MSRA: A Parameter-Efficient Spectral Correction for Local Attention — When Does It Help, and When Does It Fail?

Anonymous Authors

Abstract

Local attention with a sliding window is widely used to scale transformers to long contexts, yet it inevitably discards information beyond the window boundary. We present Multiscale Spectral Residual Attention (MSRA), a parameter-efficient correction that adds a fixed multiscale spectral residual branch to local attention, requiring only 4,600–18,000 trainable parameters per layer group. Through systematic diagnosis across GPT-2 and LLaMA models, we map the boundary between where this correction helps and where it does not. On PG-19 language modeling, MSRA yields consistent gains of 3–11% in perplexity across both architectures, with training at 1,024 tokens generalizing to evaluation at the same length. On LongBench, MSRA closes 6.9% of the full-local gap on average across 15 English tasks, with 10 of 15 tasks showing positive improvement. Mechanism analysis reveals that the correction operates as a statistical smoother rather than a retrieval mechanism: gate selectivity is near-uniform (94.5% of maximum entropy), and far-context retrieval on needle-in-haystack benchmarks shows zero recovery. We further characterize the quality–cost tradeoff across patch-layer configurations and document explicit failure modes on RULER, SCROLLS, and DeepSeek. For the GroundLM workshop, this work offers a concrete case study in evaluating the faithfulness-efficiency tradeoff: a parameter-efficient correction can improve language modeling, but those improvements do not imply faithful long-range grounding — a distinction with implications for how we evaluate and deploy efficient attention mechanisms.

1 Introduction

Scaling transformer models to long sequences remains a central challenge in language modeling. The quadratic complexity of standard self-attention makes full-context computation prohibitively expensive beyond moderate sequence lengths, driving widespread adoption of sparse and local attention mechanisms (Beltagy et al., 2020; Zaheer et al., 2020; Child et al., 2019). Among these, sliding window local attention restricts each token to attending only within a fixed window of preceding tokens, achieving linear complexity in the window size while preserving the expressive power of exact dot-product attention within that radius.

The tradeoff, however, is fundamental: information beyond the window boundary is entirely invisible to any given token. As sequence length grows relative to window size, the gap between local and full attention widens dramatically, degrading performance on tasks that require integrating evidence across long distances. Prior work has explored various strategies to mitigate this loss, including linearized attention approximations (Katharopoulos et al., 2020; Choromanski et al., 2021), frequency-domain convolution representations (Liang et al., 2025; Lee-Thorp et al., 2022), and state-space sequence models (Gu et al., 2022; Fu et al., 2023). Yet these approaches either replace the attention mechanism entirely or make strong parametric assumptions about the structure of long-range dependencies.

We take a different perspective: rather than replacing local attention, can we add a lightweight, parameter-efficient correction that partially recovers the information lost to window truncation? This question leads to Multiscale Spectral Residual Attention (MSRA), a corrective residual branch that operates in parallel with local exact attention. MSRA uses fixed multiscale exponential kernels in the frequency domain — implemented via FFT-based Toeplitz convolution — to inject a smoothed, long-range contextual correction into each attention head. The correction is added in head space before the output projection, preserving the local attention pathway intact.

Critically, our experiments reveal that the effective mechanism underlying MSRA differs sub-

stantially from our original design intent. While the method was initially conceived with query-conditioned gating over multiple spectral experts, systematic ablation shows that a single expert with uniform mixing — requiring only 4,600 trainable parameters — already captures most of the benefit. Gate selectivity sits at 94.5% of maximum entropy, indicating near-uniform expert weighting. Far-context retrieval benchmarks show zero recovery over local attention. And the method consistently fails to transfer to QA-style grounding tasks like RULER and SCROLLS.

These findings, though limiting the scope of the original claim, provide a clear and honest characterization of what spectral residual correction can and cannot do. Our contributions are thus threefold:

1. We present MSRA, a parameter-efficient spectral correction for local attention that uses 4,600–18,000 trainable parameters and consistently improves language modeling perplexity across GPT-2 and LLaMA architectures (3–11% relative gain).

2. Through systematic ablation, mechanism diagnostics, and multi-benchmark evaluation, we map the precise boundary between where MSRA helps (LM perplexity, near/mid-range context) and where it does not (far retrieval, QA transfer, instruction tasks).

3. We characterize the quality–cost tradeoff across patch-layer budgets, showing that even a 2-layer, 4.6K-parameter correction closes 6.4% of the local-full gap at 1,024 tokens.

This study is aligned with GroundLM's emphasis on evaluating the faithfulness-efficiency boundary: we characterize not only what MSRA achieves, but what it demonstrably fails to achieve, providing a concrete data point for practitioners deciding when such corrections are worth their cost.

2 Method

MSRA augments standard local attention with a far-context residual branch computed in the frequency domain. The overall output for token i combines a near-field exact attention term and a far-field spectral correction:

$$y_i = y_i^{\text{near}} + \gamma \cdot y_i^{\text{far}} \tag{1}$$

The near branch uses exact causal sliding window attention with window size w:

$$y_i^{\text{near}} = \sum_{j \in N(i)} \text{softmax}_j \left(\frac{q_i^\top k_j}{\sqrt{d}} \right) \cdot v_j \tag{2}$$

where $N(i) = \{j : i - w \leq j \leq i\}$.

For the far branch, we define M spectral experts, each associated with a long-range kernel $g_m(\Delta)$ that captures dependencies beyond the local window. Following the Toeplitz attention formulation (Qin et al., 2023), we express the far-branch computation through FFT-based convolution:

$$C_i^{(m)} = \sum_{j \in F(i)} g_m(i-j) \cdot \psi_m(k_j) \cdot v_j^\top \tag{3}$$

$$z_i^{(m)} = \sum_{j \in F(i)} g_m(i-j) \cdot \psi_m(k_j) \tag{4}$$

where $F(i) = \{j : j < i - w\}$ denotes far-context positions, $g_m(\Delta) = \exp(-\Delta/\sigma_m)$ is an exponential kernel with learnable scale σ_m, and $\psi_m(\cdot)$ is an expert-specific key projection using a positive feature map $\psi(x) = \text{ELU}(x) + 1$.

The normalized expert output combines query and key representations:

$$u_i^{(m)} = \frac{\phi(q_i)^\top C_i^{(m)}}{\phi(q_i)^\top z_i^{(m)} + \epsilon} \tag{5}$$

with $\phi(x) = \text{ELU}(x) + 1$ and $\epsilon = 10^{-6}$.

Experts are mixed via per-token gating:

$$\lambda_i = \text{softmax}(W_g \cdot q_i), \qquad y_i^{\text{far}} = \sum_{m=1}^{M} \lambda_{i,m} \cdot u_i^{(m)} \tag{6}$$

The far-branch output is scaled by a learnable LayerScale parameter γ, initialized to 10^{-5}, which allows the model to gradually incorporate the spectral correction during training. This zero-initialization ensures that at initialization, the model behaves identically to local-only attention.

2.1 Implementation

The convolutions in Equations 3 and 4 are computed via FFT in $O(n \log n)$ time using the convolution theorem. To maintain numerical stability, all linear projections and FFT operations use FP32 precision, while the main model forward pass uses FP16. We implement buffer preallocation and kernel FFT caching to reduce repeated memory allocation overhead during inference.

Model	Seeds	Window	Train Len	Local PPL	Window w	MSRA PPL	Local PPL	Gain	MSRA PPL	Gap Closed
GPT-2 (124M)	5	128	512	64	–	52.3	+3.8%		503.0	+4.0%
LLaMA-1B	4	128	512	128 (avg 4 seeds)	–	–	+11.0%		–	+11.5%
LLaMA-1B	1	128	1024	501.4	498.5	74.1	+11.2%		69.9	+6.9%
LLaMA-1B	1	256	1024	242.0	211.5	+12.6%				

Table 1: **PG-19 perplexity results.** MSRA provides consistent improvements across model scales, window sizes, and training lengths. LLaMA-1B results use the simplified M=1, gate-off configuration. GPT-2 results are averaged over 5 seeds; LLaMA over 4 seeds for 512-token training.

MSRA is applied only to the last L layers of the transformer, leaving earlier layers unchanged. This design reflects the observation that late-layer attention patterns show stronger Toeplitz consistency (mean 0.98 across heads), making them more amenable to spectral approximation.

2.2 Simplified Configuration

Through extensive ablation (Section 5), we identified a minimal effective configuration that we use as the default throughout this paper: $M = 1$ expert, no query-conditioned gating (uniform mixing), $L = 8$ patched layers, window size $w = 128$, and feature rank $r = 16$. This configuration uses only 18,448 trainable parameters for LLaMA-1B while matching or exceeding the performance of more complex variants. Unless otherwise noted, all results use this configuration.

3 Language Modeling Results

We evaluate MSRA on PG-19 (Rae et al., 2020), a collection of pre-1919 books that serves as a standard benchmark for long-range language modeling. All experiments train only the spectral branch parameters, keeping the base model frozen.

3.1 Main Results

Table 1 summarizes the main language modeling results. On GPT-2 (124M parameters), MSRA trained at 512 tokens with a 128-token window improves perplexity by 3.8% averaged over five independent seeds. The effect is substantially larger on LLaMA-1B, where the simplified configuration yields an average 11.0% improvement at the same sequence length across four seeds.

Training at longer sequences further validates the approach. With training at 1,024 tokens and evaluation at the same length, MSRA closes 11.2% of the local-full gap with a 128-token window, and 12.6% with a 256-token window. These results confirm that the spectral correction generalizes meaningfully beyond the short-sequence regime where it was first demonstrated.

Table 2: **Window size sweep on LLaMA-1B.** MSRA provides positive gain across all window sizes tested. The absolute gap shrinks with larger windows, but the relative improvement remains consistent at 4–12%.

Layers L	Trainable Params	Gap Closed	Params per % Gai
2	4,612	+6.4%	721
4	9,224	+7.7%	1,198
8	18,448	+11.5%	1,604

Table 3: **Quality–cost tradeoff across patch-layer budgets.** Even $L = 2$ layers (4.6K parameters) provide meaningful correction.

3.2 Window Size Sensitivity

Table 2 examines MSRA's behavior across different local window sizes on LLaMA-1B. The correction yields positive gains across the full range from $w = 64$ to $w = 256$, with the largest relative improvement occurring at intermediate window sizes. This pattern is consistent with MSRA acting as a window-agnostic residual correction rather than a mechanism tuned to a specific window scale.

3.3 Patch-Layer Budget Analysis

A key practical consideration is the tradeoff between correction quality and parameter budget. Table 3 reports the gap closed at 1,024 tokens for $L \in \{2, 4, 8\}$. Even the most lightweight configuration, with only 4,612 trainable parameters applied across two layers, closes 6.4% of the local-full gap — representing a parameter efficiency of 721 parameters per percentage point of improvement. Scaling to eight layers triples the gain to 11.5% while remaining well within the parameter budget of even the smallest PEFT adapters.

4 Grounding Boundary Analysis

A central question for any local attention correction is whether it meaningfully improves ground-

Model	Window	Near NLL	Mid NLL	Far NLL
Full Attention	–	0.43	0.71	1.19
Local Only	128	2.97	8.46	8.39
+ MSRA	128	2.51	8.42	8.41
Full Attention	–	0.62	0.54	1.18
Local Only	256	2.10	8.31	9.17
+ MSRA	256	3.66	8.70	9.13

Table 4: **NIAH retrieval by evidence distance.** Lower NLL indicates better retrieval. MSRA helps at near distances (within or close to the window) but provides no improvement for far-range evidence. Results at 1,024 tokens, 128-token window.

ing — the ability to retrieve and integrate evidence from distant context. We probe this boundary through controlled retrieval experiments and multi-task benchmark evaluation.

4.1 NIAH: Where Does the Correction Help?

We use a copy-based needle-in-haystack (NIAH) paradigm where a distinctive key–value pair is embedded at a controlled distance from a query prompt. Table 4 reports negative log-likelihood (NLL) of the correct answer token, bucketed by evidence distance from the query. At near range (within the local window), MSRA provides a meaningful improvement over local-only attention — a 26.5% reduction in NLL at $w = 128$. At mid-range (just beyond the window), the benefit largely vanishes. At far range, MSRA and local attention produce indistinguishable NLL.

This pattern is consistent across window sizes. With $w = 256$, the near-range benefit remains but the far-range gap is equally large. The correction, in other words, acts as a local statistical smoother rather than a mechanism for explicit long-range retrieval.

4.2 LongBench: Partial Recovery Across Tasks

LongBench (Bai et al., 2025) provides a diverse set of long-document tasks requiring evidence integration across paragraphs and pages. Table 5 reports results for all 15 English tasks using MSRA trained at 1,024 tokens on PG-19. On average, MSRA closes 6.9% of the gap between local attention ($w = 128$) and full attention. Ten of fifteen tasks show positive improvement, with particularly strong results on triviaqa (+87.8%), repobench-p (+77.2%), and trec (+77.2%).

Task	Full PPL	Local (w=128)	+ MSRA	Gap Clo
triviaqa	17	9,235	1,145	+87.8
repobench-p	9	4,896	1,135	+77.2
trec	19	4,910	1,133	+77.2
narrativeqa	24	3,170	1,441	+54.9
gov_report	6	479	353	+26.7
musique	22	7,612	5,596	+26.6
multi_news	7	1,160	925	+20.4
qmsum	21	459	373	+19.7
samsum	5	681	612	+10.2
...				
qasper	10	1,256	2,595	-107.5
lcc	37	9,755	20,059	-106.0

Table 5: **LongBench results for 15 English tasks.** MSRA (trained on PG-19 at 1,024 tokens) evaluated against local and full attention baselines. Average gap closed across all 15 tasks: 6.9%. Tasks sorted by MSRA benefit.

Task	Full PPL	Local PPL (w=128)	+ MSRA
RULER (4K truncation)			
niah_single	198	3,454	5,191
niah_multikey	138	3,456	2,662
qa_squad	148	5,881	454
qa_hotpotqa	62	2,655	18,071
SCROLLS			
contract_nli	–	0% acc	0% acc

Table 6: **RULER and SCROLLS results.** MSRA (PG-19 trained) provides no consistent benefit on QA-style grounding benchmarks. Base LMs achieve 0% accuracy on SCROLLS tasks regardless of attention type.

However, the results also reveal clear failure modes. On qasper and lcc, MSRA substantially degrades performance relative to local attention alone. These tasks involve structured question-answering over specialized documents, a setting for which the PG-19-trained spectral correction provides no useful signal. This pattern highlights an important grounding limitation: the correction transfers effectively to tasks where LM-style perplexity correlates with task performance, but fails when the task requires precise evidence extraction from domain-specific text.

4.3 RULER and SCROLLS: The Limits of Transfer

RULER (Hsieh et al., 2024) and SCROLLS (Shaham et al., 2022) represent more challenging

Configuration	Trainable Params	Local@512 PPL	Gap Closed
M=4, gate-on, ϕ(q)-on, L=4	45,096	38.3	+6.0%
M=1, gate-off, L=4	18,448	38.4	+6.0%
M=4, gate-off, L=4	43,048	38.4	+6.0%
M=1, gate-off, L=8	36,896	38.7	**+8.7%**

Table 7: **Ablation of design components on LLaMA-1B.** All variants trained at 512 tokens with $w = 128$. The simplified M=1, gate-off configuration matches or exceeds the full design. Patch-layer count provides the most significant lever.

Position Range	NLL (Δ MSRA - Local)	% Positive To
Early (0–25%)	+0.008	100%
Mid-Early (25–50%)	+0.023	100%
Mid-Late (50–75%)	+0.031	100%
Late (75–100%)	+0.031	100%

Table 8: **Per-position NLL improvement at 1,024 tokens.** All position ranges show positive mean improvement. The benefit increases modestly from early to late positions.

grounding benchmarks. As shown in Table 6, MSRA trained on language modeling data provides no consistent benefit on these tasks. The spectral correction was optimized for perplexity reduction on narrative text; the resulting smoothed representations do not transfer to the precise token-level matching required by QA-style evaluation.

This result, though negative, carries an important practical message: parameter-efficient corrections optimized for one grounding objective do not automatically transfer to others. The evaluation protocol must match the training objective, a consideration particularly relevant to the efficient grounding theme of this workshop.

5 Mechanism Diagnostics

Why does the correction work where it does, and why does it fail where it fails? We address this through systematic ablation and diagnostic analysis.

5.1 What Components Actually Matter?

Table 7 presents a controlled ablation of the key design choices in MSRA. The results contain a clear message: the number of experts, the presence of query-conditioned gating, and the $\phi(q)$ term in the denominator all have negligible impact on performance. At $L = 4$, all configurations achieve identical gap closure of 6.0%. Increasing the number of patched layers from four to eight raises this to 8.7%, representing the only design choice that materially affects outcomes.

5.2 Gate Analysis: Query-Conditioning Is Real but Weak

We measure gate entropy — the uniformity of expert mixing weights across tokens — on a trained M=4 checkpoint. The mean entropy is 1.311, which is 94.5% of the maximum possible value of $\ln 4 \approx 1.386$. This indicates that the gate distributes weight nearly uniformly across experts at almost every token position.

To verify that the gate retains some query sensitivity, we perturb the final five tokens of a sequence while keeping the preceding context fixed. The maximum absolute change in gate weights is 0.046 (on a scale where uniform weights are 0.25 per expert), confirming that the gate does respond to query changes — but the magnitude of this response is too small to create meaningful expert specialization.

This finding explains the ablation results: since the gate barely differentiates between experts, removing it entirely or reducing to a single expert incurs no performance penalty.

5.3 Per-Position Analysis: Corrections Are Broad, Not Targeted

Table 8 decomposes the perplexity improvement by token position. Every position quartile shows consistent positive gain, with 100% of tokens in each range benefiting from the correction. The improvement increases from early positions (+0.008 NLL) to late positions (+0.031 NLL), consistent with the hypothesis that later tokens have more far context to recover.

Critically, when we train at 512 tokens and evaluate at 1,024 tokens — a 2× length generalization — 90.6% of late-position tokens still show positive improvement. This out-of-distribution generalization suggests that the spectral correction learns a robust, position-agnostic pattern rather than memorizing fixed-position adjustments.

5.4 What the Correction Is — and Isn't

Synthesizing these diagnostics, MSRA behaves primarily like a statistical smoother rather than a retrieval mechanism in our experimental settings. The far branch learns to add a smoothed, long-

Seq Length	Local (tok/s)	+ MSRA (tok/s)	Overhead
512	34,254	3,738	89%
1,024	27,083	3,706	86%
2,048	17,931	3,664	80%
4,096	11,535	3,188	72%

Table 9: **Inference throughput on LLaMA-1B, RTX 4090.** MSRA with M=1, L=8. The spectral branch adds 72–89% latency overhead in the current unoptimized implementation.

range contextual bias that modestly improves token prediction, particularly for tokens that benefit from non-local statistical regularities. It does not, however, enable the model to selectively attend to specific distant tokens.

This interpretation is consistent with every observation: the near-uniform gate (no selective routing), the near-but-not-far NIAH pattern (local smoothing without distant precision), the failure on QA tasks (no precise evidence extraction), and the consistent but modest perplexity gains (broad statistical improvement rather than targeted retrieval).

6 Practical Tradeoffs

Table 9 reports inference throughput for MSRA on a consumer RTX 4090 GPU. The current implementation adds 72–89% latency overhead relative to local-only attention, primarily due to the FP32 FFT operations and the per-expert convolution loop. Buffer preallocation and kernel FFT caching provide modest improvements at short sequence lengths ($2\times$ at 512 tokens, measured on the spectral module alone), but the end-to-end overhead remains substantial.

We emphasize that MSRA in its current form is a research prototype demonstrating the *feasibility* of spectral residual correction. The overhead could be reduced through several engineering improvements: replacing the per-expert Python loop with batched operations, using mixed-precision FFT when hardware support matures, and reducing the feature rank r with minimal quality impact. These optimizations are left to future work.

7 Related Work

Our work sits at the intersection of efficient attention mechanisms, spectral sequence modeling, and parameter-efficient adaptation.

Local and sparse attention. Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), and the sparse transformer (Child et al., 2019) pioneered the use of structured sparsity patterns — including sliding windows — to reduce the quadratic cost of self-attention. Mistral (Jiang et al., 2023) and subsequent models demonstrated that sliding window attention alone, without global tokens, can serve as an effective default for large-scale training. Our work takes these designs as a starting point and asks whether a lightweight correction can recover the information they discard.

Linear and kernelized attention. Linear attention (Katharopoulos et al., 2020) and Performer (Choromanski et al., 2021) approximate the full attention matrix through kernel feature maps, reducing complexity from $O(n^2)$ to $O(n)$. RWKV (Peng et al., 2023) and RetNet (Sun et al., 2023) combine linearized attention with recurrent or decay-based formulations for efficient long-sequence processing. Our linear baseline experiment (Section 4) shows that naive linear attention underperforms local attention at the same parameter scale, motivating the hybrid local+spectral design.

Spectral and frequency-domain methods. FNet (Lee-Thorp et al., 2022) replaced self-attention entirely with Fourier transforms, demonstrating that frequency-domain mixing alone can be competitive for certain tasks. Conv-Basis (Liang et al., 2025), the closest prior work, approximates attention as a sum of convolution matrices computed via FFT. MSRA differs in preserving exact local attention while adding spectral correction only for far-context regions. State-space models (Gu et al., 2022; Fu et al., 2023) provide an alternative approach to long-range sequence modeling through structured convolutional parameterizations.

Parameter-efficient adaptation. LoRA (Hu et al., 2022) and related PEFT methods (He et al., 2022; Liu et al., 2022) showed that large pretrained models can be adapted with very few trainable parameters. MSRA shares this philosophy: only the spectral branch parameters are trained, while the base model remains frozen. At 4.6–18K parameters, MSRA is smaller than even a rank-1 LoRA adapter applied to attention projections.

Evaluation and diagnosis. Our evaluation strategy is informed by the growing emphasis on rigorous benchmarking of long-context models (Bai et al., 2025; Hsieh et al., 2024; Shaham et al., 2022), as well as recent calls for including negative results and failure analysis in NLP research (Rogers et al., 2021; Dodge et al., 2019).

8 Limitations

We document the following limitations explicitly, both to scope the contribution and to guide future work.

Far-context retrieval. MSRA does not enable selective long-range retrieval. NIAH experiments show zero recovery beyond the local window, and the method cannot substitute for full attention on tasks demanding precise distant evidence extraction.

Task transfer. The correction is optimized for language modeling perplexity and does not transfer reliably to QA or instruction-following benchmarks. RULER and SCROLLS results are either mixed or negative.

Runtime overhead. The current implementation adds 72–89% latency overhead. While we have demonstrated engineering paths to reduction (buffer caching, batched convolutions), MSRA is not currently competitive with local attention in terms of throughput.

Model coverage. Our strongest results are on GPT-2 and LLaMA-1B. On DeepSeek-R1-7B, we observe no meaningful gap between local and full attention at the tested sequence lengths, preventing meaningful evaluation of the correction.

Benchmark scope. Our multi-task evaluation is weighted toward English-language tasks and perplexity-based scoring for base language models. Performance on generation, summarization, or instruction-following tasks with fine-tuned models remains unexplored.

9 Conclusion

We have presented MSRA, a parameter-efficient spectral correction for local attention in base language models, and conducted a systematic diagnosis of where it helps and where it does not. The correction provides consistent, modest perplexity improvements across architectures and training lengths, with a parameter budget as low as 4,600 tokens. Mechanism analysis suggests that the effective operation is more consistent with statistical smoothing than explicit retrieval — a finding that constrains the scope of the contribution but also clarifies the boundary conditions for spectral attention methods. For the GroundLM community, our results offer a concrete case study in diagnosing the gap between a method's intended mechanism and its empirical behavior. The work speaks to several workshop themes simultaneously: it demonstrates that parameter-efficient architectural modifications can improve language modeling (efficiency), that these improvements do not automatically translate to faithful evidence retrieval (faithfulness), and that systematically mapping failure boundaries across diverse benchmarks provides actionable insights for practitioners (evaluation). We hope the honest characterization of both strengths and failures proves useful for future work on efficient, faithful grounding in language models.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

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

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

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *International Conference on Learning Representations*.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP-IJCNLP*.

Daniel Y. Fu, Khaled K. Saab, Aditya Thomas, Tri Dao, Atri Rudra, and Christopher Re. 2023. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*.

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

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Cheng-Ping Hsieh, Chun-Liang Li, Siyan Yeh, Hadi Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, and Chen-Yu Lee. 2024. RULER: What's the real context size of your long-context language models? In *arXiv preprint arXiv:2404.06654*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. 2022. FNet: Mixing tokens with fourier transforms. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Yingyu Liang, Heshan Liu, Zhenmei Shi, Zhao Song, Zhuoyan Xu, Jiale Zhao, and Zhen Zhuang. 2025. Conv-basis: A new paradigm for efficient attention inference and gradient computation in transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, and 15 others. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. 2023. Toeplitz neural network for sequence modeling. In *International Conference on Learning Representations*.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy Lillicrap. 2020. Compressing transformer language models with compressive transformers. In *International Conference on Learning Representations*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. In *Transactions of the Association for Computational Linguistics*.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized compaRison over long language sequences. In *Conference on Empirical Methods in Natural Language Processing*.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

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