CALIBRATING LONG-RANGE ATTENTION WITH SEMANTIC NEEDLES FOR LONG-CONTEXT PRETRAINING DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective long-context training requires documents with meaningful long-range dependencies, not merely long documents. Existing model-based selectors broadly follow two approaches: prediction-based contrast and attention-based measurement. Prediction-based selectors can measure the contribution of distant context but are costly to apply across a corpus, while attention-based selectors are more efficient yet often aggregate over heads without identifying those responsive to useful distant evidence. We propose SNAS (Semantic-Needle-guided Attention-based data Selection), which uses prediction contrast on a small calibration corpus to identify naturally occurring context-dependent target tokens and localize their distant evidence intervals. SNAS then calibrates model-specific attention heads by their enrichment over these intervals, without synthetic probes or downstream-task labels. For corpus-wide selection, it scores each document in a single forward pass using both distant-attention allocation and retrieved-content signals from the calibrated heads. Across two scoring models and three data domains, SNAS matches or numerically exceeds the strongest baseline in overall score across all six settings. With the Llama-3.1-8B scorer, SNAS achieves $1.8\times$ LongFilter's corpus-scoring throughput and uses 44% fewer GPU-hours; these measurements cover corpus scoring and exclude the one-time calibration cost.

1 INTRODUCTION

Long-context capability underpins a model's performance on tasks such as cross-document reasoning, repository-level code, and long-document summarization, and depends substantially on the long-context data seen during pretraining. Yet such data is not useful merely by being long: recent work suggests that what matters is whether a document contains meaningful long-range dependencies, and much long-form text does not, since most tokens can be predicted from local context alone (Fang et al., 2025). Selecting data that genuinely contains them is therefore a central challenge.

Existing model-based approaches to long-context data selection broadly fall into two families: prediction-based contrast and attention-based measurement. Prediction-based methods contrast next-token predictions across different context lengths. For example, LongFilter (Deng et al., 2026) applies this at the document level via a token-level KL-divergence surrogate. These methods directly measure the contribution of distant context, but require multiple forward passes per document, making them costly at billion-token scale.

Attention-based methods instead estimate long-range dependencies directly from attention, and typically require only a single forward pass per document, avoiding the repeated passes that make prediction-based contrast costly. They usually define distant keys via a positional threshold and aggregate attention over many heads and query positions, as in LongAttn and LADM. LongAttn scores documents from first-layer attention between later queries and distant earlier keys (Wu et al., 2025a), while LADM aggregates attention between fixed-length spans across all heads and layers of a separately trained Long Attention Calculator (Chen et al., 2025).

This efficiency, however, comes at the cost of a coarser selection signal that does not fully exploit long-range dependencies, owing to two main limitations. First, prior work shows that attention heads are functionally specialized: some handle copying, others retrieval, still others positional

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.

patterns (Wu et al., 2025b; Olsson et al., 2022). Aggregating over all heads can dilute the useful signal with heads that are not attuned to long-range dependencies. A few methods select data using specific heads, as in AIR and AttentionInfluence, which use retrieval heads for data selection (Liu et al., 2025; Hua et al., 2025). However, these heads are found with synthetic, copy-oriented probes and capture surface matching, which need not reflect the semantic dependencies that link distant spans in natural text, a gap that QRHead also identifies in the inference setting (Zhang et al., 2025). It therefore remains open how to identify heads that capture the distant evidence behind natural context-dependent predictions and use them for long-context data selection. Second, these selectors aggregate over broad query and key regions even though long-range signal is sparse on both sides: only a small subset of query positions benefits from distant context, and models often underuse the middle of long inputs (Fang et al., 2025; Liu et al., 2024a), while attention over keys concentrates on a few positions, including initial sinks that carry little long-range information (Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024). They also score by attention mass alone, without accounting for what is retrieved from distant keys. We hypothesize that this retrieved-content signal can better distinguish useful long-range attention from attention that merely spreads mass.

We aim for an attention-based selector that retains single-pass efficiency while overcoming the limitations above to select data for stronger long-context performance. To this end we propose SNAS (Semantic-Needle-guided Attention-based data Selection), built on the *semantic needle*: a naturally occurring token whose prediction substantially benefits from a distant evidence interval, unlike the synthetic target planted by needle-in-a-haystack probes. On a small calibration corpus, a long–short prediction contrast flags such target tokens, and SNAS localizes the evidence interval behind each, requiring no probes or labels. To select heads sensitive to genuine long-range dependencies, SNAS keeps those most enriched over the evidence intervals (addressing the first limitation). To score a document, SNAS uses one standard model forward pass and extracts attention and value statistics only from the calibrated heads, using a sparse set of query positions and their distant causal keys. It rates the document by how much attention the model allocates to distant keys and how much content that attention retrieves, not by attention mass alone (addressing the second limitation). Across both scoring models (Llama-3.1-8B-Base and Qwen3-8B-Base) and three domains (Common Crawl, arXiv, and Books), SNAS matches the strongest prior selector's quality. With the Llama scorer, corpus-wide scoring is $1.8\times$ faster and uses 44% fewer GPU-hours than LongFilter, excluding the one-time calibration cost. Overall, our main contributions are:

(1) We introduce the *semantic needle*, a naturally occurring target-token–evidence pair, and use it to calibrate model-specific attention heads whose attention is repeatedly enriched over the corresponding evidence intervals, without synthetic retrieval probes or downstream labels.

(2) On these calibrated heads and a sparse set of query positions, SNAS scores each document in a single traversal, combining attention allocation to distant causal keys with the content retrieved from them, rather than scoring by attention mass alone.

(3) Experiments across two scoring models (Llama-3.1-8B and Qwen3-8B-Base) and three domains (Common Crawl, arXiv, and Books) show that SNAS selects data of quality comparable to the strongest prior selector at substantially lower scoring cost, with ablations isolating the calibrated heads and the retrieved-content signal.

2 RELATED WORK

Long-context training data selection. Beyond positional extension and continual-pretraining strategies (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024; Fu et al., 2024; Gao et al., 2025b), recent work increasingly targets long-context training data with useful long-range dependencies. LongWanjuan (Liu et al., 2024b) evaluates general long-text quality such as coherence and cohesion, while ProLong (Chen et al., 2024) and LongPPL (Fang et al., 2025) characterize dependencies through predictive changes. LongFilter (Deng et al., 2026) further turns predictive contrast into a document selector. A complementary line scores long-range dependency from attention, including token-level distant attention in LongAttn (Wu et al., 2025a) and span-level aggregation in LADM (Chen et al., 2025). This leaves a gap between the direct predictive signals used to measure context utility and the scalable attention signals needed for corpus-scale selection.

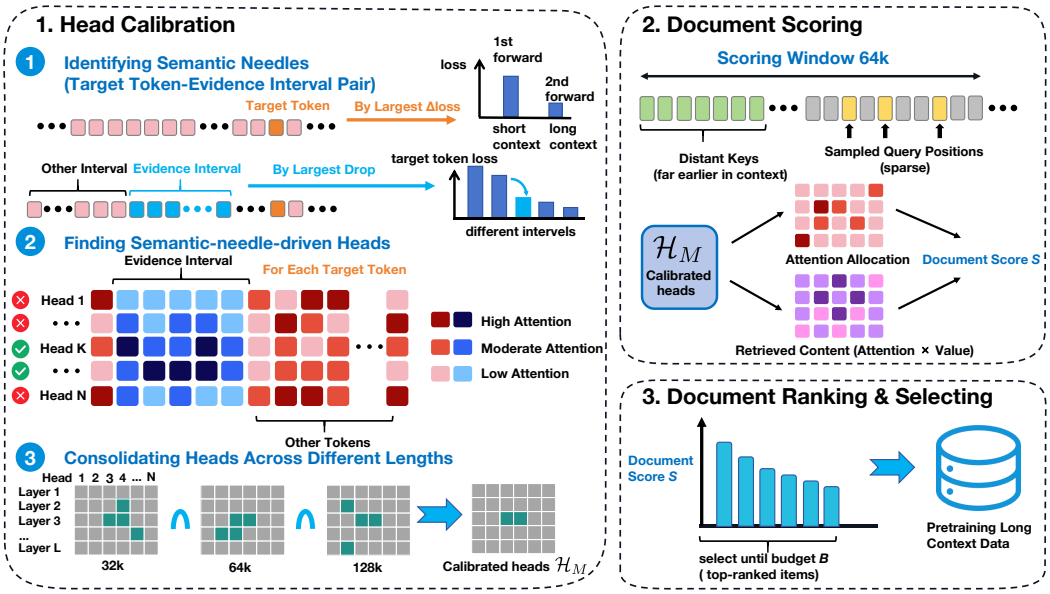

Figure 1: Overview of SNAS: (1) calibrate the semantic long-range dependency heads $\mathcal{H}_M$ from natural semantic needles; (2) one forward pass per document reads out distant attention (a) and retrieved content (v) at those heads; (3) rank-fuse them into a score S and select under the budget B (Sections 3.2–3.3).

Specialized attention heads. Attention heads exhibit functional specialization: induction heads capture matching-and-copying patterns (Olsson et al., 2022), while long-context retrieval often concentrates in a small subset of retrieval heads (Wu et al., 2025b). Subsequent work identifies heads using natural or query-conditioned retrieval signals (Zhang et al., 2025; Zhao et al., 2024) and applies capability-specific heads to data valuation (Hua et al., 2025; Liu et al., 2025). However, existing head-identification methods are largely defined through synthetic retrieval probes or task-specific query–evidence relations rather than predictive dependencies naturally occurring in unlabeled long documents. Thus, how to identify attention heads that reliably capture such dependencies and leverage them for long-context data selection remains an open problem.

3 METHODOLOGY

SNAS follows a three-stage pipeline (Figure 1). First, *head calibration*: on a small calibration corpus, SNAS identifies model-specific heads that respond to naturally occurring long-range dependencies. Second, *document scoring*: for each candidate document, it computes a long-range dependency score from these heads in a single forward pass. Third, *ranking and selection*: it ranks documents by this score and keeps the top-ranked ones within the token budget.

3.1 Problem Formulation

Let $\mathcal{D} = \{d_i\}_{i=1}^{N}$ denote a pool of long documents. Long-context data selection methods assign each candidate document a scalar score

$$s_i = S(d_i), \tag{1}$$

which induces a ranking over the candidate pool. A higher score indicates stronger long-range dependency. Our goal is to select a subset $\mathcal{S} \subseteq \mathcal{D}$ of high-scoring documents, up to a token budget B, for continual pre-training.

SNAS instantiates this score on a model-specific set of semantic long-range dependency heads:

$$s_i^{\mathrm{SNAS}} = G\big(\{\Phi(d_i; M, h)\}_{h \in \mathcal{H}_M}\big), \tag{2}$$

where M is the scoring model, $\mathcal{H}_M$ the calibrated head set (Section 3.2), $\Phi(d_i; M, h)$ the attention-derived signals from head h for d_i, and G aggregates them into the document score (Section 3.3).

3.2 Calibrating Semantic Long-Range Dependency Heads

Semantic needles in natural text. Conventional needle-based probes insert a synthetic target at a known location and test whether the model retrieves it. In contrast, we define a *semantic needle* as a naturally occurring target–evidence pair

$$\mathcal{N}_t = (x_t, \mathcal{E}_t), \tag{3}$$

where x_t is a target token whose prediction substantially improves when distant context is available, and $\mathcal{E}_t$ is the main distance interval of the prior context responsible for this improvement.

Context-dependent target tokens. Following LongPPL (Fang et al., 2025), we contrast long- and short-context likelihoods to identify target tokens whose predictions depend on distant context. For a calibration document $x_{1:T}$ and a position t, let

$$\ell_{\text{full}}(t) = -\log p_M(x_t \mid x_{<t}), \tag{4}$$

$$\ell_{\text{short}}(t) = -\log p_M(x_t \mid x_{t-L_s:t-1}), \tag{5}$$

where L_s denotes the length of the short-context. We retain x_t as a context-dependent target when

$$\ell_{\text{short}}(t) - \ell_{\text{full}}(t) > \alpha \quad \text{and} \quad \ell_{\text{full}}(t) < \beta, \tag{6}$$

for thresholds $\alpha, \beta > 0$. The first condition requires the token's prediction to improve markedly with the full context, while the second keeps it predictable once that context is given, discarding candidates whose apparent gain reflects sheer difficulty rather than genuine long-range dependence. Each x_t provides the target component of the semantic needle $\mathcal{N}_t$.

Distant-evidence intervals. For each retained target x_t, we identify the distance interval of the prior context that accounts for its predictive gain. Let $L_0 = L_s < L_1 < \cdots < L_K$ be a sequence of left-context lengths, indexed by $k = 0, \ldots, K$ and starting from the short-context window L_s, each reaching farther back from x_t than the previous one, and let

$$c_t(L_k) = -\log p_M(x_t \mid x_{t-L_k:t-1}). \tag{7}$$

We assign x_t to the interval with the largest reduction in negative log-likelihood,

$$k_t = \arg\max_{0 \le k \le K} \left[c_t(L_k) - c_t(L_{k+1}) \right], \tag{8}$$

and set the evidence interval $\mathcal{E}_t = [L_{k_t}, L_{k_t+1})$ (with L_{K+1} the full context length, so the last interval reaches the document start). Geometric spacing gives a deliberately coarse, log-scale localization of the evidence. Targets whose largest reduction falls below a minimum threshold are discarded, and each retained $\mathcal{E}_t$ completes the semantic needle $\mathcal{N}_t$. Since the ladder runs from $L_0 = L_s$ to the full context, the interval-wise reductions telescope to the long–short prediction gain of Eq. (6).

Semantic-needle-driven head scoring. Inspired by retrieval heads (Wu et al., 2025b), to identify heads important for long-range dependence, we relate the semantic needle to attention in the same way: for each target token, we measure the overlap between a head's attention distribution and its evidence interval $\mathcal{E}_t$. However, $\mathcal{E}_t$ is a distance interval, not a single token, so a head with an intrinsic positional bias can score high on it by habit alone (Voita et al., 2019; Clark et al., 2019).

To remove this bias, we score each head by how much its attention in $\mathcal{E}_t$ exceeds its own background profile, not by raw overlap. This profile is estimated from a control set $\mathcal{C}_d$ of ordinary, positionally comparable non-target positions in document d, needing no negative documents or task labels.

Let $A_h(t, j)$ be the attention weight from query t to key j for head h. To reduce the contribution of initial attention sinks, we exclude the first 128 key positions, which remain in the softmax normalization. For a distance interval b, let $\mathcal{K}_t(b) = \{\, j : t - j \in b,\ j \ge 128 \,\}$ be the set of non-sink keys lying at distance b from query t. Its attention mass and key count are

$$m_h(t, b) = \sum_{j \in \mathcal{K}_t(b)} A_h(t, j), \qquad n_t(b) = \left|\mathcal{K}_t(b)\right|, \tag{9}$$

From the control set we estimate $P_{h,d}(b)$, head h's per-key background attention on distance interval b, averaged over the control positions in document d:

$$P_{h,d}(b) = \frac{\sum_{c \in \mathcal{C}_d} m_h(c, b)}{\sum_{c \in \mathcal{C}_d} n_c(b)}. \tag{10}$$

The enrichment of a head at a target x_t compares its actual attention on $\mathcal{E}_t$ with the mass expected under this background:

$$e_h(t) = \log \frac{m_h(t, \mathcal{E}_t)}{P_{h,d}(\mathcal{E}_t)\, n_t(\mathcal{E}_t)}, \tag{11}$$

Writing $\mathcal{T}_d$ for the targets retained in document d and $\mathcal{D}_{\text{cal}}$ for the calibration corpus, we average $e_h(t)$ over the semantic needles of each document and then over documents, which gives head h's *enrichment score*

$$E_h = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{d \in \mathcal{D}_{\text{cal}}} \frac{1}{|\mathcal{T}_d|} \sum_{t \in \mathcal{T}_d} e_h(t), \tag{12}$$

where two-stage averaging prevents documents with more targets from dominating the ranking.

Multi-length head selection. To identify heads that capture long-range dependence robustly across scales rather than at a single length, we repeat the calibration above independently at 32K, 64K, and 128K context lengths, each on its own length-matched corpus $\mathcal{D}_{\text{cal}}$. Within each tier, heads are ranked by the enrichment score E_h of Eq. (12) computed on that corpus, and the top-κ heads are retained. The final model-specific head set $\mathcal{H}_M$ is their strict intersection

$$\mathcal{H}_M = \mathcal{H}_{M,\kappa}^{32\text{K}} \cap \mathcal{H}_{M,\kappa}^{64\text{K}} \cap \mathcal{H}_{M,\kappa}^{128\text{K}}. \tag{13}$$

3.3 Document Scoring and Selection

We turn each document d into a scalar score in three steps: select query positions and their distant keys, compute two document-level signals, then rank-fuse them. To keep scoring tractable on large-scale long-document corpora, we compute this score on the first 64K tokens of each document and take this prefix as a proxy for the whole document; shorter documents are scored in full.

Query selection. Long-range dependence is sparse in both the query and the key dimension, so both signals restrict attention to a sparse set of queries and their distant context rather than aggregating over all positions. The query set $\mathcal{Q}$ consists of 2048 tokens at equal intervals, from token 4096 to the end of the scoring window; each signal reads its distant keys as specified below.

Attention allocation (a). The first signal is based on LongAttn's long-range dependency statistic (Wu et al., 2025a), computed on the calibrated heads $\mathcal{H}_M$ and over the query positions in $\mathcal{Q}$. Keeping this component identical to the strongest attention-based baseline isolates our contribution to the calibrated heads and the second signal below. For each head h it gives a proportion $\bar{a}_h(d)$ of attention mass on long-distance context and the variance $\sigma_h^2(d)$ of that allocation. We average each over $\mathcal{H}_M$ and, since the two lie on different scales, min-max normalize each over the candidate pool $\mathcal{D}$, giving pool-relative $\widetilde{a}(d)$ and $\widetilde{\sigma^2}(d)$; the signal is

$$a(d) = \widetilde{a}(d) - \tfrac{1}{2}\,\widetilde{\sigma^2}(d). \tag{14}$$

This rewards heads that consistently allocate substantial attention to distant context while penalizing highly variable allocation.

Retrieved content (v). For each query $q \in \mathcal{Q}$ and head $h \in \mathcal{H}_M$, its *distant* keys are the positions more than 4096 tokens before q. The second signal measures how much content this distant attention actually retrieves: grouping the distant keys into geometric distance bands $\{b\}$, we sum the norm of the value content retrieved within each band and discount the initial-sink attention,

$$r_h(q) = \Big(\sum_b \big\| \textstyle\sum_{j \in b} A_h(q, j)\, \mathbf{z}_j \big\| \Big) \big(1 - \text{sink}_h(q)\big), \tag{15}$$

where $\mathbf{z}_j$ is the value vector at key j and $\text{sink}_h(q)$ is the fraction of q's distant attention that falls on the first key block (the initial 512 tokens). We reduce r_h to a per-head document score by averaging its top 1% of queries, $u_h(d) = \text{mean}_{q \in \mathcal{Q}_h^{1\%}} r_h(q)$, and average over heads,

$$v(d) = \frac{1}{|\mathcal{H}_M|} \sum_{h \in \mathcal{H}_M} u_h(d). \tag{16}$$

Keeping each head's strongest queries reflects sparse long-range dependence within a document.

Fusion and selection. Because $a(d)$ and $v(d)$ still lie on different scales, we combine them by rank fusion over the candidate pool $\mathcal{D}$,

$$S(d) = \mathrm{rank}\big(a(d)\big) + \mathrm{rank}\big(v(d)\big), \tag{17}$$

with $\mathrm{rank}(\cdot) \in (0, 1]$ the normalized rank across $\mathcal{D}$. We select whole documents in descending $S(d)$ order until the token budget B is reached. A selected document contributes all of its tokens to the budget. For documents longer than 64K, we score only the first 64K tokens, using this window as a proxy for the whole document's long-range quality.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Head selection. We calibrate model-specific head sets for Llama-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B-Base (Yang et al., 2025), the latter extended beyond its native 32K with YaRN (Peng et al., 2024). Calibration uses Common Crawl, arXiv, and Books documents at 32K/64K/128K ($\approx 400/400/200$ retained, ≤ 48 target tokens each). For target detection (Eq. 6) we set $L_s = 4096$, $\alpha = \beta = 2$, and drop targets whose largest per-band NLL reduction is under 1 nat. The evidence ladder doubles across $\{4096, 8192, 16384, 32768\}$ at 32K/64K, adding 65536 at 128K, with an open final band to full context. For the background profile, attention is binned by doubling distance to the tier length, and up to 128 control positions per document are drawn from 8 position bins, excluding ± 16 tokens around each target (Eq. 11). We keep the top $\kappa = 50$ heads per tier by enrichment and take their strict three-tier intersection as $\mathcal{H}_M$ (Eq. 13).

Data selection and cpt. Following LongFilter (Deng et al., 2026), we draw a $\sim$10B-token SlimPajama (Shen et al., 2023) candidate pool per domain, at document lengths of 32K (Common Crawl), 64K (arXiv), and 128K (Books). Within each domain, all selectors use the same budget: each selects whole documents from its pool by descending score to 1.6B tokens, then combines them with a shared 0.4B-token set of same-domain short documents (<32K) to form the same 2B-token 8:2 mixture. Selection and training deliberately use different models. Documents are scored with two backbones, Llama-3.1-8B and Qwen3-8B-Base, while every 2B set is used to continually pre-train the same Llama-3-8B-Base. We extend the training context to 128K with adjusted RoPE base frequency (ABF) (Su et al., 2024; Xiong et al., 2024; Gao et al., 2025a). Before evaluation, all models are instruction-tuned on UltraChat-200K (Ding et al., 2023), and we repeat the training over three seeds and report the mean. Full details are in Appendix A.4.

Baselines and evaluation. We compare against the prediction-based LongFilter (Deng et al., 2026) and the attention-based LongAttn (Wu et al., 2025a) and LADM (Chen et al., 2025). All baselines use the same scoring backbone as SNAS. For LADM, we additionally report a variant using its own separately trained attention calculator. We evaluate on HELMET* (Yen et al., 2025) at 32K/64K/128K, RULER (Hsieh et al., 2024) from 4K to 128K, LongBench v2 (Bai et al., 2025), and MRCR-v2 (OpenAI, 2025). HELMET* retains representative rule-scored tasks while excluding LLM-judged free-form tasks. RULER uses variable-tracking and word-extraction tasks while omitting NIAH and QA to avoid overlap with HELMET*. Full evaluation details are in Appendix B.

4.2 THE CALIBRATED HEAD SET

Before turning to downstream results, we first examine the calibrated head set. For Llama-3.1-8B, the three-tier intersection retains only 30 of 1024 heads (2.9%), with 27 concentrated in layers L9–L16 (Figure 2). The 44 synthetic retrieval heads (Wu et al., 2025b), by contrast, spread more broadly across layers and overlap $\mathcal{H}_M$ in only 9 heads (Jaccard 0.14). Semantic-needle calibration thus identifies a related but distinct subset of long-range attention heads.

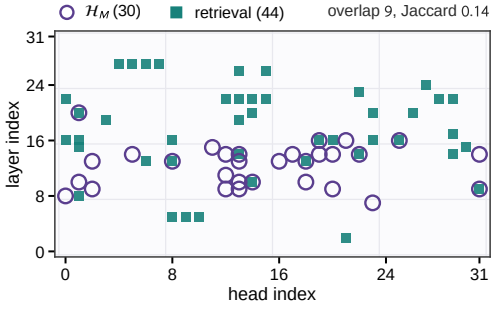

Figure 2: Llama-3.1-8B heads: $\mathcal{H}_M$ (violet circles) vs. retrieval heads (teal squares).

Table 1: **Main results.** Column groups indicate the scoring model used for selection (§4.1); the last block averages over the three domains. Best per column within each block in **bold**, second best underlined. LADM†: LADM with its own attention calculator. Means over three runs.

Method	Scorer: Llama-3.1-8B					Scorer: Qwen3-8B-Base				
	HELMET*	LB v2	RULER	MRCR-v2	Ovr.	HELMET*	LB v2	RULER	MRCR-v2	Ovr.
Common Crawl										
LongFilter	**65.40**	25.30	74.57	**19.90**	46.29	**65.05**	26.40	74.12	**17.95**	45.88
LongAttn	57.52	24.20	**76.52**	12.36	42.65	61.65	26.40	**75.48**	13.57	44.28
LADM	65.02	25.50	74.21	18.79	45.88	62.16	24.90	74.79	16.38	44.56
LADM†	63.33	24.80	76.16	17.28	45.39	–	–	–	–	–
SNAS (ours)	64.88	**25.95**	75.81	19.65	**46.57**	64.42	**27.00**	74.51	17.92	**45.96**
arXiv										
LongFilter	**64.54**	25.95	**74.04**	**10.15**	43.67	**65.07**	26.70	72.81	7.68	43.07
LongAttn	54.17	24.40	71.97	7.45	39.50	61.72	26.70	73.50	**8.20**	42.53
LADM	59.82	**27.60**	73.80	9.62	42.71	58.59	23.40	72.72	7.44	40.54
LADM†	60.88	27.10	72.91	7.55	42.11	–	–	–	–	–
SNAS (ours)	64.47	**27.60**	73.60	9.43	**43.77**	64.93	**27.30**	**75.40**	7.08	**43.68**
Books										
LongFilter	64.92	27.15	**76.09**	17.45	46.40	64.91	**28.00**	74.44	**19.23**	46.65
LongAttn	64.63	25.80	74.07	15.95	45.12	64.30	25.90	75.59	17.87	45.92
LADM	64.65	25.40	75.33	16.32	45.43	63.83	24.90	74.66	15.76	44.79
LADM†	62.43	24.60	74.74	13.52	43.82	–	–	–	–	–
SNAS (ours)	**65.55**	**27.70**	76.05	**19.03**	**47.08**	**65.33**	26.20	**76.47**	**19.23**	**46.81**
Average over the three domains										
LongFilter	64.95	26.13	74.90	15.83	45.45	**65.01**	**27.03**	73.79	**14.95**	45.20
LongAttn	58.77	24.80	74.19	11.92	42.42	62.56	26.33	74.86	13.21	44.24
LADM	63.16	26.17	74.45	14.91	44.67	61.53	24.40	74.06	13.19	43.30
LADM†	62.21	25.50	74.60	12.78	43.77	–	–	–	–	–
SNAS (ours)	**64.97**	**27.08**	**75.15**	**16.04**	**45.81**	64.89	26.83	**75.46**	14.74	**45.48**

4.3 Main Results

Overall performance. Table 1 shows that SNAS matches or numerically exceeds the strongest baseline in Ovr. across all six domain–scorer settings. With Llama-3.1-8B as the scorer, it scores 46.57, 43.77, and 47.08 on Common Crawl, arXiv, and Books, respectively, yielding a three-domain mean of 45.81 versus 45.45 for LongFilter, the strongest baseline. SNAS also has the numerically highest three-domain mean on each of the four evaluation suites, though several margins are small. With Qwen3-8B-Base as the scorer, it leads in all three domains and averages 45.48 Ovr. versus 45.20 for LongFilter. Across both scoring models, SNAS thus delivers selection quality on par with the strongest baseline, with numerical leads in every evaluated domain.

Domain-level results. The largest Ovr. margin over LongFilter occurs on Books with the Llama scorer (47.08 vs. 46.40), where SNAS also scores higher on MRCR-v2 (19.03 vs. 17.45). With Qwen as the scorer, SNAS leads on arXiv (43.68 vs. 43.07), including higher LB v2 and RULER scores, but trails LongFilter on HELMET* and MRCR-v2 in that domain. These results support a claim of first-tier selection quality, rather than a substantial accuracy advantage over LongFilter; the controlled ablations below examine the contributions of head selection and score fusion.

4.4 Scoring Efficiency

We measure corpus-wide scoring cost in terms of throughput and GPU-hours using the same Llama-3.1-8B scorer. These measurements cover corpus scoring and exclude the one-time head-calibration stage, which uses roughly 1,000 retained documents per scorer and is amortized over the much larger ∼10B-token candidate pool. Although SNAS, LongAttn, and LADM each require only one forward pass per document, they differ in the amount of attention information they process. LongAttn is fastest because it reads only first-layer attention. SNAS restricts extraction to calibrated heads and sampled queries. LADM is slowest because it recomputes sliding-window statistics across all layers and heads. LongFilter, in contrast, requires three forward passes per document.

We compare SNAS against two reference points. Against LongFilter, the closest baseline in quality (45.45 vs. 45.81 Ovr.), SNAS reaches a comparable score while scoring $1.8\times$ faster with 44% fewer GPU-hours. Against LongAttn, the fastest baseline, SNAS raises Ovr. from 42.42 to 45.81 at some cost in throughput. SNAS thus offers a favorable quality–efficiency trade-off for corpus-scale selection, with LongAttn remaining the fastest overall.

Table 2: Quality vs. scoring cost. Ovr. is the three-domain mean; cost on Common Crawl.

Method	Ovr.	doc/s	GPU-h
LongFilter	45.45	0.125	348.10
LongAttn	42.42	0.484	90.10
LADM	44.67	0.064	678.60
SNAS (ours)	**45.81**	0.225	194.20

4.5 Ablations

Table 3: Head-selection ablation on Common Crawl with Llama-3.1-8B. All rows share the full SNAS scoring rule and the same budget. Best per column in **bold**.

Head set	Num. Heads	HELMET*	LB v2	RULER	MRCR-v2	Ovr.
Random heads	30	63.99	25.10	73.59	15.76	44.61
Retrieval heads	44	64.43	**26.10**	74.86	16.01	45.35
First-layer heads	32	61.64	25.00	**76.91**	15.76	44.83
SNAS heads	30	**64.88**	25.95	75.81	**19.65**	**46.57**

Head selection. Table 3 isolates the role of head selection by replacing $\mathcal{H}_M$ with random, retrieval, or first-layer heads while keeping the SNAS scoring rule and data budget fixed. Under this controlled comparison, the calibrated SNAS heads achieve the highest overall performance (46.57 Ovr.), ahead of retrieval heads (45.35), while alternative head sets remain competitive on individual suites: retrieval heads are best on LB v2, and first-layer heads are best on RULER. The advantage of SNAS heads is most pronounced on MRCR-v2 (19.65 vs. 16.01 for retrieval heads and 15.76 for random and first-layer heads), which is consistent with our calibration objective: $\mathcal{H}_M$ is selected for enrichment on naturally occurring evidence intervals that improve token prediction, rather than for synthetic copy retrieval or layer identity alone. Together with the limited overlap between calibrated and retrieval heads (Section 4.2), these results suggest that semantic-needle calibration identifies a partially overlapping but distinct subset of heads better suited to long-context data selection.

Table 4: Scoring ablation on Common Crawl (Llama-3.1-8B, $\mathcal{H}_M$ fixed).

Scoring rule	HELMET*	LB v2	RULER	MRCR-v2	Ovr.
a only (LongAttn on $\mathcal{H}_M$)	63.95	**27.20**	75.21	17.60	45.99
$a + v$ **(SNAS, full)**	**64.88**	25.95	**75.81**	**19.65**	**46.57**

Scoring variants. Since v is computed from the same single forward pass and calibrated heads used by a, we test whether it adds anything beyond attention allocation: holding $\mathcal{H}_M$ and the budget fixed, we compare a alone against the full score $a + v$ (Table 4). Adding v raises the overall score from 45.99 to 46.57, most clearly on MRCR-v2 (17.60 to 19.65). This is consistent with how v is designed: a sums attention *mass*, while v takes the norm of the attention-weighted *value* and downweights attention sinks, rewarding documents whose distant attention actually assembles substantive content, as multi-round co-reference requires. a alone is higher only on LongBench v2 (27.20 vs. 25.95), but the fused score is better overall, so we adopt $a + v$ as the final scoring rule.

4.6 Robustness to Document Length

Because selection scores can correlate with length, differences between methods could reflect the length of the documents they select rather than their long-range dependencies. We therefore build a fixed-length Common Crawl pool by truncating documents of at least 64K to exactly 64K. SNAS still leads at 44.48 Ovr., compared with 44.06 for LongAttn and 43.71 for LongFilter (Table 5), underscoring the strength of its calibrated, content-aware scoring even in this stricter comparison.

Under fixed length, LongAttn is the only method whose score rises (from 42.65 to 44.06 Ovr. on Common Crawl, Table 1), while every other selector drops. This suggests that LongAttn's length-normalized rule is better suited to scoring uniformly long inputs than to selecting a top-scoring subset from a variable-length pool.

Table 5: Fixed-length 64K comparison on Common Crawl with Llama-3.1-8B.

Method	HELMET*	LB v2	RULER	MRCR-v2	Ovr.
LongFilter	60.99	**24.60**	75.85	13.41	43.71
LongAttn	**61.72**	24.20	77.17	13.16	44.06
LADM	60.80	24.20	74.40	14.96	43.59
SNAS (ours)	61.06	**24.60**	**77.20**	**15.05**	**44.48**

4.7 Head Stability and Generality

Stability. To test whether calibration depends on the specific documents, we recompute the per-tier top-50 heads using a disjoint calibration split of the same size ($\sim$400/400/200 documents at 32K/64K/128K). The resulting head sets are stable: per-tier Jaccard overlap ranges from 0.75 to 0.92, and the final three-tier intersection $\mathcal{H}_M$ reaches 0.84 for Llama-3.1-8B and 0.75 for Qwen3-8B (Table 6). Even the lowest value (0.75) corresponds to about 86% of heads shared between disjoint splits, far above the near-zero overlap expected by chance. This indicates that the calibrated heads are largely insensitive to the particular calibration sample.

Table 6: Head-set stability across disjoint splits.

Scorer	32K	64K	128K	$\mathcal{H}_M$
Qwen3-8B	0.85	0.92	0.92	0.75
Llama-3.1-8B	0.75	0.92	0.92	0.84

Cross-model generality. Beyond the two main scoring models, we repeat calibration on two auxiliary architectures, Mistral-Nemo-12B and Qwen3.5-9B-Base, for testing the generalization of SNAS. Qwen3.5-9B-Base uses hybrid attention, so we calibrate only its eight full-attention layers. Table 7 shows that SNAS remains close to the strongest baseline across scoring models of different sizes and architectures: it slightly exceeds LongFilter with Mistral-Nemo-12B (46.18 vs. 45.96 Ovr.) and trails it slightly with Qwen3.5-9B-Base (46.33 vs. 46.54). This suggests that the calibration-and-selection procedure extends beyond the two main scoring models when recalibrated for each model.

Table 7: Cross-model generality on Common Crawl with auxiliary scoring models; the downstream model is a fixed Llama-3-8B-Base. Best score per column within each scorer in **bold**.

Scorer	Method	HELMET*	LB v2	RULER	MRCR-v2	Ovr.
Mistral-Nemo-12B	LongFilter	**64.73**	**27.30**	75.31	16.50	45.96
	SNAS	64.60	25.70	**75.67**	**18.76**	**46.18**
Qwen3.5-9B-Base	LongFilter	**65.02**	**26.50**	73.26	**21.40**	**46.54**
	SNAS	64.49	26.40	**75.59**	18.83	46.33

5 Conclusion

We presented SNAS, which applies prediction-level contrast only on a small calibration corpus to mine natural semantic needles and calibrate model-specific semantic long-range dependency heads, and then scores the full corpus with a single traversal per document using both distant-attention allocation and retrieved content. Under an identical continual-pretraining protocol, SNAS matches the strongest prior selector's average quality across Common Crawl, arXiv, and Books at substantially lower scoring cost, and holds up on a fixed-length pool, with ablations validating its main components.

AI USE STATEMENT

Generative AI tools were used to assist with writing and implementation. They were used to improve the clarity and readability of author-written text and to accelerate routine programming tasks such as code completion, debugging, and refactoring. The research ideas, methodological design, experimental protocol, and interpretation of results were developed by the authors. All AI-assisted text and code were reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the calibration, data-selection, training, and evaluation procedures in the main text and appendices, including the relevant hyper-parameters, data construction, sequence packing, and evaluation protocols. All main results use matched candidate pools, budgets, and training configurations and are averaged over three random seeds. The anonymous supplementary material contains the core code for semantic-needle mining, head calibration, document scoring, and data selection. Upon publication, we intend to publicly release the corresponding code.

REFERENCES

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. Ladm: Long-context training data selection with attention-based dependency measurement for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3076–3090, 2025.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, 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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8222–8234, 2024.

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 276–286, 2019.

Haoran Deng, Yingyu Lin, Zhenghao Lin, Xiao Liu, Yizhou Sun, Yian Ma, and Yeyun Gong. Beyond length: Quantifying long-range information for long-context llm pretraining data. 2026: 12983–12995, 2026.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, 2023.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. 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*, volume 2025, pp. 94541–94563, 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. 2024.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. Nextlong: Toward effective long-context training without long documents. 2025a.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7376–7399, 2025b.

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

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

Kai Hua, Steven Wu, Ge Zhang, and Ke Shen. Attentioninfluence: Adopting attention head influence for weak-to-strong pretraining data selection. *arXiv preprint arXiv:2505.07293*, 2025.

Mario Michael Krell, Matej Kosec, Sergio P Perez, and Andrew Fitzgibbon. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv preprint arXiv:2107.02027*, 2021.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. volume 37, pp. 22947–22970, 2024.

Jinrui Liu, Jeff Wu, Xuanguang Pan, Gavin Cheung, Shuai Ma, and Chongyang Tao. Air: Post-training data selection for reasoning via attention head influence. *arXiv preprint arXiv:2512.13279*, 2025.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. volume 12, pp. 157–173, 2024a.

Xiaoran Liu, Kai Lv, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. Longwanjuan: Towards systematic measurement for long text quality. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5709–5725, 2024b.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

OpenAI. OpenAI MRCR: Long context multiple needle in a haystack benchmark. Hugging Face Datasets, 2025. URL `https://huggingface.co/datasets/openai/mrcr`. MRCR stands for multi-round co-reference resolution.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. Slimpajama-dc: Understanding data combinations for llm training, 2023.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

Longyun Wu, Dawei Zhu, Guangxiang Zhao, Zhuocheng Yu, Junfeng Ran, Xiangyu Wong, Lin Sun, and Sujian Li. Longattn: Selecting long-context training data via token-level attention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19367–19380, 2025a.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025b.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 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 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*, volume 2025, pp. 98914–98965, 2025.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23802–23816, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. volume 36, pp. 34661–34710, 2023.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding synthetic context extension via retrieval heads. 2024.

A Training Setup and Design-Choice Ablations

This appendix is organised around one question: which of our training decisions actually matter, and by how much? We first fix what we average (Appendix A.1), then describe how the corpus is built and packed (Appendix A.2), settle how many tokens to spend (Appendix A.3), and close with the full hyper-parameter listing (Appendix A.4).

A.1 Composite Scores

As in the main paper, every result in this appendix reports the four long-context suites—HELMET*, LongBench v2, RULER and MRCR-v2—together with their unweighted mean *Ovr.*

A.2 Corpus Construction and Sequence Packing

Mixture. The continued-pretraining corpus is drawn from three SlimPajama domains—ArXiv, Book and CommonCrawl—and each domain is treated as an independent training set. The long side reuses the LongFilter candidate pools, which total $\sim$10B tokens of long documents across the three domains (Section 4.1); the short side is one fixed set of same-domain documents under 32K tokens per domain.

What varies and what does not. Selection acts *only* on the long pool: a method ranks the long candidate pool of a domain and keeps the top 1.6B. The short side is shared by every method and frozen at 0.4B tokens per domain. Long and short are then combined at 8:2 into a 2B-token per-domain training set, mixed and packed once, and consumed in the same order by every run. The only variable across runs is thus which long documents enter training, which is what makes the comparison in the main paper an attribution to the selection method rather than to the data recipe.

Routing by length. Every document is tokenized and wrapped with the model's BOS/EOS tokens, then routed by length. Documents longer than the 128K window are cut into consecutive full-window blocks, each of which becomes a training sequence on its own; the trailing remainder re-enters the packing pool if it still holds at least $10{,}240$ tokens and is discarded otherwise. Everything no longer than the window, plus the recycled remainders, forms the packing pool.

Packing. The pool is packed into $131{,}072$-token (128K) sequences by Best-Fit-Decreasing (BFD) packing (Krell et al., 2021): documents are visited in decreasing length order and each is placed in the bin with the smallest remaining capacity that can still hold it, a new bin being opened only when none can. BFD never splits a document in order to fill a bin—it prefers to leave padding—and still reaches a token fill rate above 0.94 in every domain.

Independence inside a packed sequence. Co-packed documents must not see each other, so position IDs restart at zero for every document and attention is restricted by a block-diagonal mask; a token attends only to earlier tokens of the *same* document. This is the document-masked packing of ProLong (Gao et al., 2025b), and it prevents positional and attention signal from leaking across document boundaries—which matters especially here, since a long-context run whose packed neighbours bleed into each other would be rewarded for the wrong behaviour. For context parallelism each document is padded to a multiple of $2 \times \mathrm{CP} = 8$ tokens with the end-of-sequence token; the loss label of every padding position is set to -100, as is the label of the first token of each document, which removes the boundary loss induced by the standard causal label shift.

A.3 How Many Tokens Should Be Spent? (2B per Domain)

Setup. Every main-paper model is continued-pretrained on 2B tokens *per domain* (three independent per-domain models whose scores are then averaged), so we sweep the per-domain budget from 1B to 5B at the fixed 8:2 mixture, instruction-tune every checkpoint identically, and evaluate (Table 8). The 2B point of the sweep is thus exactly the training maturity of all main runs.

One step dominates. The $1 \rightarrow 2$B step is the largest single movement anywhere in the sweep: $+1.87$ Ovr. At 1B the model is simply under-trained, and the long-context behaviour that the continued-pretraining stage is supposed to install has not yet appeared (MRCR-v2 stands at 20.70 versus 26.09 at 2B).

Everything after it is marginal. Past 2B the curve flattens. The individual steps are -0.94, $+0.86$ and $+0.43$ Ovr., and the *entire* 2B$\rightarrow$5B tail—more than doubling the compute—is worth only $+0.35$ Ovr., less than a quarter of what the first step delivers on its own. This is precisely the argument for stopping at 2B: it captures almost all of the attainable gain at two fifths of the cost, and it avoids over-training on a fixed long-document pool, which would compress the very differences between selection methods that the main experiments are meant to expose.

A flat composite hides two opposing trends. The per-task columns are worth reading, because the plateau is not a plateau per task. HELMET* and RULER continue to improve slowly with budget ($61.39 \rightarrow 63.38$ and $76.20 \rightarrow 78.58$), while LongBench v2 and MRCR-v2 peak at 2B and then give back 1.10 and 1.88 points. More tokens of the same long pool keep sharpening long-range *retrieval* while slowly eroding the dispersed-evidence skills, and the two effects cancel in the composite. This is another reason not to buy the extra tokens.

A.4 Hyper-parameters

Our base models are continually pretrained from their native context length to 128K and then instruction-tuned on UltraChat-200K with prompt-token masking (loss on responses only). Both stages run in MS-Swift with a Megatron backend, and both use the same optimiser family, schedule and parallel layout, so we list them side by side in Table 9; entries that do not apply to a stage are marked "—". Every number reported in this appendix and in the main paper uses this configuration, with the single exception of the ablation being studied in the corresponding subsection. Throughput and wall-clock time are measured on one representative Common Crawl run and count all packed sequence slots, including padding.

Table 8: Continued-pretraining token budget on our Llama-3-8B base under the 128K CPT+SFT pipeline (the budget is the only variable). Arrows give the change in Ovr. relative to the previous budget (↑ gain, ↓ loss); best per column in **bold**. The $1 \rightarrow 2$B step gains $+1.87$ Ovr., whereas the whole 2B→5B tail adds only $+0.35$.

CPT budget	HELMET*	LB v2	RULER	MRCR-v2	Ovr.
1B	61.70	25.10	76.12	20.70	45.90
2B (ours)	61.39	**27.40**	76.20	**26.09**	47.77 ↑1.87
3B	62.43	25.20	78.40	21.30	46.83 ↓0.94
4B	62.48	26.00	**78.68**	23.61	47.69 ↑0.86
5B	**63.38**	26.30	78.58	24.21	**48.12** ↑0.43

Table 9: Training hyper-parameters for both stages. All base models share one configuration, so a single column per stage suffices.

Hyper-parameter	Continued pretraining	Instruction tuning
Training data	SlimPajama (3 domains)	UltraChat-200K
Context length (train)	128K (packed to 135,168)	128K (packed to 132,096)
RoPE base θ	2×10^8 (ABF)	inherited
Budget	2B tokens / domain	1 epoch
Global batch size	8 seq (≈ 1M tok)	8 seq
Micro batch size	1	1
Peak learning rate	8×10^{-6}	8×10^{-6}
Min learning rate	8×10^{-7}	8×10^{-7}
LR schedule / warmup	cosine / 10%	cosine / 10%
Optimizer	AdamW (0.9, 0.95)	AdamW
Weight decay	0.1	—
Gradient clipping	1.0	—
Loss masking	padding + doc. BOS	prompt tokens
Precision	bfloat16	bfloat16
Attention backend	FlashAttention-3	FlashAttention-3
Activation recomputation	full	full
Parallelism (TP/CP/PP/DP)	2/4/1/4	2/4/1/4
Hardware	4× H800 (32 GPUs)	4× H800 (32 GPUs)
Throughput (tokens/s/GPU)	≈3.3K	≈7.1K
Wall-clock time / run	5.5 h	0.3 h
GPU-hours / run	≈175	≈10

B Evaluation Suite Details

We detail the four long-context suites underlying the results in Section 4.3. Every reported number comes from a deterministic, rule-based metric on a 0–100 scale; a suite score is the equally weighted mean over its constituent tasks, and *Ovr.* averages the four suite scores (Table 10).

HELMET*. HELMET* (Yen et al., 2025) denotes the nine HELMET tasks we run at 32K/64K/128K (Table 11). They span five capability groups—retrieval-augmented QA, passage re-ranking, citation-grounded answering, synthetic recall, and many-shot in-context learning—each scored by a deterministic, non-LLM-judge metric (SubEM, NDCG@10, STR-EM/Rec@5, recall, or accuracy). We skip HELMET's long-document QA and summarization tasks, whose official protocols grade free-form text with a model judge. The suite score is the mean over these nine tasks.

RULER. RULER (Hsieh et al., 2024) is a synthetic probe with fully controllable sequence length. We evaluate it from 4K to 128K on variable tracking and common/frequent-word extraction, and ex-

Table 10: The four evaluation suites used in this work. Each suite contributes a single aggregate score on a 0–100 scale, and *Ovr.* is their unweighted mean.

Suite	Context	Scoring	Capability probed
HELMET*	32/64/128K	SubEM, NDCG@10, recall, accuracy	Retrieval, re-ranking, citation, ICL
RULER	4K–128K	String-match accuracy	Variable tracking, word aggregation
LongBench v2	8K–2M w	Six-domain macro-accuracy	Realistic understanding & reasoning
MRCR-v2	16K–128K	String-similarity ratio	Multi-round needle disambiguation

Table 11: The nine HELMET* tasks retained in our evaluation, grouped by capability and reported at 32K/64K/128K.

Capability	Task	Metric	Probe
Retrieval-augmented QA	PopQA	SubEM	Entity-centric open-domain lookup
	JSON KV	SubEM	Value retrieval from a large JSON map
Passage re-ranking	MS MARCO	NDCG@10	Relevance ordering of retrieved passages
Citation	ALCE ASQA	STR-EM	Attributed answers to ambiguous questions
	ALCE QAMPARI	Rec@5	List answers with supporting citations
Synthetic recall	RULER NIAH MK-2	Recall	Numeric needle amid distractor keys
	RULER NIAH MK-3	Recall	UUID-valued needle amid distractor keys
	RULER NIAH MV	Recall	Several values tied to one key
Many-shot ICL	CLINC150	Accuracy	Intent labelling under many demos

clude its needle-in-a-haystack and QA families because HELMET* already exercises closely related retrieval skills; scores use RULER's string-match accuracy.

LongBench v2. LongBench v2 (Bai et al., 2025) contributes realistic, human-curated multiple-choice questions over inputs ranging up to millions of words. It spans single- and multi-document QA, long in-context learning, long-dialogue and code-repository understanding, and structured-data reasoning; its score is the unweighted mean of multiple-choice accuracy across the six domains.

MRCR-v2. MRCR-v2 (OpenAI, 2025) casts multi-round co-reference resolution as a long multi-turn dialogue in which the model must locate and reproduce the correct one among several near-identical needles. We use its 16K–128K split and measure the string-similarity ratio between the prediction and the reference.