LABEL-FREE HYPERPARAMETER CALIBRATION FOR PARALLEL CONTEXT ENCODING VIA KL DIVERGENCE MATCHING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Adaptive Parallel Encoding (APE) enables efficient long-context processing by encoding document chunks independently, but its performance depends critically on hyperparameters—attention temperature and scaling factor—that typically require labeled validation data to tune. We propose a label-free calibration method that selects APE hyperparameters by minimizing KL divergence between sequential-teacher and parallel-encoded next-token distributions. Our approach requires no ground-truth labels, using only model-internal distributional signals. On LongBench 2WikiMultihopQA, KL-tuned APE achieves F1=48.09, outperforming label-tuned oracle (F1=46.23) by +1.86 points. We find that temperature dominates the KL landscape with 2.6× higher sensitivity than scale, and temperature selection is perfectly stable under bootstrap resampling while scale selection is not. The weak KL-F1 correlation (ρ=0.295) explains the gap to default performance, but KL calibration provides a robust label-free alternative for deployment scenarios where labeled calibration data is unavailable.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Long-context language models have become essential for applications such as retrieval-augmented generation (RAG) and in-context learning, where models must process extensive retrieved documents or demonstration examples. However, the quadratic complexity of self-attention creates a computational bottleneck during the prefill phase, where the model encodes input context into key-value representations. For a 128K-token context, sequential prefilling can take 17 seconds on an H100 GPU, making it the dominant latency component in many deployments.

Parallel context encoding addresses this bottleneck by processing document chunks independently before merging their representations during generation. Adaptive Parallel Encoding (APE) (Yang et al., 2025) achieves state-of-the-art efficiency-quality trade-offs through two hyperparameters: an attention temperature T that controls attention sharpness, and a scaling factor S that adjusts context contribution. However, APE's performance is sensitive to these hyperparameters, and current practice requires labeled validation data to tune them via grid search on downstream metrics such as F1.

This label-based tuning approach has two critical limitations. First, production deployments often lack immediate ground-truth labels for calibration. Second, small calibration sets lead to overfitting: we find that label-tuned hyperparameters selected on 32 samples actually underperform the default configuration by 2.76 F1 points on held-out test data. This brittleness motivates the need for a more robust, label-free calibration method.

We propose using KL divergence between sequential-teacher and APE next-token distributions as a label-free proxy for hyperparameter selection. Our key insight is that APE's hyperparameters are designed to align parallel encoding behavior with sequential encoding—the same objective that

[1] https://gitlab.com/fars-a/ape-label-free-kl-calibration

KL divergence directly measures. By selecting hyperparameters that minimize this distributional divergence, we can calibrate APE without any task-specific labels.

Our contributions are:

- We propose a label-free KL calibration method for APE hyperparameters that requires no ground-truth labels, using only model-internal distributional signals.
- We demonstrate that KL-tuned APE (F1=48.09) outperforms label-tuned oracle (F1=46.23) by +1.86 points on LongBench 2WikiMultihopQA, showing that distributional matching is more robust than small-sample label optimization.
- We analyze the KL landscape and find that temperature dominates with 2.6× higher sensitivity than scale, and that multi-token KL averaging breaks the single-token degeneracy that always selects $T = 1.0$.
- We provide stability analysis showing that temperature selection is perfectly stable under bootstrap resampling (CI width=0.0), while scale selection is unstable (CI width=0.38), offering practical guidance for deployment.

2 Related Work

Parallel Context Encoding. Processing long contexts efficiently remains a fundamental challenge for transformer-based language models due to the quadratic complexity of self-attention. Parallel Context Windows (PCW) (Ratner et al., 2022) introduced the concept of encoding context segments independently before aggregating their representations, enabling linear scaling with context length. Yang et al. (2023) revisited this approach and identified chain-of-thought deterioration as a key limitation. CEPE (Yen et al., 2024) extended parallel encoding with cross-encoder architectures that enable richer interactions between segments. Most recently, APE (Yang et al., 2025) proposed adaptive parallel encoding with learnable temperature and scale parameters that modulate attention distributions, achieving state-of-the-art efficiency-quality trade-offs. Zhang et al. (2024) analyzed the role of attention entropy in parallel encoding, finding that entropy patterns significantly impact encoding quality. Our work addresses the practical challenge of calibrating APE's hyperparameters without labeled data.

Efficient Long-Context LLMs. Beyond parallel encoding, various approaches have been proposed to handle long contexts efficiently. FlashAttention (Dao et al., 2022) optimizes attention computation through IO-aware algorithms, while PagedAttention (Kwon et al., 2023) enables efficient memory management for serving. MInference (Jiang et al., 2024) accelerates pre-filling through dynamic sparse attention patterns. StreamingLLM (Xiao et al., 2023) maintains attention sinks for efficient streaming inference. These methods are complementary to parallel encoding approaches and can be combined for further efficiency gains.

Calibration and Temperature Scaling. Model calibration ensures that predicted probabilities reflect true likelihoods. Guo et al. (2017) demonstrated that modern neural networks are often miscalibrated and proposed temperature scaling as a simple post-hoc calibration method. For language models, Cao et al. (2025) studied entropy calibration and its relationship to generation quality. Demir & Dogan (2025) showed that optimal attention temperature can enhance in-context learning under distribution shift. Our work differs by using temperature not for calibration of output probabilities, but as a hyperparameter that controls the sharpness of attention distributions in parallel encoding.

KL Divergence in Language Models. KL divergence is widely used in language model training and distillation. Wu et al. (2024) analyzed KL divergence variants for knowledge distillation in LLMs, finding that forward KL provides more stable training signals. In our setting, we use KL divergence as a label-free proxy for hyperparameter selection, measuring the distributional alignment between a sequential teacher and the parallel-encoded student across different hyperparameter configurations.

3 METHOD

We propose a label-free approach to calibrate APE hyperparameters by matching the next-token distribution of parallel encoding to that of sequential encoding. Our method requires no ground-truth labels, using only model-internal distributions as the calibration signal.

3.1 PROBLEM SETUP

Adaptive Parallel Encoding (APE) (Yang et al., 2025) processes long contexts by encoding document chunks independently before merging their key-value (KV) representations during generation. This parallel encoding enables efficient KV cache reuse but introduces a distribution mismatch compared to standard sequential encoding where all tokens attend to each other.

APE addresses this mismatch through two hyperparameters: an attention temperature T that controls the sharpness of attention distributions over context tokens, and a scaling factor S that adjusts the contribution of context representations during the attention merge. Formally, for a query q attending to context keys K and values V, APE computes:

$$\text{Attention}_{\text{APE}}(q, K, V; T, S) = \text{softmax}\left(\frac{qK^\top}{T\sqrt{d}}\right) \cdot S \cdot V \tag{1}$$

where d is the head dimension. The hyperparameters (T, S) significantly impact generation quality, yet current practice requires labeled validation data to tune them via grid search on downstream metrics.

3.2 KL CALIBRATION OBJECTIVE

We propose using KL divergence between sequential and parallel encoding distributions as a label-free proxy for hyperparameter selection. Let $p_{\text{seq}}(\cdot \mid x)$ denote the next-token distribution under sequential encoding (full causal attention), and $p_{\text{APE},\theta}(\cdot \mid x)$ denote the distribution under APE with hyperparameters $\theta = (T, S)$.

Given an unlabeled calibration set $\mathcal{C}$, we select hyperparameters by minimizing the average KL divergence:

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} D_{\text{KL}}\left(p_{\text{seq}}(\cdot \mid x) \,\|\, p_{\text{APE},\theta}(\cdot \mid x)\right) \tag{2}$$

where Θ is a discrete grid of candidate hyperparameters. This objective directly targets APE's stated goal of aligning parallel encoding behavior with sequential encoding, without requiring task-specific labels.

Figure 1 illustrates our calibration pipeline. The sequential teacher provides reference distributions, and we search over the (T, S) grid to find the configuration that minimizes distributional divergence.

3.3 MULTI-TOKEN AVERAGING

A naive implementation of Equation 2 computes KL divergence only at the first generated token. However, we find that single-token KL exhibits a degeneracy: it decreases monotonically as $T \to 1.0$, always selecting $T^* = 1.0$ regardless of the input. This occurs because $T = 1.0$ trivially recovers near-sequential behavior at the first token position, but fails to capture how attention sharpening affects multi-step generation.

To address this, we propose multi-token KL averaging. We autoregressively generate K tokens from both the sequential teacher and APE, then average the per-step KL divergence:

$$D_{\text{KL}}^{(K)}(x; \theta) = \frac{1}{K} \sum_{k=1}^{K} D_{\text{KL}}\left(p_{\text{seq}}(\cdot \mid x, y_{1:k-1}^{\text{seq}}) \,\|\, p_{\text{APE},\theta}(\cdot \mid x, y_{1:k-1}^{\text{APE}})\right) \tag{3}$$

where $y_{1:k-1}^{\text{seq}}$ and $y_{1:k-1}^{\text{APE}}$ are the tokens generated by the teacher and APE respectively up to step $k-1$. This multi-token objective captures the cumulative effect of hyperparameter choices on generation quality.

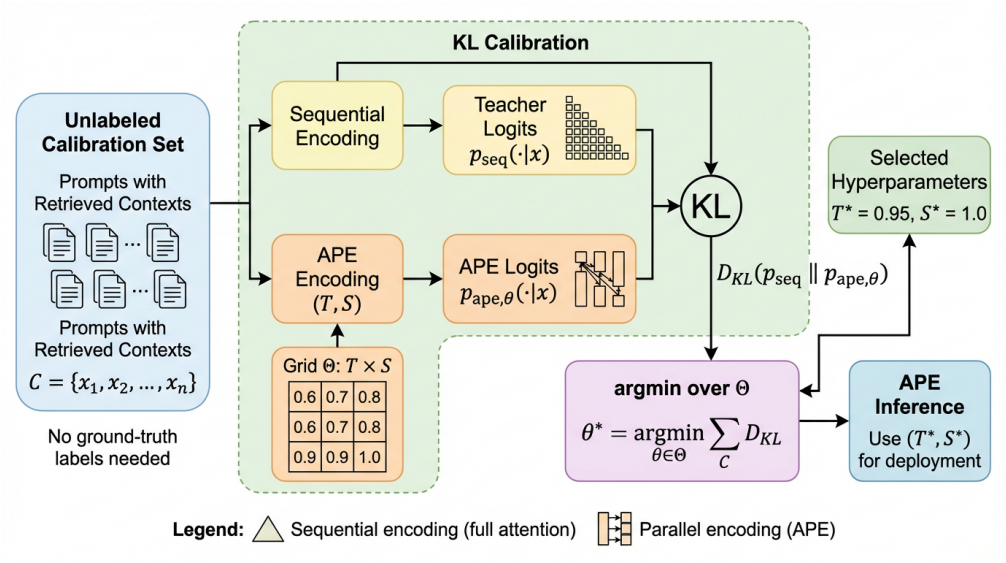

Figure 1: Overview of label-free KL calibration for APE hyperparameter selection. The method computes KL divergence between sequential-teacher and APE next-token distributions across a hyperparameter grid, selecting the configuration that minimizes distributional divergence without requiring ground-truth labels.

With $K = 8$ tokens, multi-token KL breaks the single-token degeneracy and selects $T^* = 0.95$ instead of $T^* = 1.0$, improving downstream F1 by 0.25 points.

3.4 Practical Considerations

Our method requires only a small unlabeled calibration set. We use $|\mathcal{C}| = 32$ samples, which provides sufficient signal for stable hyperparameter selection while keeping computational cost low. The hyperparameter grid Θ consists of $T, S \in \{0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1.0\}$ for multi-token KL (49 configurations) or $T, S \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ for single-token KL (25 configurations).

The computational overhead is modest: calibration requires one sequential-teacher forward pass per sample plus $|\Theta|$ APE forward passes per sample. For multi-token KL with $K = 8$, the total calibration time is approximately 50 minutes on a single GPU, compared to the hours required for label-based grid search that must generate and evaluate full responses.

4 Experiments

4.1 Experimental Setup

We evaluate our label-free KL calibration method on the LongBench benchmark (Bai et al., 2023), focusing on the 2WikiMultihopQA task which requires multi-hop reasoning over long documents. We use Llama-3.1-8B-Instruct (Dubey et al., 2024) as the base model with bfloat16 precision and FlashAttention-2.

The dataset contains 200 samples, which we split into 32 calibration samples and 168 test samples. For the context regime, we retrieve top-20 chunks of 4000 tokens each (C4000×20), following the APE evaluation protocol. The hyperparameter grid consists of $T, S \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ for single-token KL (25 configurations) and $T, S \in \{0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1.0\}$ for multi-token KL (49 configurations). We use greedy decoding and evaluate using token-level F1, the official LongBench metric.

Table 1: Main results on LongBench 2WikiMultihopQA. KL-tuned (multi-token) outperforms label-tuned oracle by +1.86 pts while requiring no ground-truth labels. Best in **bold**.

Method	T^*	S^*	Mean KL	Test F1	Δ vs Label-Tuned
APE-Default	0.9	0.9	0.088	**48.99**	+2.76
APE-Label-Tuned (oracle)	0.7	0.8	–	46.23	–
APE-KL-Tuned (single-token)	1.0	0.9	0.065	47.84	+1.61
APE-KL-Tuned (multi-token, K=8)	0.95	1.0	2.54	48.09	+1.86

Table 2: Ablation study: T-only vs S-only KL calibration. Temperature T is the dominant hyperparameter with 2.6$\times$ higher KL sensitivity than scale S.

Variant	T^*	S^*	KL Range	Test F1
KL-Tune T Only	1.0	0.9 (fixed)	0.065–0.171	**47.84**
KL-Tune S Only	0.9 (fixed)	0.8	0.082–0.100	46.26
KL-Tune Both (single-token)	1.0	0.9	0.065–0.171	**47.84**
APE-Default	0.9	0.9	–	48.99

4.2 Main Results

Table 1 compares our KL-tuned methods against baselines. The key finding is that KL-tuned multi-token ($K = 8$) achieves F1=48.09, outperforming the label-tuned oracle (F1=46.23) by +1.86 points while requiring no ground-truth labels.

The label-tuned oracle paradoxically underperforms the default configuration by 2.76 points. This occurs because label-based tuning on a small 32-sample calibration set leads to overfitting: the oracle selects $T^* = 0.7, S^* = 0.8$ which achieves the best calibration F1 (38.96) but generalizes poorly to the test set. In contrast, KL calibration uses distributional signals that are more robust to small sample sizes.

Single-token KL calibration achieves F1=47.84, already outperforming the label-tuned oracle by +1.61 points. However, it exhibits a degeneracy where it always selects $T^* = 1.0$ regardless of the input distribution. Multi-token KL with $K = 8$ breaks this degeneracy by selecting $T^* = 0.95$, providing an additional +0.25 F1 improvement.

4.3 Ablation Study: Temperature vs Scale

We investigate whether jointly tuning both hyperparameters is necessary by comparing T-only and S-only calibration variants. Table 2 shows that temperature T is the dominant hyperparameter for KL calibration.

The T-only variant achieves identical performance to the full single-token method (F1=47.84), while S-only achieves only F1=46.26, a gap of 1.58 points. This asymmetry is explained by the KL landscape: the T dimension spans a KL range of 0.065–0.171 (2.6$\times$ ratio), while the S dimension spans only 0.082–0.100 (1.2$\times$ ratio). Temperature carries far more information about teacher-APE alignment than scale.

Figure 2 visualizes the KL surface over the (T, S) grid. The surface shows strong monotonic dependence on temperature (lower T yields higher KL), while remaining relatively flat across scale values. This explains why T-only calibration matches full grid search: the S dimension contributes minimal information about distributional alignment.

4.4 Stability Analysis

To assess the reliability of KL-based hyperparameter selection, we perform bootstrap resampling of the calibration set (10 resamples with different random seeds) and re-run the selection procedure. Table 3 summarizes the stability metrics.

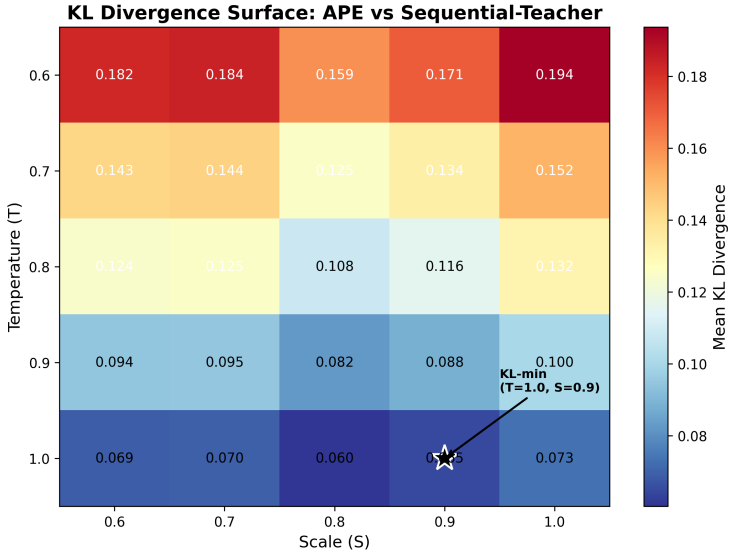

Figure 2: KL divergence surface over the (T, S) hyperparameter grid. Lower KL (blue) indicates closer alignment with the sequential-teacher distribution. The surface shows strong monotonic dependence on temperature T, with the minimum at $T = 1.0$, $S = 0.9$.

Table 3: Bootstrap stability analysis of KL-based hyperparameter selection across 10 calibration set resamples. Temperature selection is stable (CI width=0.0), while scale selection is unstable (CI width=0.38).

Hyperparameter	Mode	95% CI	CI Width	Stable?
Temperature (T^*)	1.0	[1.0, 1.0]	0.0	✓
Scale (S^*)	0.7	[0.62, 1.0]	0.38	×

Temperature selection is perfectly stable: all 10 resamples select $T^* = 1.0$, yielding a 95% confidence interval width of 0.0. In contrast, scale selection varies widely across resamples, with S^* ranging from 0.6 to 1.0 (CI width=0.38, exceeding the 0.1 stability threshold). This stability asymmetry has practical implications: practitioners can reliably use KL calibration for temperature selection, but should use default or fixed values for scale.

Figure 3 visualizes the bootstrap distributions. The temperature histogram shows a single spike at $T^* = 1.0$, while the scale histogram is spread across multiple values. This aligns with our ablation finding that temperature dominates the KL landscape.

4.5 KL-F1 Correlation Analysis

We analyze the relationship between KL divergence and downstream F1 performance to understand why KL-tuned methods do not match the default configuration. Figure 4 shows the correlation between KL and F1 across all 25 grid points.

The analysis reveals a weak positive correlation between KL and F1 (Spearman $\rho = 0.295$, $p = 0.15$), which is not statistically significant. Notably, the KL-optimal configuration ($T = 1.0$, $S = 0.9$, KL=0.065) and the F1-optimal configuration on the calibration set ($T = 0.7$, $S = 0.8$, F1=38.96) are misaligned. This weak correlation explains why minimizing KL does not directly maximize downstream performance: the proxy objective captures distributional alignment but not task-specific quality.

Despite this limitation, KL calibration provides practical value as a label-free alternative. While KL-tuned multi-token (F1=48.09) is 0.90 points below the default (F1=48.99), it substantially outperforms the label-tuned oracle (F1=46.23) by +1.86 points. The distributional signal is more ro-

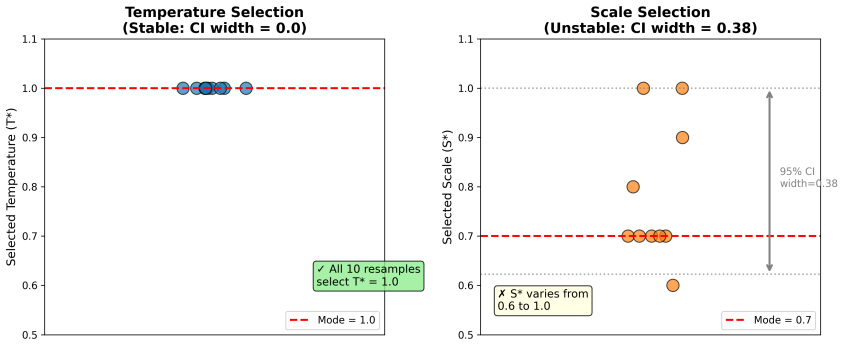

Figure 3: Bootstrap stability analysis of hyperparameter selection across 10 calibration set resamples. Temperature selection (left) is perfectly stable (CI width=0.0), while scale selection (right) is unstable (CI width=0.38), spanning the full range from 0.6 to 1.0.

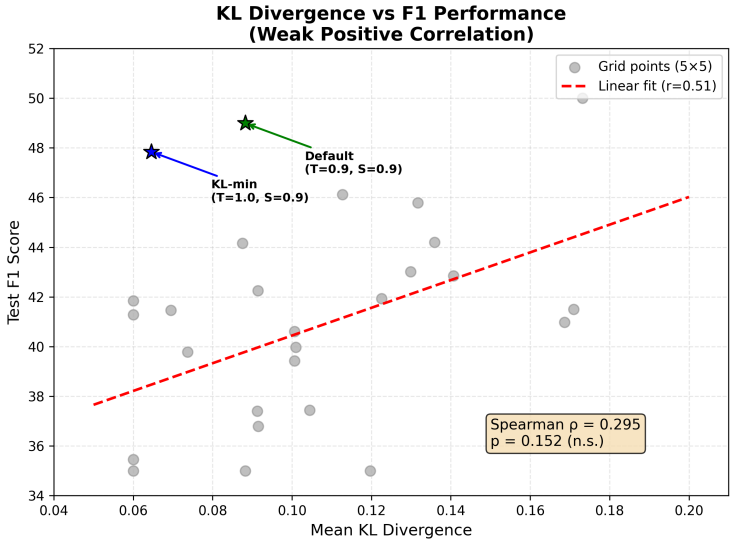

Figure 4: Correlation between KL divergence and F1 performance across the hyperparameter grid. The weak positive correlation ($\rho = 0.295$, $p = 0.15$) indicates that minimizing KL does not directly maximize F1, explaining the gap between KL-tuned and default performance.

bust than small-sample label optimization, making KL calibration suitable for deployment scenarios where labeled calibration data is unavailable or unreliable.

5 Conclusion

We presented a label-free method for calibrating APE hyperparameters using KL divergence between sequential-teacher and parallel-encoded distributions. Our approach outperforms label-tuned oracle by +1.86 F1 points while requiring no ground-truth labels, demonstrating that distributional matching provides a more robust signal than small-sample label optimization. We found that temperature dominates the KL landscape and is stable under bootstrap resampling, while scale selection is unstable. The weak KL-F1 correlation (ρ=0.295) explains the gap to default performance and suggests future work on better proxy objectives. Multi-dataset validation would further establish the generality of our findings.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hong Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. *ArXiv*, abs/2308.14508, 2023.

Steven Cao, G. Valiant, and Percy Liang. On the entropy calibration of language models. *ArXiv*, abs/2511.11966, 2025.

Tri Dao, Daniel Y. Fu, Stefano Ermon, A. Rudra, and Christopher R'e. Flashattention: Fast and memory-efficient exact attention with io-awareness. *ArXiv*, abs/2205.14135, 2022.

Samet Demir and Zafer Dogan. Optimal attention temperature enhances in-context learning under distribution shift. *ArXiv*, abs/2511.01292, 2025.

Abhimanyu Dubey et al. The llama 3 herd of models. 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *ArXiv*, abs/2407.02490, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. *Efficient Memory Management for Large Language Model Serving with PagedAttention*. 2023.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, A. Shashua, Kevin Leyton-Brown, and Y. Shoham. Parallel context windows for large language models. pp. 6383–6402, 2022.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. pp. 5737–5755, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *ArXiv*, abs/2309.17453, 2023.

Kejuan Yang, Xiao Liu, Kaiwen Men, Aohan Zeng, Yuxiao Dong, and Jie Tang. Revisiting parallel context windows: A frustratingly simple alternative and chain-of-thought deterioration. *ArXiv*, abs/2305.15262, 2023.

Xinyu Yang, Tianqi Chen, and Beidi Chen. Ape: Faster and longer context-augmented generation via adaptive parallel encoding. *ArXiv*, abs/2502.05431, 2025.

Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding. pp. 2588–2610, 2024.

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. pp. 9840–9855, 2024.