FOCUSFT: BILEVEL OPTIMIZATION FOR DILUTION-AWARE LONG-CONTEXT FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can now process increasingly long inputs, yet their ability to effectively use information spread across long contexts remains limited. We trace this gap to how attention budget is spent during supervised fine-tuning (SFT) on long sequences: positional biases and attention sinks cause the model to allocate most of its attention to positionally privileged tokens rather than semantically relevant content. This *training-time attention dilution* (the starvation of content tokens in the attention distribution) weakens the gradient signal, limiting the model's ability to learn robust long-context capabilities. We introduce FOCUSFT, a bilevel optimization framework that addresses this problem at training time. An inner loop adapts lightweight fast-weight parameters on the training context to form a parametric memory that concentrates attention on relevant content, and the outer loop performs SFT conditioned on this sharpened representation. Both loops apply bidirectional attention over context tokens while preserving causal masking for responses, reducing the causal asymmetry that gives rise to attention sinks and aligning inner-outer behavior. On BABILong, FOCUSFT improves accuracy by up to +14pp across 4K–32K context lengths; on RULER, it raises CWE aggregation from 72.9% to 81.1% at 16K; and on GPQA with agentic tool use, it yields a 24% relative gain in pass@1. Attention analysis shows that FOCUSFT reduces attention sink mass by 529$\times$ and triples context engagement during training.

1 INTRODUCTION

Many applications of large language models (LLMs) rely on long-context capabilities: analyzing scientific corpora, synthesizing documents, maintaining coherent multi-turn dialogues, and reasoning over large code repositories (Jimenez et al., 2023; Bai et al., 2024b; Pei et al., 2025). Recent advances in positional encoding, distributed training, and architectural design have expanded context windows by orders of magnitude (Dubey et al., 2024; Liu et al., 2023b; Ding et al., 2024; Team et al., 2024). On the surface, the long-context problem appears largely solved: modern frontier models can ingest far more tokens than ever before.

However, a growing body of empirical evidence exposes a fundamental gap between context capacity and context utilization. RULER (Hsieh et al., 2024a) showed that many models scoring highly on simple needle-in-a-haystack retrieval suffer large performance drops as task complexity increases, with most models failing to maintain effective performance at their advertised context lengths. The "Lost in the Middle" phenomenon (Liu et al., 2024) revealed a U-shaped accuracy curve: LLMs attend well to the beginning and end of the input but systematically neglect content in the middle. These findings point to a common conclusion: a larger context window does not imply a larger reliable working memory. For instance, Qwen2.5-7B (Yang et al., 2024) supports 128K tokens, yet already struggles on reasoning tasks at 4K–32K, well within its native capacity. The problem is particularly acute in agentic settings, where the context comprises complex multi-turn dialogues including system prompts, user instructions, tool calls and their outputs, and prior assistant responses. In such scenarios, the model must attend to relevant information dispersed across structurally heterogeneous turns, making it especially vulnerable to positional biases.

The root causes are well studied at inference time: positional biases direct attention toward the beginning and end of the context (Liu et al., 2024; Hsieh et al., 2024b), and attention sinks consume a large share of the budget on a handful of initial tokens (Xiao et al., 2023). We refer to the resulting

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.

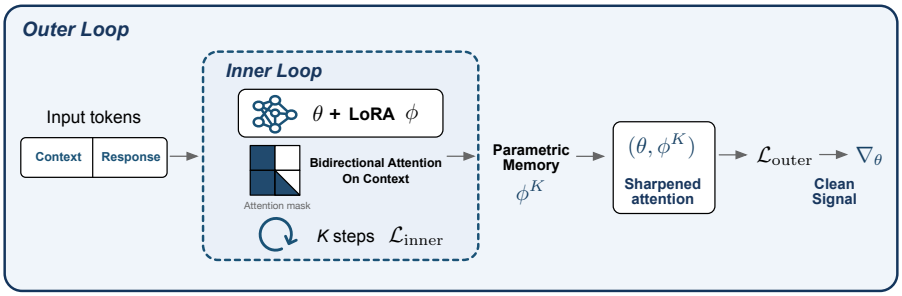


Figure 1: Overview of FOCUSFT. The inner loop (dashed box) adapts LoRA fast weights ϕ with bidirectional context attention to form parametric memory $\phi^{(K)}$. The outer loop performs response-token SFT conditioned on the sharpened representations, updating $\boldsymbol{\theta}$. Fast weights are discarded after each step; only $\boldsymbol{\theta}$ persists.

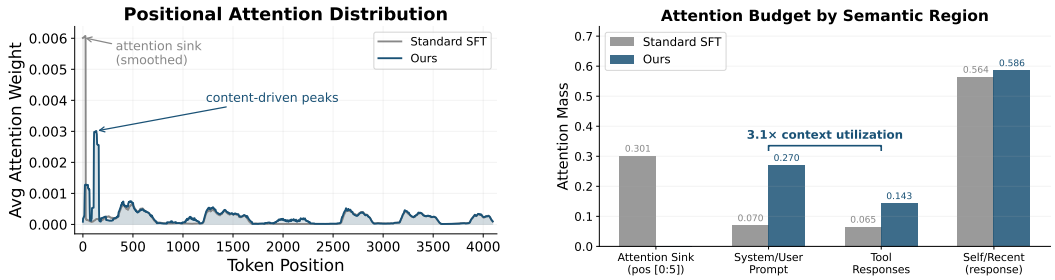
starvation of content tokens as *attention dilution* (formalized in Section 2.1). Existing remedies overwhelmingly target inference (positional calibration (Hsieh et al., 2024b), dynamic scaling (Ye et al., 2026), test-time training (Bansal et al., 2025; Tandon et al., 2025)) or require pretraining from scratch (Child et al., 2019; Ye et al., 2024).

What remains largely unexplored is whether the fine-tuning procedure itself contributes to this gap. We present evidence that it does: during standard SFT on a long sequence, the same biases and sink patterns govern the forward pass that produces the training loss. The gradient signal is computed from representations where most attention goes to positionally privileged tokens rather than content. Longer training sequences may reinforce rather than correct these patterns, creating a vicious cycle in which training-time dilution leads to poor long-context learning.

We propose FOCUSFT, a dilution-aware fine-tuning framework that breaks this cycle through bilevel optimization (Figure 1). In the inner loop, lightweight fast-weight adapters (Hinton & Plaut, 1987; Ba et al., 2016) perform a small number of gradient steps on the training context, forming a parametric memory that concentrates attention on relevant content. The outer loop then performs standard SFT conditioned on the sharpened representations produced by this memory, so that the gradient signal reflects actual context content rather than a diluted approximation. To mitigate the sink mechanism, both loops apply bidirectional attention over context tokens while preserving causal masking for responses (Du et al., 2022): when all context tokens can attend to each other, the asymmetric visibility that drives sinks is reduced. A key design principle, inner-outer consistency, aligns both loops to share the same attention structure and objective, so that the sharpened representations remain compatible with how the model processes context during training and inference.

Our main contributions are as follows:

- We identify training-time attention dilution (the starvation of content tokens due to positional biases and learned sinks) as a previously under-explored bottleneck for long-context learning, and characterize the vicious cycle between diluted training signals and poor long-context utilization.
- We propose FOCUSFT, a bilevel optimization framework in which an inner loop constructs parametric memory via fast-weight adaptation and an outer loop performs SFT conditioned on the sharpened representation. Both loops employ bidirectional context attention that reduces the causal asymmetry linked to attention sinks, unified by an inner-outer consistency principle.
- We demonstrate consistent improvements across benchmarks: up to +14pp on BABILong at 4K–32K, +8.2pp on RULER CWE aggregation at 16K, and +3.8pp pass@1 on GPQA agentic reasoning. Attention analysis shows a 529× reduction in sink mass and 3.1× higher context engagement.



(a) Positional attention distribution

(b) Attention budget by semantic region

Figure 2: Training-time attention on a 4096-token multi-turn agentic sample, averaged over response queries and all 28 layers. (a) FOCUSFT replaces the dominant sink with content-driven attention. (b) Context attention increases from 13.5% to 41.4% (3.1×).

2 PRELIMINARIES AND MOTIVATION

2.1 ATTENTION MECHANISMS AND LONG-CONTEXT FAILURE MODES

We briefly review scaled dot-product attention and the failure modes that arise as context length grows.

Given a sequence of T tokens with hidden representations $\{\mathbf{h}_i\}_{i=1}^{T} \in \mathbb{R}^d$, each transformer layer computes query, key, and value projections $\mathbf{q}_i = W_Q\mathbf{h}_i$, $\mathbf{k}_j = W_K\mathbf{h}_j$, $\mathbf{v}_j = W_V\mathbf{h}_j$, where $W_Q, W_K \in \mathbb{R}^{d_k \times d}$ and $W_V \in \mathbb{R}^{d_v \times d}$. The attention logits $z_{i,j} = \mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_k}$ are normalized via softmax to yield attention weights:

$$\alpha_{i,j} = \frac{\exp(z_{i,j})}{\sum_{\ell=1}^{T} \exp(z_{i,\ell})}, \qquad \mathbf{o}_i = \sum_{j=1}^{T} \alpha_{i,j}\, \mathbf{v}_j. \tag{1}$$

In autoregressive models, causal masking restricts the sum to $j \le i$.

Positional attention bias. LLMs exhibit a well-documented U-shaped positional bias (Liu et al., 2024): tokens at the beginning and end of the context receive systematically higher attention, irrespective of their relevance. This effect is shaped by positional encoding schemes such as RoPE distance decay and reinforced by the training data distribution. As a result, information placed in the middle of the context may be effectively invisible to the model, even when the context window comfortably accommodates it.

Attention sinks. Under causal masking, the first few tokens are the only positions visible to all subsequent tokens in the sequence. Models learn to exploit this unique global visibility by using these initial tokens as attention sinks (Xiao et al., 2023): destinations that absorb excess attention mass when no other token is strongly relevant. This is not a defect but a functional mechanism: the model needs somewhere to place the probability mass that softmax forces it to allocate, and the globally visible initial tokens are a natural choice. However, the consequence is that a substantial fraction of the attention budget is consumed by a handful of tokens that carry no semantic relevance.

Attention dilution. Together, positional bias and learned sinks cause the attention budget available for semantically relevant content to be *diluted*: the model allocates most of its attention to positionally privileged tokens, leaving content tokens underattended. We refer to this overall phenomenon as *attention dilution*.

2.2 TRAINING-TIME ATTENTION DILUTION: THE OVERLOOKED BOTTLENECK

Prior work has treated attention dilution primarily as an inference-time phenomenon. However, inference-time attention is a product of the learned parameters, which are shaped by training-time attention. We argue that a critical bottleneck lies in the training process itself, and provide empirical evidence in Figures 2 and 3.

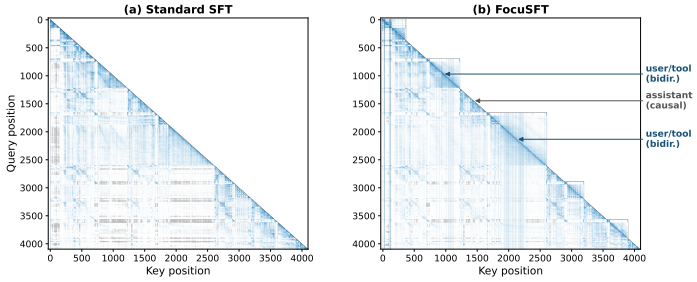
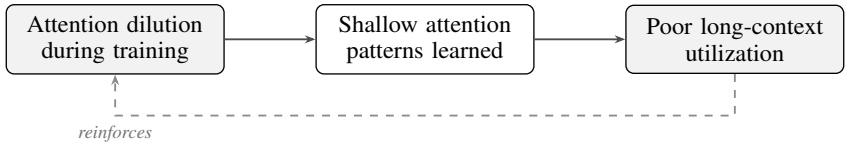


Figure 3: Middle-layer attention on a 4096-token multi-turn sample. (a) Standard SFT exhibits an initial-token sink. (b) FOCUSFT reveals bidirectional context blocks and causal response blocks.

Consider a standard SFT step on a long training sequence of length T. During the forward pass, the attention mechanism computes weights $\alpha_{i,j}$ over the entire sequence. Because of the patterns described above (positional bias and learned sinks), the output representation $\mathbf{o}_i$ at the prediction position is dominated by positionally privileged tokens rather than semantically relevant context. The cross-entropy loss computed from this representation therefore provides a gradient signal that reflects a diluted view of the training data: the model sees the relevant information in its context window but cannot attend to it.

Figure 2a makes this concrete: under standard SFT, nearly all attention mass concentrates at position 0 with negligible weight elsewhere. Figure 2b quantifies the consequence: the attention sink absorbs 30.1% of the budget on just 5 tokens, while the entire context content (system/user prompt and tool responses combined) receives only 13.5%. Figure 3 further reveals the structural pattern: a bright sink column at the initial tokens dominates across all query positions, obscuring the underlying dialogue structure. Over many training steps, the model converges to these attention patterns, failing to learn the sharp, content-specific focus required for reliable long-context utilization. (For comparison, these figures also show the corresponding patterns under FOCUSFT; we analyze these in Section 4.4.)

This creates a vicious cycle:



A natural response is to simply train on longer sequences. However, longer sequences exacerbate rather than alleviate dilution: the attention sink absorbs an even larger share of the budget, and the model has more distractor tokens competing for what remains. Empirically, models trained with longer context windows often show improved performance at short contexts but diminishing gains at the lengths they were trained on (Hsieh et al., 2024a; Kuratov et al., 2024).

These observations motivate a training-time approach that addresses two complementary aspects of the problem. First, the root cause: under causal masking, the asymmetric visibility structure creates attention sinks that waste attention budget. Bidirectional context attention addresses this asymmetry: when all context tokens can attend to each other, initial tokens are no longer uniquely privileged, reducing the pressure that drives the sink mechanism. Second, addressing the mask structure alone is insufficient: the model must still learn to concentrate attention on semantically relevant content, which requires active guidance during training. Bilevel optimization with an inner-loop parametric memory provides this guidance by sharpening the attention distribution, so that the outer-loop gradient signal better reflects actual context content. In the next section, we present how FOCUSFT realizes these ideas.

3 METHODOLOGY

3.1 BILEVEL OPTIMIZATION FRAMEWORK

Let $\boldsymbol{\theta}$ denote the base model parameters and $\mathbf{x} = [\mathbf{x}_{\text{ctx}}; \mathbf{x}_{\text{resp}}]$ a training sequence of T tokens, where $\mathcal{C}$ and $\mathcal{R}$ denote the sets of context and response token positions ($\mathcal{C} \cup \mathcal{R} = \{1, \ldots, T\}$). Let $\boldsymbol{\phi}$ be a

set of lightweight fast-weight parameters (LoRA adapters (Hu et al., 2021)) that are re-initialized at each training step. FOCUSFT decomposes each step into two nested optimization levels:

$$\min_{\boldsymbol{\theta}} \; \mathcal{L}_{\text{outer}}\big(\boldsymbol{\theta}, \boldsymbol{\phi}^{(K)}(\boldsymbol{\theta})\big), \qquad \text{where} \quad \boldsymbol{\phi}^{(K)} = \textsc{InnerLoop}\big(\boldsymbol{\phi}^{(0)}, \boldsymbol{\theta}, \mathbf{x}\big). \tag{2}$$

The inner loop runs K gradient steps to adapt $\boldsymbol{\phi}$, producing $\boldsymbol{\phi}^{(K)}$; the outer loop then performs standard SFT on the response tokens conditioned on these adapted fast weights. This structure is inspired by meta-learning (Finn et al., 2017; Nichol et al., 2018) and fast-weight mechanisms (Hinton & Plaut, 1987; Ba et al., 2016), repurposed to counteract attention dilution during training.

3.2 Inner Loop: Parametric Memory

The fast weights $\boldsymbol{\phi}$ are LoRA adapters applied to the feed-forward network (FFN) of a selected subset of transformer layers, re-initialized to zero at each training step. The inner loop performs K gradient steps to minimize an adaptation loss $\mathcal{L}_{\text{inner}}$ on the response tokens, using the same next-token prediction objective as the outer loop: $\mathcal{L}_{\text{inner}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = -\sum_{i \in \mathcal{R}} \log p_{\boldsymbol{\theta}, \boldsymbol{\phi}}(x_i \mid \mathbf{x}_{<i})$. The update rule is:

$$\boldsymbol{\phi}^{(k+1)} = \boldsymbol{\phi}^{(k)} - \eta_{\text{in}} \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{inner}}\big(\boldsymbol{\theta}, \boldsymbol{\phi}^{(k)}; \mathbf{x}\big), \qquad k = 0, \ldots, K-1, \tag{3}$$

where η_{in} is the inner learning rate. By optimizing the same response prediction objective, the inner loop forces $\boldsymbol{\phi}$ to encode context information that is directly useful for generating accurate responses. After K steps, the adapted $\boldsymbol{\phi}^{(K)}$ modifies the model's intermediate representations, indirectly reshaping the attention distribution in subsequent layers to better concentrate on the salient content of the current sample.

We adopt a first-order approximation (Nichol et al., 2018; Rajeswaran et al., 2019): the outer loss treats $\boldsymbol{\phi}^{(K)}$ as a constant, avoiding second-order derivatives through the inner-loop graph. This reduces memory and compute overhead while preserving the core benefit.

3.3 Outer Loop: SFT with Sharpened Attention

The outer loop performs a standard autoregressive forward pass using the combined parameters $(\boldsymbol{\theta}, \boldsymbol{\phi}^{(K)})$ and computes cross-entropy on the response token positions $\mathcal{R}$:

$$\mathcal{L}_{\text{outer}}(\boldsymbol{\theta}, \boldsymbol{\phi}^{(K)}) = -\sum_{i \in \mathcal{R}} \log p_{\boldsymbol{\theta}, \boldsymbol{\phi}^{(K)}}(x_i \mid \mathbf{x}_{<i}). \tag{4}$$

Only $\boldsymbol{\theta}$ receives gradient updates; $\boldsymbol{\phi}^{(K)}$ is discarded after each step. Because the fast weights shift attention toward relevant context, the gradient signal reaching $\boldsymbol{\theta}$ better reflects actual content rather than a diluted approximation.

3.4 Bidirectional Context Attention

Following GLM-style attention (Du et al., 2022), we apply bidirectional attention over context tokens while preserving causal masking for responses. The attention mask $M \in \{0, -\infty\}^{T \times T}$ is defined as:

$$M_{i,j} = \begin{cases} 0 & \text{if } i, j \in \mathcal{C}, \\ 0 & \text{if } i \in \mathcal{R} \text{ and } j \leq i, \\ -\infty & \text{otherwise.} \end{cases} \tag{5}$$

This mask is applied identically across all attention heads and in both the inner and outer loops.

Bidirectional context attention addresses the root cause of attention sinks identified in Section 2.1: under causal masking, initial tokens are the only globally visible positions and absorb excess attention mass. When all context tokens can attend to each other, this asymmetry vanishes and the sink mechanism becomes unnecessary. This is particularly beneficial for the inner loop, where a complete view of the context enables more effective parametric memory formation.

3.5 INNER-OUTER CONSISTENCY

The effectiveness of FOCUSFT depends on alignment between the two loops, a principle we call inner-outer consistency. If the inner loop operates under conditions that differ from the outer loop (e.g., mismatched attention masks or objectives), the fast weights may encode representations that are incompatible with the outer loop, producing distortion rather than sharpening.

We enforce consistency along two dimensions. **Objective**: the inner loop minimizes the same next-token prediction loss on response tokens as the outer loop (Equation (4)), so the fast weights are optimized to encode context representations that directly improve response generation. **Attention pattern**: both loops use the same attention mask (Equation (5)), with bidirectional attention over context tokens and causal masking for responses. Because the fast weights are LoRA adapters on the same FFN layers used by the outer loop, the sharpened representations are directly compatible by construction. The inner loop thus operates as a preview of the outer loop under identical conditions, forcing the fast weights to encode context representations that are directly compatible with the outer-loop objective. At inference time, no inner-loop computation is required: the fine-tuned $\boldsymbol{\theta}$ is used with standard autoregressive decoding, and the bilevel training produces attention patterns that are more content-focused even under standard causal masking.

4 EXPERIMENTS

We evaluate FOCUSFT on long-context understanding benchmarks spanning synthetic reasoning, retrieval-aggregation, real-world QA, and agentic reasoning.

4.1 EXPERIMENTAL SETUP

Base model and training data. We use Qwen2.5-7B (Yang et al., 2024) as the base model. Training uses 3K multi-turn agentic SFT samples (Yu et al., 2025) with a maximum sequence length of 4096 tokens, trained for 5 epochs with an effective batch size of 32 (8 GPUs, gradient accumulation 4). The outer-loop optimizer is AdamW (Loshchilov & Hutter, 2017) with learning rate 1×10^{-5}, cosine schedule with 10% warmup, and weight decay 0.01. All training uses BF16 mixed precision.

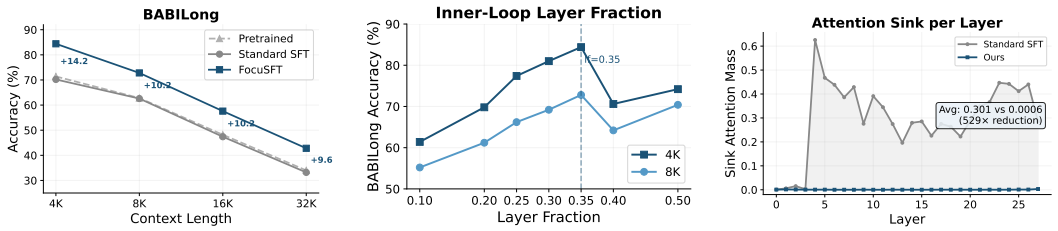
Bilevel hyperparameters. The inner loop performs $K{=}2$ gradient steps with learning rate 1.0 on LoRA (Hu et al., 2021) adapters (rank 32, $\alpha{=}64$) applied to the FFN layers of the top 35% of transformer layers. Inner gradients are clipped at norm 1.0. Full hyperparameter details are provided in Section A.

Baselines. We compare against: (1) the pretrained Qwen2.5-7B without fine-tuning, and (2) Standard SFT with identical data, model, and training budget but no bilevel optimization. For ablations, we additionally test SFT with bidirectional context attention (no bilevel) and causal bilevel (bilevel without bidirectional context).

Evaluation benchmarks. **BABILong** (Kuratov et al., 2024): reasoning-in-a-haystack tasks at context lengths 4K–32K, testing fact retrieval and multi-hop reasoning within long narratives. **RULER** (Hsieh et al., 2024a): multi-category benchmark covering retrieval (NIAH-MultiValue), aggregation (CWE), and multi-hop tracing (VT). **LongBench** (Bai et al., 2024b): real-world QA tasks (HotpotQA, MultifieldQA, NarrativeQA, Qasper) at 8K context. **GPQA** (Rein et al., 2023): graduate-level science reasoning (198 Diamond problems) evaluated with multi-turn agentic tool use via Open-AgentRL (Yu et al., 2025) ($n{=}32$ rollouts per problem).

4.2 MAIN RESULTS

BABILong. Figure 4a presents the main BABILong results. FOCUSFT outperforms Standard SFT by +14.2, +10.2, +10.2, and +9.6pp at 4K, 8K, 16K, and 32K respectively. Standard SFT provides essentially no improvement over the pretrained model at any length, suggesting that naive fine-tuning with diluted attention fails to teach long-context reasoning. FOCUSFT maintains its advantage well beyond the 4K training length, demonstrating that dilution-aware training produces representations that generalize to longer sequences. The gains are largest on multi-hop subtasks that require connecting dispersed facts across the context (Section B.1), the setting where attention dilution is most harmful.



(a) Accuracy vs. context length. (b) Inner-loop layer fraction. (c) Attention sink mass per layer.

Figure 4: BABILong performance and attention analysis. (a) FOCUSFT improves accuracy across context lengths. (b) Accuracy peaks at layer fraction 0.35. (c) Mean sink mass drops from 30.1% to 0.06% (529× reduction).

Table 1: RULER results by task category at 4K/8K/16K context lengths.

	NIAH-MV			CWE			VT		
Method	4K	8K	16K	4K	8K	16K	4K	8K	16K
Pretrained	98.4	95.1	94.0	98.0	83.2	66.7	99.6	96.5	93.7
Standard SFT	97.6	94.6	94.6	97.5	85.4	72.9	100.0	97.4	94.6
FOCUSFT	**98.9**	**95.4**	**95.6**	**97.8**	**88.8**	**81.1**	100.0	**97.5**	94.4

Table 2: Downstream evaluation: LongBench QA (F1, 8K context) and GPQA Diamond (agentic tool use, n=32).

	LongBench QA					GPQA	
Method	HotpotQA	Multifield	Narrative	Qasper	Avg	pass@1	pass@32
Standard SFT	10.7	28.6	8.7	12.2	15.0	15.6	78.8
FOCUSFT	**11.6**	**33.8**	**11.3**	**12.8**	**17.4**	**19.4**	**80.8**

Table 3: BABILong ablations and training cost (8× GPU, 469 steps). The 2×2 design isolates bilevel optimization and bidirectional context attention. All methods share the same inference cost.

			BABILong accuracy				Training cost		
Method	**Bilevel**	**Bidir**	**4K**	**8K**	**16K**	**32K**	**s/step**	**Time**	**Overhead**
Standard SFT	×	×	70.2	62.6	47.4	33.2	3.64	0.47h	1.00×
SFT + Bidir.	×	✓	65.4	57.6	46.0	33.0	3.64	0.47h	1.00×
Causal Bilevel	✓	×	82.4	72.6	56.6	39.0	5.52	0.72h	1.52×
FOCUSFT	✓	✓	**84.4**	**72.8**	**57.6**	**42.8**	6.21	0.81h	1.71×

RULER. Table 1 shows per-task RULER results. The improvement is most pronounced on CWE, where FOCUSFT achieves 81.1% vs. 72.9% (+8.2pp) at 16K. CWE requires aggregating information spread across the full context, consistent with the benefit of reduced attention dilution. On NIAH-MV and VT, both methods perform near ceiling at shorter lengths; FOCUSFT shows a modest edge on NIAH-MV at 16K (+1.0pp).

Downstream tasks. Table 2 reports results on LongBench QA and GPQA. On LongBench, FOCUSFT improves average F1 by +2.4pp, with the largest gain on MultifieldQA (+5.2pp), which requires cross-document evidence aggregation. On GPQA, FOCUSFT achieves 19.4% vs. 15.6% pass@1 (+3.8pp), suggesting that training-time attention improvements can transfer to complex agentic reasoning over long heterogeneous contexts.

4.3 Ablation Study

Table 3 presents the 2×2 factorial results. Bilevel optimization is the primary driver of improvement, accounting for +12.2, +10.0, +9.2, and +5.8pp over Standard SFT at 4K/8K/16K/32K, supporting the hypothesis that the inner-loop parametric memory concentrates attention on salient context during training. Bidirectional context attention alone (SFT + Bidir.) actually degrades performance by 4–5pp at shorter lengths: without the inner loop to leverage the richer representations, the train–eval mismatch between bidirectional training and causal inference introduces a distribution shift. However, when combined with bilevel optimization, bidirectional attention provides an additional +2.0, +0.2, +1.0, and +3.8pp over Causal Bilevel. The gap widens at 32K, where attention dilution is most severe and bidirectional encoding enables fuller aggregation of dispersed evidence; the combined gain of +9.6pp exceeds the sum of individual effects (+5.8 and -0.2), indicating a positive interaction between the two components.

Layer fraction sensitivity. Figure 4b shows BABILong accuracy as a function of the fraction of layers receiving LoRA adaptation in the inner loop. Performance exhibits a clear inverted-U shape, peaking at lf$=0.35$. Too few adapted layers (lf$\leq$0.20) under-capacitate the parametric memory, limiting its ability to encode context-specific representations. Too many (lf$\geq$0.40) cause the inner loop to over-specialize, disrupting the base model's pretrained representations and degrading the outer-loop gradient signal.

Other bilevel hyperparameters. The number of inner steps K and the inner learning rate exhibit similar sensitivity. $K=2$ is optimal; increasing to $K=3$ causes the inner loop to overfit to the current training sample, producing an outer gradient that no longer reflects general context utilization and leading to significant performance degradation. The inner learning rate must be sufficiently high (we use 1.0) to enable meaningful adaptation within only two gradient steps; substantially lower values underadapt, while excessively high values destabilize training. These interactions are consistent with the general principle that the inner loop should form a rough context sketch rather than memorize the sample.

Training efficiency. Table 3 reports training cost. The bilevel inner loop is the dominant overhead (1.52×); bidirectional attention adds negligible cost since it only changes the mask without additional parameters. FocuSFT totals 1.71× wall time (a modest premium for +14.2pp on BABILong) and incurs zero inference overhead, as the inner-loop adapters are discarded after training.

4.4 Attention Analysis

Section 2.2 identified training-time attention dilution in standard SFT. Here we examine how FocuSFT reshapes these patterns.

Per-layer sink reduction. Figure 4c shows that the attention sink under standard SFT is not confined to a few layers but is pervasive across all 28 layers. FocuSFT reduces the per-layer sink mass to 0.06% (a 529× reduction), confirming that bidirectional context attention removes the causal asymmetry that sustains the sink mechanism throughout the network.

Attention budget redistribution. With the sink eliminated, the freed attention budget is redirected to semantically relevant content (Figure 2b). System and user prompt attention rises from 7.0% to 27.0%, and tool response attention from 6.5% to 14.3%, yielding 3.1× higher total context engagement (13.5% → 41.4%). The positional profile (Figure 2a) reflects this shift: the spike at position 0 disappears and is replaced by content-driven peaks aligned with semantically meaningful turns.

Structural transformation. The attention heatmaps (Figure 3) reveal the structural consequence. Under standard SFT, a uniform sink column dominates across all query positions, leaving the middle context substantially underattended. Under FocuSFT, attention is visibly spread across the middle context positions that were previously ignored: response queries now attend to content tokens throughout the sequence, not just the initial positions. The heatmap also reveals the multi-turn dialogue structure: context turns exhibit bidirectional attention blocks, while assistant responses maintain causal masking. This shift from positionally concentrated to content-distributed attention is the mechanism through which the budget redistribution above translates into improved long-context utilization.

5 Related Work

Long-context architectures. Extending the context window of transformer-based LLMs has been an active area of research. Positional encoding improvements such as RoPE scaling (Liu et al., 2023b; Ding et al., 2024; Peng et al., 2023) and ALiBi (Press et al., 2021) extrapolate to longer sequences, while distributed mechanisms like ring attention (Liu et al., 2023a) address memory constraints. Sparse attention patterns (Child et al., 2019; Beltagy et al., 2020; Yuan et al., 2025) reduce the quadratic cost while preserving long-range connectivity, and Differential Transformer (Ye et al., 2024) subtracts two softmax maps to cancel attention noise. These methods expand context *visibility* or efficiency but do not address the attention dilution that limits context *utilization* (Hsieh et al., 2024a; Liu et al., 2024). FocuSFT is orthogonal to these architectural advances: it keeps the standard attention mechanism intact and modifies only the training procedure.

Inference-time long-context methods. A growing family of methods address long-context failures at inference time. "Found in the Middle" (Hsieh et al., 2024b) applies post-hoc positional calibration; DySCO (Ye et al., 2026) dynamically up-weights retrieval-head scores during decoding. Test-time training (TTT) approaches (Sun et al., 2020; Bansal et al., 2025; Tandon et al., 2025) are methodologically closest to FocuSFT: they also adapt model parameters on the input context via gradient steps before generation. The key difference is that TTT operates at inference time on each new input, incurring per-sample compute overhead, while FocuSFT uses the same inner-loop mechanism during training only, producing a base model that requires no additional inference-time computation.

Long-context fine-tuning. LongLoRA (Chen et al., 2024) enables efficient long-context fine-tuning via shifted sparse attention during training with full attention at inference, and LongAlpaca (Chen et al., 2023) provides instruction-following data for long-context adaptation. LongAlign (Bai et al., 2024a) further studies long-context alignment data construction and training efficiency. These approaches increase the model's ability to process longer inputs but do not specifically address how attention budget is distributed during fine-tuning. FocuSFT targets a complementary axis: rather than extending the context window, it improves how the model uses the context already within its window during SFT.

Bilevel optimization and fast weights. Bilevel optimization is well established in meta-learning (Thrun & Pratt, 1998; Finn et al., 2017; Nichol et al., 2018; Rajeswaran et al., 2019), where MAML (Finn et al., 2017) uses inner-loop adaptation to learn task-generalizing initializations. Recent work has applied bilevel optimization to LLM data reweighting (Pan et al., 2025). Fast weights (Hinton & Plaut, 1987; Ba et al., 2016) maintain a secondary set of parameters updated on a faster timescale; Fast Weight Layers (Clark et al., 2022) express this as linear attention, and MemoryLLM (Wang et al., 2024) integrates context into model parameters for knowledge retention. FocuSFT aims at a different goal: not cross-task generalization or knowledge storage, but *intra-sequence* attention sharpening during long-context SFT.

6 Limitations

FocuSFT requires 1.71× the training time due to inner-loop updates, with unchanged inference cost (Table 3). Evaluation covers only Qwen2.5-7B and a fixed 3K-sample corpus; scaling across model sizes and larger datasets remains unexplored. The method targets SFT; integration with reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) has not been investigated.

7 Conclusion

Training-time attention dilution contributes to poor long-context utilization. FocuSFT uses inner-loop fast weights to form content-focused parametric memory and outer-loop SFT to learn from sharpened representations. Bidirectional context attention reduces sink-inducing causal asymmetry, while inner-outer consistency aligns representations with downstream use. Together, these mechanisms mitigate the cycle linking diluted training to poor long-context learning.

AI USE STATEMENT

Generative AI tools were used only to edit the manuscript text and assist with LaTeX formatting, including migration to the ICLR 2027 template. They were not used to develop research ideas, implement research code, conduct experiments, or analyze results. The authors take responsibility for the final content of the manuscript.

REFERENCES

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29, 2016.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1376–1395, 2024a.

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

Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, Sai Surya Duvvuri, Devvrit Khatri, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Sanjay Kale, et al. Let's (not) just put things in context: Test-time training for long-context llms. *arXiv preprint arXiv:2512.13898*, 2025.

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

Yukang Chen, Shaozuo Yu, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Long alpaca: Long-context instruction-following models. https://github.com/dvlab-research/LongLoRA, 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 *The International Conference on Learning Representations (ICLR)*, 2024.

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

Kevin Clark, Kelvin Guu, Ming-Wei Chang, Panupong Pasupat, Geoffrey Hinton, and Mohammad Norouzi. Meta-learning fast weight language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9751–9757, 2022.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

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

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Geoffrey E Hinton and David C Plaut. Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the Cognitive Science Society*, pp. 177–186, 1987.

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? *arXiv preprint arXiv:2404.06654*, 2024a.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 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. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. Scalebio: Scalable bilevel optimization for llm data reweighting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31959–31982, 2025.

Zehua Pei, Hui-Ling Zhen, Shixiong Kai, Sinno Jialin Pan, Yunhe Wang, Mingxuan Yuan, and Bei Yu. Scope: Prompt evolution for enhancing agent effectiveness. *arXiv preprint arXiv:2512.15374*, 2025.

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

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

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Arnuv Tandon, Karan Dalal, Xinhao Li, Daniel Koceja, Marcel Rød, Sam Buchanan, Xiaolong Wang, Jure Leskovec, Sanmi Koyejo, Tatsunori Hashimoto, et al. End-to-end test-time training for long context. *arXiv preprint arXiv:2512.23675*, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pp. 3–17. Springer, 1998.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian J. McAuley. MEMORYLLM: towards self-updatable large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=p0lKWzdikQ.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv e-prints*, pp. arXiv–2412, 2024.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

Xi Ye, Wuwei Zhang, Fangcong Yin, Howard Yen, and Danqi Chen. Dysco: Dynamic attention-scaling decoding for long-context lms. *arXiv preprint arXiv:2602.22175*, 2026.

Zhaochen Yu, Ling Yang, Jiaru Zou, Shuicheng Yan, and Mengdi Wang. Demystifying reinforcement learning in agentic reasoning. *arXiv preprint arXiv:2510.11701*, 2025.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097, 2025.

A Additional Experimental Details

Training hyperparameters. Table 4 lists all hyperparameters for the main FocuSFT configuration.

Table 4: Hyperparameters for FocuSFT training.

Parameter	Value
Outer loop (SFT)	
Base model	Qwen2.5-7B
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Learning rate	1×10^{-5}
LR schedule	Cosine with 10% linear warmup
Weight decay	0.01
Max gradient norm	1.0
Epochs	5
Effective batch size	32 (8 GPUs $\times$ 1 $\times$ GA 4)
Max sequence length	4096
Precision	BF16
Inner loop (parametric memory)	
Inner steps (K)	2
Inner learning rate	1.0
Inner gradient clip	1.0
Adapter type	LoRA
LoRA rank	32
LoRA α	64
LoRA dropout	0.0
Target modules	`gate_proj`, `up_proj`, `down_proj`
Layer fraction	0.35 (top 10 of 28 layers)
Attention mode	Bidirectional context (GLM-style)
Other	
Seed	1234
Gradient checkpointing	Enabled
Training samples	3,000
Hardware	8$\times$ NVIDIA GPU

GPQA evaluation. We evaluate GPQA Diamond using the Open-AgentRL (Yu et al., 2025) framework with multi-turn tool-use rollouts in Hermes chat format. Each of the 198 problems receives n=32 independent rollouts with temperature 1.0 and top-p 0.6, allowing up to 16 assistant turns per episode. We report pass@1 (majority vote) and pass@32 (oracle across all rollouts).

B Additional Results

B.1 Per-Task BABILong Breakdown

Table 5 shows per-task BABILong accuracy at 4K and 16K context lengths. The largest gains appear on QA2 (two-fact reasoning, +26pp at 4K) and QA3 (temporal reasoning, +31pp at 4K, +30pp at 16K), both of which require the model to locate and connect multiple dispersed facts across the context, precisely the scenario where attention dilution is most damaging.

B.2 RULER 2 $\times$ 2 Ablation

Table 6 extends the BABILong ablation (Table 3) to RULER, confirming the same pattern: bilevel optimization provides the primary benefit on aggregation tasks (CWE), while bidirectional attention alone slightly degrades performance at 16K.

Table 5: Per-task BABILong accuracy. QA2 and QA3 require multi-hop reasoning over dispersed facts.

		4K			16K		
Task	**Type**	SFT	FocuSFT	Δ	SFT	FocuSFT	Δ
QA1	Single fact	78.0	84.0	+6.0	57.0	52.0	−5.0
QA2	Two facts	51.0	77.0	**+26.0**	19.0	31.0	+12.0
QA3	Temporal	53.0	84.0	**+31.0**	37.0	67.0	**+30.0**
QA4	Spatial	82.0	89.0	+7.0	63.0	69.0	+6.0
QA5	Argument	87.0	88.0	+1.0	61.0	69.0	+8.0
Average		70.2	**84.4**	+14.2	47.4	**57.6**	+10.2

Table 6: RULER 2×2 ablation (average accuracy across NIAH-MV, CWE, VT).

Method	**Bilevel**	**Bidir**	**Avg@4K**	**Avg@8K**	**Avg@16K**
Standard SFT	×	×	98.3	92.5	87.3
SFT + Bidir.	×	✓	98.5	92.1	84.9
Causal Bilevel	✓	×	97.9	91.9	90.3
FocuSFT	✓	✓	**98.9**	**93.9**	**90.4**

B.3 Inference-Time Adaptation

A natural question is whether the inner-loop adapters can also be applied at inference time for additional gains. We tested this by performing a single inner-loop gradient step on the test input before generation. Table 7 shows that inference-time adaptation does not consistently improve over the base FocuSFT model and slightly hurts on most benchmarks.

Table 7: Effect of inference-time adaptation. The bilevel-trained model already internalizes the attention-sharpening benefit; additional test-time adaptation is unnecessary.

	BABILong			
Method	4K	8K	16K	32K
FocuSFT	**84.4**	**72.8**	**57.6**	**42.8**
FocuSFT + Inf. Adapt.	83.4	71.6	53.4	41.6

FocuSFT therefore incurs zero inference overhead: the bilevel training procedure improves the base model weights directly, and the inner-loop adapters are discarded after training.

C Attention Visualization Details

The attention heatmaps (Figure 3) and per-layer analysis (Figure 4c) are computed from a representative 4096-token multi-turn agentic sample containing 5 context turns (system prompt + 4 tool responses) and 5 assistant response turns. Attention weights are extracted from layer 14 (middle of the network) for heatmaps, and averaged across all attention heads. The sink mass metric sums attention weight on positions [0:5] from all response-query positions.