1+\epsilon\right]. \tag{6}$$

Here, N includes all packed positions in each method's archive, and weights are not rescaled. For example, $\epsilon = 0.05$ counts weights above 1.05 or below $1/1.05$. The Gini coefficient measures inequality among all weights, including unchanged weights of one:

$$G(w) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|w_i - w_j|}{2N\sum_{i=1}^{N} w_i}. \tag{7}$$

It is zero when all weights are equal. These measures distinguish how often weights change from how unequally they are distributed.

Results. The archived distributions show that LongFocus retains weight one for most tokens and increases the weights of a subset, whereas LongCE changes weights both above and below one (Figure 7). In the 64K archives, the active-token ratio at the 5% tolerance is 20.04% for LongFocus and 62.51% for LongCE; Gini is 0.1665 and 0.3881, respectively.

C.2 Reweighting Quality Analysis

This section tests whether LongFocus's PreSelect and Reweight procedure turns coarse SegRatio scores into useful training weights. We assess *reweighting quality*: whether extra weight favors token predictions that benefit from remote context. Two controls examine the information in our token choices and the value of our weighting procedure; we then check whether this preference persists when segment boundaries move.

Setup. We fix Llama-3-8B-Instruct as the scoring model and select 40 passages of 32K-64K tokens from the ProLong-64K training corpus. A passage is a continuous text fragment identified by the dataset's document boundaries. Weighting hyperparameters follow the main experiments. For each original packed input, we reproduce LONGFOCUS's production weighting rule to get the final weight w_i. We compute these weights on the complete packed input before extracting the evaluated passage and targets. All comparisons below use these same frozen-model loss caches.

How to construct a reference token dependency signal? To evaluate the final weight assignments, we need a reference answer to a simple question: *which token predictions improve when more remote context is available?* Effective extra weighting should favor tokens with larger independently measured improvements. The texts have no annotated ground-truth dependencies, so we construct *a fine-grained reference to measure context benefit*. We predict the same token twice, once with a shorter context(2K) and once with a longer context(8K), keeping the model and target unchanged. For each evaluated token, we define

$$g_i = \ell_i^{\text{short}} - \ell_i^{\text{long}}, \tag{8}$$

where the two losses use shorter and longer contexts, respectively. A positive g_i means that additional context improves prediction. For efficiency, targets are evaluated in blocks of 64: target j sees $2048 + j$ or $8192 + j$ preceding tokens, for $j = 0, \ldots, 63$. The reference is used only for evaluation and does not determine w_i.

Measuring reweighting quality. We use the token-level reference gains g_i to evaluate the final weights from each assignment. The positive extra weight $(w_i - 1)_+ = \max(w_i - 1, 0)$ measures how much additional emphasis token i receives. We calculate the *weighted reference gain* U within each passage:

$$U = \frac{\sum_i (w_i - 1)_+ g_i}{\sum_i (w_i - 1)_+}. \tag{9}$$

This is the average reference gain per unit of extra weight. A larger U means that extra weight is more concentrated on predictions with greater measured benefit. Taking the positive part lets us evaluate upweighting consistently even when a control also downweights tokens; it leaves LONGFOCUS and Random unchanged. All assignments are evaluated on the same sampled targets with observable full-context, segmented, and reference losses, including targets below LONGFOCUS's loss floor. Such targets receive no extra weight from LONGFOCUS or Random, so their inclusion does not change those two assignments' U values. Segment-start targets without an observable segmented loss are excluded for all assignments. We report the mean over passages, weighting every passage equally. Confidence intervals(CI) (95%) are computed by bootstrapping passages within each domain (2,000 resamples, sampled with replacement). For each comparison, we use the same resampled passages for both assignments to estimate the paired difference.

Comparison design and purpose. On the same frozen model and passages, we compare LONGFOCUS with two variants that answer different questions. *Random* tests whether the token choices carry useful information when the amount and values of extra weight are fixed. It preserves LONGFOCUS's selected token count and extra-weight values on each complete passage, but assigns them to randomly chosen tokens passing the loss floor. These candidates are not restricted to LONGFOCUS's top-b or positive-score tokens; unselected tokens keep weight one. For Random only, we average U over 32 independent random assignments per passage. Extra-weight totals are matched on complete passages, but can differ on the sampled evaluation targets. SEGRATIO + *LongCE-style mapping* tests whether our complete PreSelect and Reweight procedure converts the coarse scores into more useful extra weight than direct score mapping(like LongCE). It replaces the loss-difference input of LongCE's exponential formula with the same SEGRATIO score r_i, giving $w_i = \min\{\exp(r_i), 4\}$ for every token with observable full-context and segmented losses. This variant uses the same ratio denominator protection, but no loss-floor filter, top-b selection, positive-score gate, or percentile conversion. Negative scores can produce weights below one; total extra weight is not matched. This tests our complete procedure on the same coarse signal, rather than rank mapping alone or the original LongCE method.

Comparison Results. LONGFOCUS achieves $U = 0.3874$, compared with 0.0843 for Random (Table 15). With the same number and values of extra weights, our token choices yield greater

Assignment	Weighted reference gain U	LONGFOCUS minus baseline U difference [95% CI]
LONGFOCUS	0.3874	—
Random	0.0843	0.3032 [0.2075, 0.4031]
LONGFOCUS with LongCE-style mapping	0.1975	0.1899 [0.1180, 0.2651]

Table 15: Reweighting quality comparison. Random controls the selected count and extra-weight values per passage. LongCE-style mapping controls the input scores while replacing our complete filtering, selection, and rank-based weighting procedure.

Offset	U	Change in U from offset 0 [95% CI]	Reference overlap	
			$w_i \geq 2$	$w_i \geq 3$
0	0.3874	—	29.02% (15.27%)	16.49% (6.34%)
64	0.3849	-0.0025 $[-0.0245, 0.0180]$	27.78% (15.49%)	17.14% (6.54%)
128	0.3875	$+0.0001$ $[-0.0376, 0.0347]$	28.59% (15.83%)	16.71% (6.97%)
192	0.4070	$+0.0196$ $[-0.0292, 0.0678]$	28.68% (16.22%)	17.81% (7.03%)
256	0.4058	$+0.0184$ $[-0.0112, 0.0450]$	27.02% (14.60%)	17.58% (6.23%)
320	0.4006	$+0.0132$ $[-0.0180, 0.0447]$	28.53% (15.07%)	17.45% (6.15%)
384	0.3773	-0.0101 $[-0.0484, 0.0291]$	27.55% (15.39%)	17.42% (6.26%)
448	0.3796	-0.0079 $[-0.0322, 0.0156]$	26.79% (15.11%)	16.51% (6.22%)

Table 16: Reweighting quality across eight segment offsets. Intervals are pointwise 95% paired bootstrap confidence intervals.

remote-context benefit. The enrichment therefore reflects information in the weight assignment, not merely a larger total weight. This supports an informative training signal, while the frozen probe alone cannot attribute the main training gains or rule out optimization effects.

Using the same coarse scores, direct LongCE-style mapping yields $U = 0.1975$ while LONGFOCUS improves this by 0.1899 (95% CI $[0.1180, 0.2651]$). Thus the complete PreSelect and Reweight procedure makes the coarse signal more useful: filtering, top-b selection, and rank-based mapping together concentrate extra loss weight on predictions with greater remote-context benefit. This is an improvement in benefit per unit of *additional* weight; U does not evaluate the direct variant's downweighting or overall training performance.

Sensitivity to segment offset. We test whether reweighting quality depends on a particular placement of the artificial segment boundaries. On the same frozen model and 40 passages, we evaluate eight evenly spaced offsets from 0 to 448 tokens, covering one 512-token segment period. For each offset, we recompute the segmented scores and the complete PreSelect and Reweight procedure, keeping the text, full-context losses, and independent reference gains unchanged. We retain the initial short segment rather than discarding text. We compute U on eligible sampled targets at each offset, following the preceding comparisons.

In addition to U, we measure whether high-weight tokens agree with the independent reference. For these set comparisons, we use the same 13,655 sampled targets eligible under every offset. For each passage, we take the targets with $w_i \geq 2$ or $w_i \geq 3$ and compare them with the same number of targets having the largest g_i. We report the fraction of the high-weight set shared with this reference set, together with the expected overlap for uniform random selection of the same size.

Across the eight offsets, U stays between 0.3773 and 0.4070, around the offset-0 value of 0.3874. The paired confidence intervals for changes from offset 0 all include zero. Reference overlap also remains comparable across offsets: 26.79-29.02% for $w_i \geq 2$ and 16.49-17.81% for $w_i \geq 3$, consistently above the respective chance expectations (Table 16). Thus both the benefit per unit of extra weight and the reference enrichment of high-weight tokens persist throughout the tested offsets. These results support that reweighting quality is not tied to a particular boundary placement, even though the identities of individual high-weight tokens may change.

D Detailed Experiment Results

Table 17: Hyperparameter ablation of LongFocus on Llama-3-8B-Instruct at the 64K and 64K+SFT stages. Long-Avg averages HELMET and RULER at 64K, and HELMET, RULER, and BABILong at 64K+SFT.

					64K			64K+SFT			
Variant	b	γ	S	PreSelect	HELMET	RULER	**Long-Avg**	HELMET	RULER	BABILong	**Long-Avg**
Vanilla CE	–	–	–	–	45.40	86.07	65.73	47.96	87.76	48.79	61.50
LongFocus (default)	0.5	2	512	✓	47.34	87.52	67.43	50.20	88.50	49.62	62.78
Reweighting fraction b											
b=0.1	**0.1**	2	512	✓	46.65	87.32	66.98	50.56	88.44	49.12	62.70
b=0.7	**0.7**	2	512	✓	47.59	87.30	67.45	51.16	88.75	48.41	62.77
Reweighting exponent γ											
γ=0.5	0.5	**0.5**	512	✓	47.80	87.52	67.66	50.45	87.37	48.69	62.17
γ=1	0.5	**1**	512	✓	46.54	86.80	66.67	50.61	88.38	48.08	62.36
Segment-view length S											
S=128	0.5	2	**128**	✓	47.25	87.44	67.34	49.98	88.57	48.68	62.41
S=2048	0.5	2	**2048**	✓	47.27	87.73	67.50	50.63	88.79	48.92	62.78
S=8192	0.5	2	**8192**	✓	47.04	86.02	66.53	50.10	86.76	48.25	61.71
PreSelect gate											
w/o PreSelect	0.5	2	512	×	47.22	86.58	66.90	50.12	88.16	48.87	62.38

We report the full benchmark results of the mid-training experiments described in Section 4 in this section.

D.1 Llama-3-8B-Instruct full results

Table 18: Full short-context benchmark results on Llama-3-8B-Instruct.

		Knowledge and commonsense QA					Math		Reasoning		Instr.-Follow.	
Stage	**Method**	ARC-C	HellaSwag	MMLU	MMLU-Pro	WinoGrande	GSM8K	MATH-hard	BBH	MuSR	IFEval	**Short-Avg**
64K	Vanilla CE	60.58	81.93	**66.31**	32.70	78.69	**59.74**	6.72	48.32	40.74	10.91	48.66
	LongCE	**61.95**	**82.46**	65.67	32.41	**79.48**	57.39	5.66	48.01	**45.11**	**12.20**	**49.03**
	LongFocus (ours)	59.90	81.79	66.21	**33.25**	77.98	**59.74**	**6.87**	**48.65**	41.27	11.46	48.71
64K+SFT	Vanilla CE	**66.21**	**82.59**	64.20	32.56	**78.69**	**70.20**	**6.27**	49.87	42.33	38.45	53.14
	LongCE	66.04	82.57	64.28	32.80	78.61	69.14	5.82	49.87	**42.46**	42.51	53.41
	LongFocus (ours)	65.78	82.48	**64.53**	**33.05**	78.53	69.37	6.04	**50.20**	42.20	**43.99**	**53.62**
64K+512K	Vanilla CE	59.39	81.95	**65.75**	32.75	77.82	58.15	**6.42**	47.75	40.61	10.54	48.11
	LongCE	**62.29**	**82.68**	64.46	32.30	**77.90**	56.48	5.36	**48.67**	**46.03**	11.09	**48.73**
	LongFocus (ours)	59.13	81.71	65.50	**32.90**	77.03	**59.74**	6.12	48.57	41.93	**11.46**	48.41
64K+512K+SFT	Vanilla CE	**64.93**	82.43	63.82	**32.89**	77.98	**67.63**	**6.12**	**49.42**	41.40	36.97	52.36
	LongCE	64.42	82.36	**64.16**	32.55	**78.45**	65.35	5.82	49.07	**42.72**	**42.33**	**52.72**
	LongFocus (ours)	**64.93**	**82.58**	64.04	32.52	78.30	66.49	5.44	48.62	41.53	39.56	52.40

D.2 MoE-16B-A3B full results

D.3 MoE-125B-A10B full results

E Case Study

We present several case study examples in Figures 8 to 10.

Table 19: HELMET results by task category and context length on Llama-3-8B-Instruct (four-category average).

Method	**Mean**	RAG	ICL	Cite	Re-rank
64K mid-training (no SFT)					
Vanilla CE	45.4	61.0	76.8	9.0	34.8
LongCE	**47.8**	59.5	**79.2**	**16.1**	36.4
LONGFOCUS	47.3	**61.3**	78.7	12.5	**36.9**
64K+SFT					
Vanilla CE	48.0	66.0	72.7	11.6	**41.5**
LongCE	47.4	65.4	**77.9**	8.4	38.0
LONGFOCUS	**50.2**	**66.9**	76.0	**16.5**	**41.5**
64K+512K mid-training (no SFT)					
Vanilla CE	45.2	61.0	79.0	8.7	32.3
LongCE	**47.4**	59.6	**79.8**	**17.9**	32.4
LONGFOCUS	46.4	**61.5**	78.7	12.4	**33.2**
64K+512K+SFT					
Vanilla CE	48.4	64.9	75.8	14.8	38.3
LongCE	47.7	**65.5**	**79.2**	8.2	38.0
LONGFOCUS	**49.8**	65.2	77.2	**16.6**	**40.1**

(a) By task category

Method	8K	16K	32K	64K	128K
64K mid-training (no SFT)					
Vanilla CE	51.5	47.0	44.6	38.4	
LongCE	**54.5**	**50.3**	45.6	**40.8**	
LONGFOCUS	53.6	49.1	**46.7**	39.9	
64K+SFT					
Vanilla CE	56.3	49.4	45.7	40.4	
LongCE	55.9	48.0	46.1	39.8	
LONGFOCUS	**57.6**	**53.3**	**47.6**	**42.3**	
64K+512K mid-training (no SFT)					
Vanilla CE	49.9	47.5	45.3	43.0	40.5
LongCE	**53.9**	**51.8**	**47.5**	**43.4**	40.5
LONGFOCUS	51.5	49.9	46.1	43.3	**41.4**
64K+512K+SFT					
Vanilla CE	57.1	53.2	47.4	43.8	40.5
LongCE	55.7	49.7	46.9	44.4	42.0
LONGFOCUS	**58.1**	**54.7**	**48.9**	**44.8**	**42.3**

(b) By context length

Table 20: RULER results by task category and context length on Llama-3-8B-Instruct. The mean is computed over lengths and shown in the by-length subtable.

Method	NIAH	VT	CWE	FWE	QA
64K mid-training (no SFT)					
Vanilla CE	96.5	80.6	66.1	88.9	55.8
LongCE	95.7	80.8	76.6	80.1	55.6
LONGFOCUS	**96.8**	**81.0**	**78.5**	**89.7**	**57.0**
64K+SFT					
Vanilla CE	**95.7**	**86.1**	68.1	**91.5**	64.8
LongCE	94.9	80.1	80.4	86.1	61.9
LONGFOCUS	95.4	81.8	**82.0**	91.1	**66.3**
64K+512K mid-training (no SFT)					
Vanilla CE	**99.3**	**99.6**	58.1	90.3	55.8
LongCE	98.1	98.7	64.3	85.9	**58.1**
LONGFOCUS	98.5	**99.6**	**66.1**	**90.5**	57.9
64K+512K+SFT					
Vanilla CE	**98.8**	92.6	58.4	87.8	66.0
LongCE	97.9	**95.5**	**65.1**	85.7	**67.2**
LONGFOCUS	98.1	94.4	61.5	**88.0**	66.6

(a) By task category

Method	**Mean**	4K	8K	16K	32K	64K	128K
64K mid-training (no SFT)							
Vanilla CE	86.1	**94.6**	92.0	**93.0**	87.6	63.2	
LongCE	85.7	93.2	91.5	91.4	**88.7**	63.5	
LONGFOCUS	**87.5**	94.4	**92.6**	92.9	88.6	**69.1**	
64K+SFT							
Vanilla CE	87.8	**95.0**	**93.2**	93.3	88.0	69.3	
LongCE	86.9	94.3	93.0	92.9	**90.6**	63.7	
LONGFOCUS	**88.5**	94.3	93.0	**93.5**	89.6	**72.1**	
64K+512K mid-training (no SFT)							
Vanilla CE	88.8	94.7	91.3	91.0	89.9	84.0	**81.5**
LongCE	88.4	93.3	91.4	92.0	90.3	83.7	79.9
LONGFOCUS	**89.3**	**95.1**	**91.9**	**92.6**	**91.1**	**84.3**	80.5
64K+512K+SFT							
Vanilla CE	89.3	94.0	**92.3**	92.4	90.5	84.3	82.3
LongCE	**89.6**	**94.1**	91.6	92.0	**91.6**	**85.5**	**82.5**
LONGFOCUS	89.4	93.7	92.2	**93.1**	**91.6**	84.3	81.4

(b) By context length

Table 21: BABILong results by context length on Llama-3-8B-Instruct.

Method	**Mean**	0K	1K	2K	4K	8K	16K	32K	64K	128K	256K	512K
64K+SFT												
Vanilla CE	48.8	50.7	**50.9**	50.4	50.5	49.7	48.4	46.1	**43.6**			
LongCE	48.0	**52.7**	48.0	48.8	49.6	49.5	48.4	**46.5**	40.4			
LONGFOCUS	**49.6**	52.3	50.8	**50.9**	**51.8**	**51.1**	**50.2**	**46.5**	43.4			
64K+512K+SFT												
Vanilla CE	45.5	51.6	51.5	**51.5**	50.4	49.3	49.8	48.3	46.6	41.6	36.0	23.6
LongCE	44.2	**54.0**	48.5	47.3	48.4	48.6	49.2	46.1	46.0	41.0	36.6	20.8
LONGFOCUS	**46.7**	52.7	**52.3**	51.4	**51.5**	**50.4**	**51.4**	**49.6**	**48.8**	**43.0**	**38.0**	**24.8**

Table 22: Full short-context benchmark results on MoE-16B-A3B.

		Knowledge and commonsense QA					Math		Reasoning		Instr.-Follow.	
Stage	**Method**	ARC-C	HellaSwag	MMLU	MMLU-Pro	WinoGrande	GSM8K	MATH-hard	BBH	MuSR	IFEval	**Short-Avg**
64K	Vanilla CE	**63.74**	**81.67**	**72.78**	**43.38**	75.69	60.96	23.79	**45.11**	41.27	18.67	52.71
	LONGFOCUS (ours)	63.65	81.56	72.23	43.18	**75.77**	**73.92**	**25.91**	44.19	**42.06**	**20.15**	**54.26**
64K+SFT	Vanilla CE	**65.10**	81.74	**72.76**	**43.55**	75.85	**79.38**	**26.59**	**52.27**	**44.71**	21.44	**56.34**
	LONGFOCUS (ours)	64.59	**81.83**	72.68	43.48	**76.56**	77.86	25.53	51.68	**44.71**	**23.66**	56.26
64K+512K	Vanilla CE	**63.74**	81.56	**72.62**	42.55	**75.22**	68.46	**23.56**	44.59	41.40	18.67	53.24
	LONGFOCUS (ours)	63.57	**81.59**	71.95	**43.58**	73.88	**72.10**	**23.56**	**44.65**	**41.93**	**20.15**	**53.70**
64K+512K+SFT	Vanilla CE	**65.19**	81.44	**72.60**	42.90	**76.09**	**77.63**	**25.23**	51.74	**44.31**	22.55	**55.97**
	LONGFOCUS (ours)	64.93	**81.45**	72.30	**43.13**	74.90	77.33	24.24	**51.76**	44.18	**23.11**	55.73

Table 23: HELMET results by task category and context length on MoE-16B-A3B (four-category average). The four-category mean excludes the near-saturated Recall category.

Method	**Mean**	RAG	ICL	Cite	Re-rank
64K mid-training (no SFT)					
Vanilla CE	**35.4**	30.4	**76.0**	**4.2**	31.0
LONGFOCUS	34.6	**31.3**	72.0	3.3	**31.8**
64K+SFT					
Vanilla CE	43.4	59.4	**74.4**	7.7	32.2
LONGFOCUS	**44.2**	**60.3**	73.6	**8.3**	**34.5**
64K+512K mid-training (no SFT)					
Vanilla CE	30.8	14.2	**81.3**	1.4	26.3
LONGFOCUS	**34.5**	**26.5**	80.3	**2.1**	**29.0**
64K+512K+SFT					
Vanilla CE	43.7	**59.6**	**81.2**	4.6	29.3
LONGFOCUS	**44.2**	58.9	80.6	**5.8**	**31.6**

(a) By task category

Method	8K	16K	32K	64K	128K
64K mid-training (no SFT)					
Vanilla CE	**36.4**	**36.3**	**34.0**	34.9	
LONGFOCUS	33.9	35.2	33.6	**35.7**	
64K+SFT					
Vanilla CE	49.5	45.2	41.1	37.7	
LONGFOCUS	**50.2**	**46.2**	**41.5**	**38.9**	
64K+512K mid-training (no SFT)					
Vanilla CE	33.0	31.6	29.2	29.0	31.1
LONGFOCUS	**34.5**	**34.9**	**33.9**	**33.9**	**35.1**
64K+512K+SFT					
Vanilla CE	50.5	45.0	42.9	40.8	39.2
LONGFOCUS	**50.6**	**46.5**	**43.3**	**41.1**	**39.6**

(b) By context length

Table 24: RULER results by task category and context length on MoE-16B-A3B. For task types, NIAH averages the needle variants and QA averages HotpotQA and SQuAD. The mean is computed over lengths and shown in the by-length subtable (64K-family models stop at 64K while 512K-family models reach 128K; blank = length not evaluated).

Method	NIAH	VT	CWE	FWE	QA
64K mid-training (no SFT)					
Vanilla CE	96.2	92.4	56.3	88.2	**54.2**
LONGFOCUS	**96.9**	**94.2**	**65.1**	**89.3**	53.9
64K+SFT					
Vanilla CE	**97.1**	**98.1**	57.8	88.7	57.6
LONGFOCUS	96.5	97.9	**62.6**	**90.6**	**61.0**
64K+512K mid-training (no SFT)					
Vanilla CE	95.9	**91.3**	44.0	66.3	51.7
LONGFOCUS	**97.2**	89.0	**59.5**	**78.5**	**53.8**
64K+512K+SFT					
Vanilla CE	**97.4**	**99.0**	41.5	85.3	**58.1**
LONGFOCUS	97.1	97.2	**55.8**	**89.0**	57.8

(a) By task category

Method	**Mean**	4K	8K	16K	32K	64K	128K
64K mid-training (no SFT)							
Vanilla CE	85.7	93.6	89.6	87.6	83.7	74.3	
LONGFOCUS	**87.0**	**94.0**	**89.8**	**87.7**	**84.8**	**78.9**	
64K+SFT							
Vanilla CE	87.4	95.3	91.0	88.9	84.3	77.6	
LONGFOCUS	**88.1**	**95.4**	**92.1**	**89.6**	**85.4**	**77.9**	
64K+512K mid-training (no SFT)							
Vanilla CE	82.5	92.8	88.7	86.9	81.9	75.9	68.7
LONGFOCUS	**85.6**	**93.0**	**89.6**	**88.2**	**85.8**	**82.1**	**74.7**
64K+512K+SFT							
Vanilla CE	86.2	**95.2**	91.4	87.1	85.4	81.6	76.7
LONGFOCUS	**87.3**	95.0	**92.2**	**90.1**	**86.7**	**82.4**	**77.2**

(b) By context length

Table 25: BABILong results by context length on MoE-16B-A3B.

Method	**Mean**	0K	1K	2K	4K	8K	16K	32K	64K	128K	256K	512K
64K+SFT												
Vanilla CE	29.5	**50.1**	**31.6**	24.8	23.1	24.7	26.4	28.0	27.4			
LONGFOCUS	**33.4**	49.4	31.3	**29.0**	**28.3**	**30.9**	**31.9**	**33.9**	**33.0**			
64K+512K+SFT												
Vanilla CE	30.2	**48.4**	**33.6**	29.4	26.3	26.7	26.7	29.9	32.4	**31.8**	25.6	21.0
LONGFOCUS	**33.2**	47.7	32.8	**33.0**	**31.3**	**33.2**	**34.7**	**35.4**	**34.2**	31.2	**29.6**	**21.6**

Table 26: Full short-context benchmark results on MoE-125B-A10B. Short-Avg is the equally weighted mean of all ten tasks.

		Knowledge and commonsense QA					Math		Reasoning		Instr.-Follow.	
Stage	**Method**	ARC-C	HellaSwag	MMLU	MMLU-Pro	WinoGrande	GSM8K	MATH-hard	BBH	MuSR	IFEval	**Short-Avg**
64K	Vanilla CE	**64.51**	**82.64**	**73.54**	43.67	77.11	**76.20**	**20.39**	**47.51**	**50.00**	23.66	**55.92**
	LONGFOCUS (ours)	64.33	82.39	73.52	**44.57**	**77.66**	74.40	20.24	47.23	48.15	**24.40**	55.69
64K+SFT	Vanilla CE	64.33	82.63	72.87	43.80	75.85	**67.60**	16.62	46.97	47.88	21.07	53.96
	LONGFOCUS (ours)	**65.44**	**82.77**	**73.34**	**45.20**	**77.66**	60.20	**21.22**	**51.36**	**52.25**	**21.26**	**55.07**
64K+512K	Vanilla CE	64.59	**83.25**	**72.95**	44.45	76.32	71.70	19.79	**47.77**	**49.74**	**23.84**	**55.44**
	LONGFOCUS (ours)	**65.02**	82.50	72.92	**44.56**	**77.51**	**72.00**	**20.54**	46.03	48.94	22.92	55.29
64K+512K+SFT	Vanilla CE	63.40	82.62	72.41	44.48	75.93	50.04	17.37	48.17	49.60	19.78	52.38
	LONGFOCUS (ours)	**65.70**	**83.11**	**73.09**	**45.00**	**77.90**	**68.61**	**20.39**	**51.73**	**52.12**	**20.89**	**55.85**

Table 27: HELMET results by task category and context length on MoE-125B-A10B (four-category average). The near-saturated Recall category is excluded.

Method	**Mean**	RAG	ICL	Cite	Re-rank
64K mid-training (no SFT)					
Vanilla CE	30.0	5.3	72.8	**5.9**	35.8
LONGFOCUS	**31.7**	**5.6**	**75.3**	5.1	**40.9**
64K+SFT					
Vanilla CE	36.3	58.5	53.4	3.5	29.5
LONGFOCUS	**46.3**	**65.3**	**73.7**	**3.7**	**42.5**
64K+512K mid-training (no SFT)					
Vanilla CE	30.0	**3.6**	81.6	3.7	31.2
LONGFOCUS	**31.3**	3.5	**82.6**	**4.5**	**34.7**
64K+512K+SFT					
Vanilla CE	40.0	50.9	78.6	3.0	27.4
LONGFOCUS	**47.5**	**65.4**	**82.6**	**4.4**	**37.6**

(a) By task category

Method	8K	16K	32K	64K	128K
64K mid-training (no SFT)					
Vanilla CE	31.8	30.6	29.0	28.5	
LONGFOCUS	**32.7**	**33.2**	**30.2**	**30.7**	
64K+SFT					
Vanilla CE	38.1	36.6	34.8	35.6	
LONGFOCUS	**48.5**	**48.6**	**45.1**	**43.1**	
64K+512K mid-training (no SFT)					
Vanilla CE	33.1	31.6	30.1	28.4	27.0
LONGFOCUS	**33.8**	**31.8**	**31.6**	**30.1**	**29.2**
64K+512K+SFT					
Vanilla CE	40.5	40.3	41.1	39.8	38.2
LONGFOCUS	**51.4**	**49.4**	**47.1**	**46.1**	**43.5**

(b) By context length

Table 28: RULER results by task category and context length on MoE-125B-A10B. NIAH averages the needle variants; QA averages HotpotQA and SQuAD. The mean is computed over lengths and shown in the by-length subtable.

Method	NIAH	VT	CWE	FWE	QA
64K mid-training (no SFT)					
Vanilla CE	**98.1**	**98.0**	70.0	94.3	64.9
LONGFOCUS	97.4	94.9	**73.2**	**95.7**	**69.5**
64K+SFT					
Vanilla CE	**98.2**	**97.6**	71.8	**96.3**	**67.2**
LONGFOCUS	96.7	95.6	**76.1**	**96.3**	65.2
64K+512K mid-training (no SFT)					
Vanilla CE	96.2	**97.0**	61.4	96.0	59.2
LONGFOCUS	**97.0**	92.4	**66.8**	**96.6**	**63.9**
64K+512K+SFT					
Vanilla CE	**96.3**	**95.8**	59.2	96.3	63.0
LONGFOCUS	96.2	93.7	**70.1**	**97.1**	**69.2**

(a) By task category

Method	**Mean**	4K	8K	16K	32K	64K	128K
64K mid-training (no SFT)							
Vanilla CE	90.5	**95.4**	94.0	91.9	88.7	82.8	
LONGFOCUS	**90.9**	95.0	**94.3**	**92.6**	**89.4**	**83.2**	
64K+SFT							
Vanilla CE	**91.2**	**95.9**	94.6	**93.3**	**89.7**	**82.5**	
LONGFOCUS	90.1	95.6	**94.7**	91.6	86.9	81.8	
64K+512K mid-training (no SFT)							
Vanilla CE	87.9	**94.6**	**93.2**	91.6	87.3	80.9	79.7
LONGFOCUS	**89.2**	94.3	93.1	**92.1**	**89.7**	**84.0**	**81.8**
64K+512K+SFT							
Vanilla CE	88.3	**95.1**	93.7	**92.7**	86.3	83.2	78.8
LONGFOCUS	**89.9**	94.8	**93.8**	92.5	**91.0**	**86.3**	**80.9**

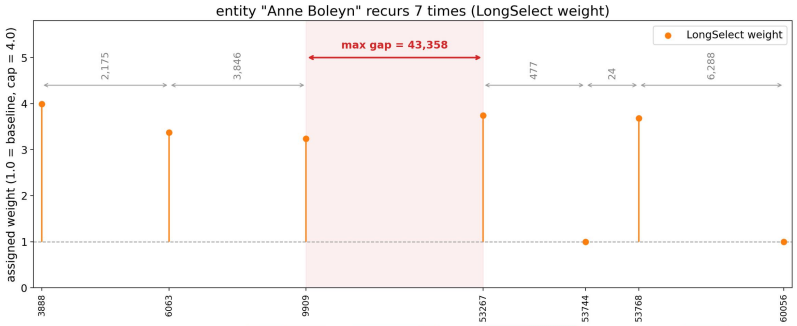
(b) By context length

Table 29: BABILong results by context length on MoE-125B-A10B.

Method	**Mean**	0K	1K	2K	4K	8K	16K	32K	64K	128K	256K	512K
64K+SFT												
Vanilla CE	30.8	23.6	40.1	34.0	30.9	30.3	29.1	29.3	29.2			
LONGFOCUS	**43.7**	**33.2**	**46.4**	**46.1**	**47.3**	**46.1**	**44.9**	**43.7**	**41.4**			
64K+512K+SFT												
Vanilla CE	32.1	35.3	**46.7**	41.6	41.7	39.5	36.7	32.4	26.8	22.4	20.8	**9.4**
LONGFOCUS	**39.4**	**38.0**	45.6	**46.2**	**46.2**	**45.8**	**44.8**	**43.9**	**42.6**	**37.4**	**34.0**	9.0



Figure 8: Case Study of LONGFOCUS's on sample 1



...(1567 tokens omitted)... There is no room for mistakes at court.'\n\n# Spring 1522\n\n'I am going to France tomorrow ...(1483 tokens omitted)... pair we shall be. What man could resist us?'\n\nI laughed, she could always make me laugh. ...(781 tokens omitted)... sister has come to join our company,' the queen said. 'This is Anne Boleyn.'\n\n'George's sister?' the king asked.\n\nMy brother bowed. ' ...(1206 tokens omitted)... voices.\n\n'She must be losing hope of it,' George said. 'What is she now? Thirty ...(490 tokens omitted)... him heir if there's no issue from the queen.'\n\n'A bastard on the throne?'\n\n'He wasn't ...(1002 tokens omitted)... dumped, dizzy and giggly, on the ground. When we ladies came out from the castle and danced ...(2863 tokens omitted)... There was a brief silence.\n\n'I'll speak with Carey's father ,' my father volunteered. 'William will ...(342 tokens omitted)... and the wreckage of his own hopes for a marriage and a son.\n\n'Anne Boleyn,' He bowed briefly. 'Have you come to help your sister onwards and ...(328 tokens omitted)... was a relief to go to bed at night with Anne. We drew the curtains around us so that we ...(1254 tokens omitted)... yes! And you all but fainted when I unmasked and you discovered me. Who did you think ...(3799 tokens omitted)... her bidding, and I saw why she was sending me away. The king was coming slowly across the grass ...(164 tokens omitted)... . I just looked at him. If Anne or George had been close by they could have prompted me with ...(1029 tokens omitted)... ,' the king said abruptly. 'M istress Carey should not be running round the jousting ground ...(1009 tokens omitted)... .'\n\nGeorge bowed and went to the back of the tent. I drifted after him. 'Give it me ...(789 tokens omitted)... oath at Peter for pinching him with a buckle.'\n\nHenry laughed. 'God be praised!'\n\nTwo ...(540 tokens omitted)... 'You've got to get it.' She pushed me in the small of my back and I stepped forward ...(954 tokens omitted)... . He wore my kerchief under his breastplate.'\n\n'He dropped it and forgot it,' Anne pointed ...(1672 tokens omitted)... king dropped it and his page gave my scarf to Queen Mary,' I said, my throat tight with nervous ...(910 tokens omitted)... 'What matters is that one of us catches the king. It hardly matters which one. If his taste ...(2984 tokens omitted)... my days were very long.\n\nI had no-one to talk to . I had no company to speak of ...(4688 tokens omitted)... Very pretty.'\n\n'He'll send for you,' Anne said that night as we undressed in the room ...(1184 tokens omitted)... bedroom while I was snug and loved in the King of England's bed. 'My God, you must ...(3294 tokens omitted)... breakfast and smiled as he took his seat. The queen's pale eyes took in the rosy colour of my face, the rich gleam of my cream gown, and looked away. She called for some musicians ...(380 tokens omitted)... boy and I beamed back. My mare, Jes mond, was like a coiled spring, and ...(2638 tokens omitted)... about the room, unpacking my clothes from the chest and laying them out for the morning.\n\n'You ...(15330 tokens omitted)... onto the top of her head and admired her face in the mirror. 'Anne Boleyn in love,' he said thoughtfully. 'Who'd have believed it?'\n\n ...(6755 tokens omitted)... in shabby clothes stood uneasily on the threshold.\n\n'A letter for Mistress Anne Boleyn,' he said. 'For none but her, and the young lord said ...(5461 tokens omitted)...

Figure 9: Case Study of LONGFOCUS's on sample 2



Figure 10: Case Study of LONGFOCUS's on sample 3