Delta Attention

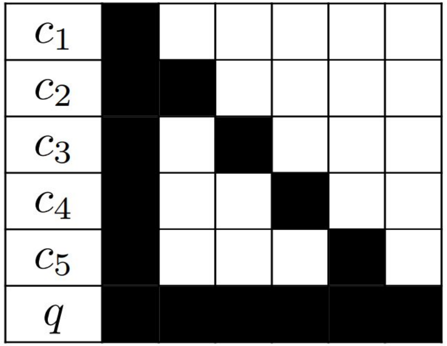
NeurIPS 2025

Jeffrey Willette[1], Heejun Lee[1], Sung Ju Hwang[1-2]

1. KAIST 2. Deepauto.ai

NEURAL INFORMATION PROCESSING SYSTEMS

Delta Attention - Motivation



Star Attention [1]

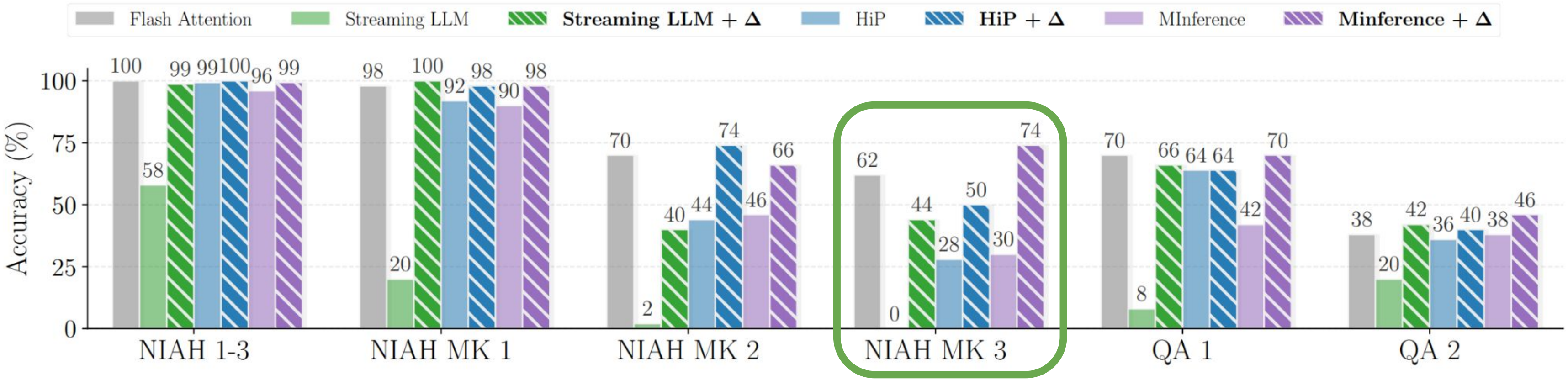
- **Linear prefill** (like Streaming LLM)
- Designed with distributed RAG in mind.
- Dense decode **maintains all tokens in the KV cache** for possible future usage.

- When working with sparse prefill / dense decode we made a surprising discovery
- Even if **relevant context fits in the window** (like NIAH tasks)
- The dense decode can **fail to find the relevant tokens**

[1] Acharya, S., Jia, F., & Ginsburg, B. (2024). Star attention: Efficient llm inference over long sequences. arXiv preprint arXiv:2411.17116.

Delta Attention - Motivation

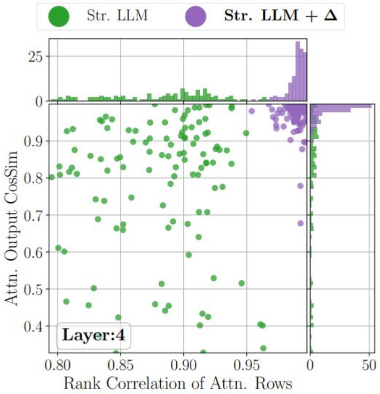
- However, we discovered a simple and effective way to fix this with minimal added overhead.
- Our method works in conjunction with existing sparse attention kernels, for example in RULER MK3 we increase Streaming LLM from 0% → 44% accuracy.
- In the standard setting, we maintain 98.5% sparsity, which makes our model 32X faster than Flash Attention 2 at 1M token prefills.



[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Delta Attention - Introduction

Sparse Attention introduces a shift in the attention distribution and thus a shift in the distribution of cosine similarities of the outputs.



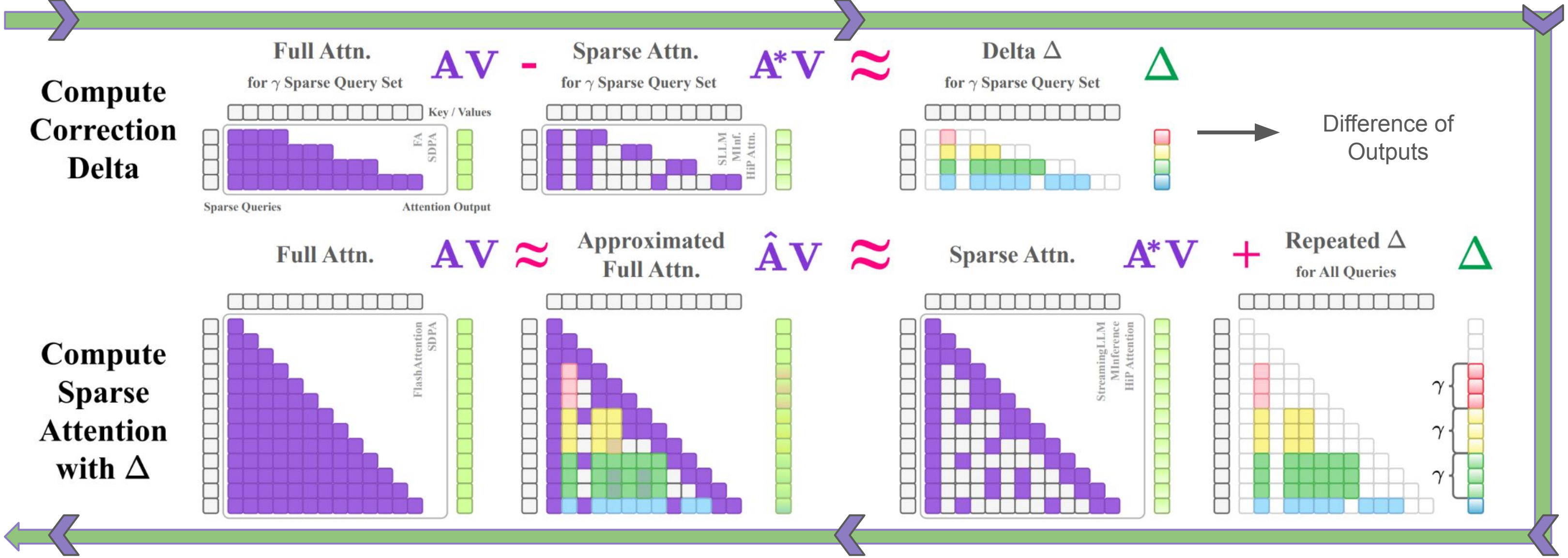
- Both methods compute cosine similarity and rank correlation with respect to full attention.
- Therefore, a cluster in the top-right corner (1, 1) would be a good approximation to full attention.
- Streaming LLM causes a large shift from full attention.
- **Thought experiment: How does this distribution affect the transition from sparse prefill to dense decode?**

From layer (l) to layer (l+1), the outputs determine the (Q, K, V) matrices of the next layer. This means that **decode queries will no longer align with prefill keys and therefore will fail to retrieve important context from the prompt**…

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Delta Attention - Method

Delta Attention assumed we have access to some existing key-sparse attention kernel such as Streaming LLM [3], MInference [4], or HiP Attention [5].



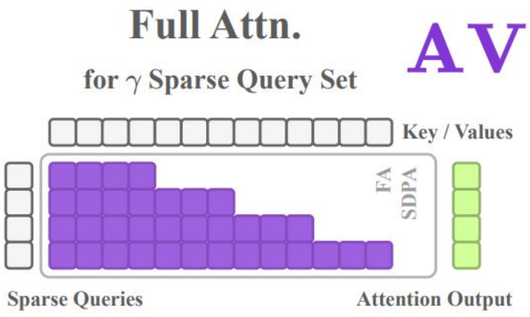
1. We perform query sparse attention **2.** We take a difference of outputs with an arbitrary sparse attention (making a delta) **3.** Repeat the delta and sum with sparse attention output.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

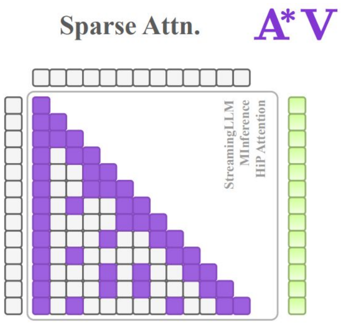
Delta Attention - Method

This requires only two attention kernel calls. **1.** A query sparse attention kernel and **2.** An arbitrary key-sparse attention kernel.

1. Query Sparse Attention



1. Key Sparse Attention

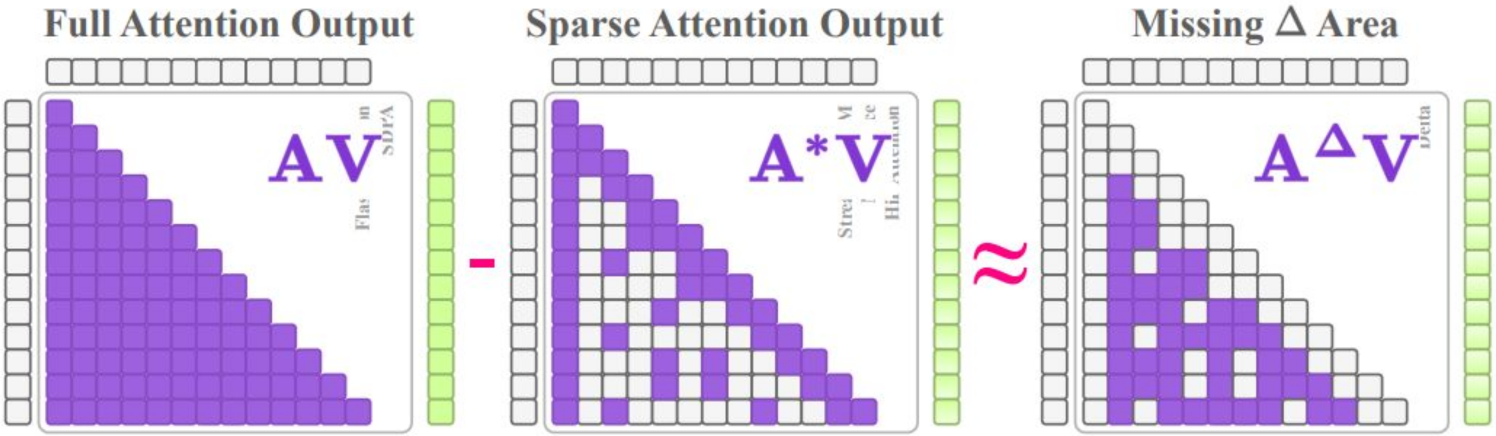


After these two kernel calls, everything else is simple and lightweight sum in the output space. **Therefore, delta attention creates minimal added overhead.**

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Delta Attention - Method

Intuitively, we view the difference of dense/sparse outputs as representing the portion of the attention which is missing from the sparse output



Therefore, as there is likely not a large change in the queries and sparse mask from q_i to q_{i+1}, we may re-use this "missing attention" in subsequent rows of the outputs.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Delta Attention - Method

We study the analytical difference between the delta correction and the “missing attention” region. With the sorted attention scores and **H** being the head attention scores (smaller) and **T** being the tail scores (larger).

Lemma 1. *w.l.o.g. Consider an arbitrary row in the attention matrix* $\mathbf{a}$ *and arbitrary column of the values* $\mathbf{v}$*, with both* $\mathbf{a}$ *and* $\mathbf{v}$ *being sorted according to rank of* $\mathbf{a}$ *such that* $\mathbf{a} = (a_{r(1)} \leq a_{r(2)} \leq \cdots \leq a_{r(N)})$*. For a top-*$k$ *sparse attention matrix which only computes the top-*k *attention scores, one only needs to compute* $\mathbf{a^*}^\top \mathbf{v} = \sum_{N-k+1}^{N} \mathbf{a^*}_i \mathbf{v}_i$*. With* $\mathbf{\Delta} = \mathbf{a}^\top \mathbf{v} - \mathbf{a^*}^\top \mathbf{v}$*, we may bound the error of our attention approximation as,*

$$\left| \mathbf{\Delta} - \sum_{i=1}^{N-k} \mathbf{a}_i \, \mathbf{v}_i \right| \leq \frac{H}{H+T} \max_{i > N-k} |v_i|$$

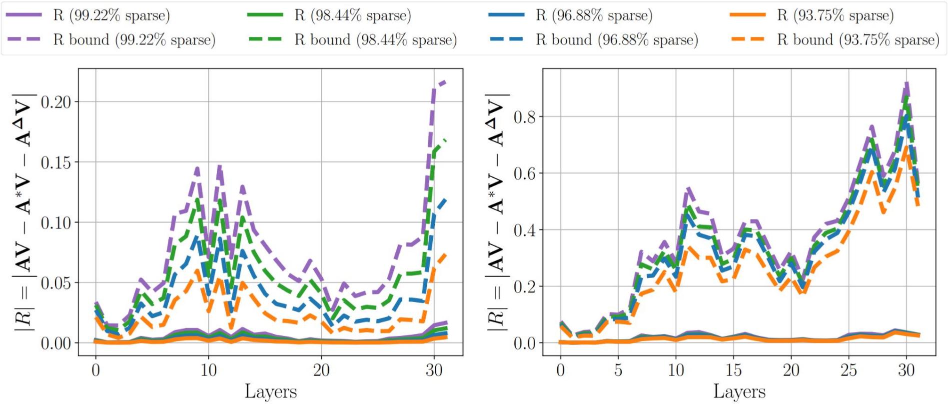
Proof. See Appendix G. □

We found that the difference is bounded by the value vector magnitude and a function of the normalization constants of the sparse and dense attention matrices.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Delta Attention - Method

Looking at the bound given in the previous slide, we studied it empirically for both an oracle top-k sparse attention, and Streaming LLM.



(a) Oracle top-k sparse attention.

(b) Streaming LLM.

We found that the real empirical error tends to be much lower than the upper bound.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Positive Attributes of Delta Attention

- Can be applied to **existing sparse attention methods**.
- Performs well when calculating 1/64 queries **(98.5% sparsity)**
- **32X faster than Flash Attention 2 for prefilling 1M** tokens with Streaming LLM.
- At RULER 131K with Str. LLM, **we recover 95% of the accuracy of full attention (up from 31%)**
- Since we work in the attention output space, **we can utilize all existing low-level sparse attention kernels** (MInference, HiP, Str. LLM, etc.)
- **Very simple to implement**, and only requires a simply modified flash attention kernel for the query sparse attention.

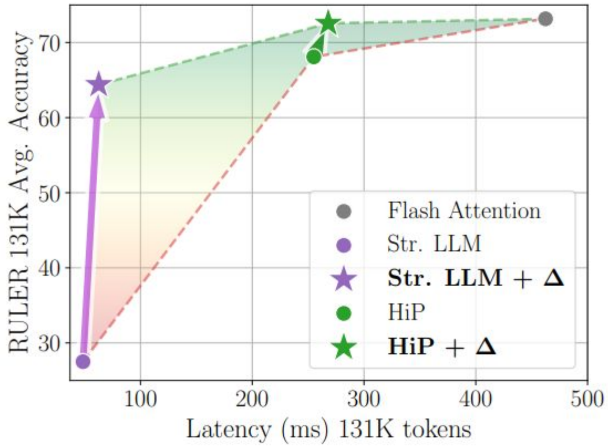
[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

RULER: Long Context Needle-In-a-Haystack Tasks

- Overall, Delta Attention is better.
- **(Left)** Largest increase in performance is always at the longest context length.
- **(Right)** Performance gains incur minimal added latency overhead.

Model	Llama 3.1 8B Instruct										Mistral NeMo 12B				
Attn. Method	Flash Attn.	Str. LLM	Str. LLM	Str. LLM	Str. LLM	**Str. LLM+Δ**	MInf.	**MInf.+Δ**	HiP	**HiP+Δ**	Flash Attn.	Str. LLM	**Str. LLM+Δ**	HiP	**HiP+Δ**
Wind.	-	2K	4K	16K	32K	2K	3K	3K	3K	3K	-	2K	2k	3K	3K
4K	96.74	90.52	96.71	96.71	96.71	96.54	96.74	96.71	96.78	96.82	90.60	71.01	90.42	90.36	90.55
8K	93.25	60.53	93.76	93.76	93.76	92.25	93.65	93.69	94.44	94.32	87.67	44.89	85.38	88.36	87.69
16K	90.99	38.13	68.07	91.15	91.15	88.66	92.32	91.34	94.02	94.02	81.82	33.28	78.07	78.07	81.08
32K	85.84	30.25	43.38	56.32	85.83	81.27	86.75	85.96	86.75	91.12	62.54	12.27	34.76	58.76	60.38
65K	85.25	18.59	34.08	41.28	58.35	75.22	84.43	83.67	79.30	82.91	46.89	03.28	16.22	35.87	41.56
131K	73.16	27.45	30.32	40.51	49.17	64.40	65.73	73.31	68.09	72.56	18.09	02.25	01.44	10.10	10.93
Avg.	87.54	44.25	61.05	69.96	79.16	**83.06**	86.60	**87.44**	86.56	**88.62**	64.60	27.83	**51.05**	60.25	**62.03**

RULER Benchmark

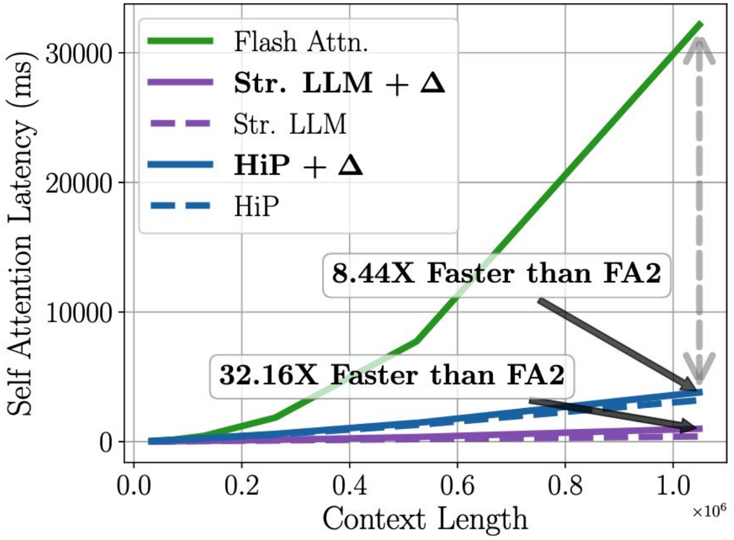


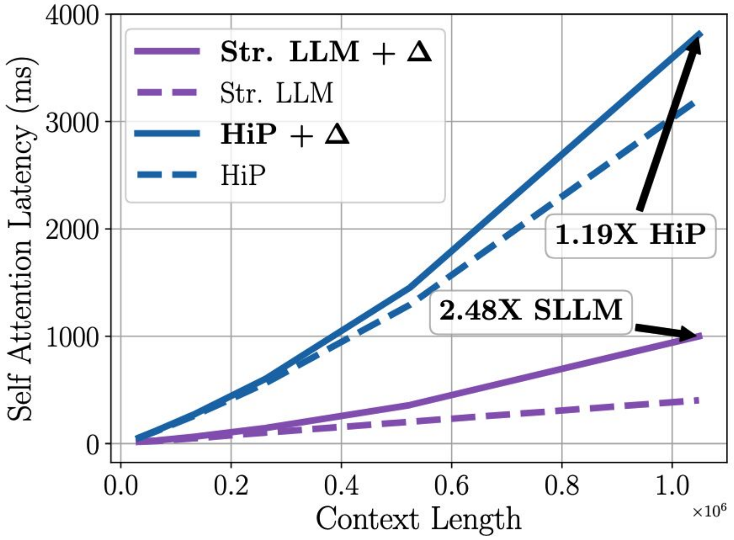
RULER 131K vs Latency

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Delta Attention: Latency Overhead

- We add latency overhead because we use (sparse attn. + delta correction).
- However, the latency overhead is negligible.
- Str. LLM is 32X faster than FA2 at 1M.

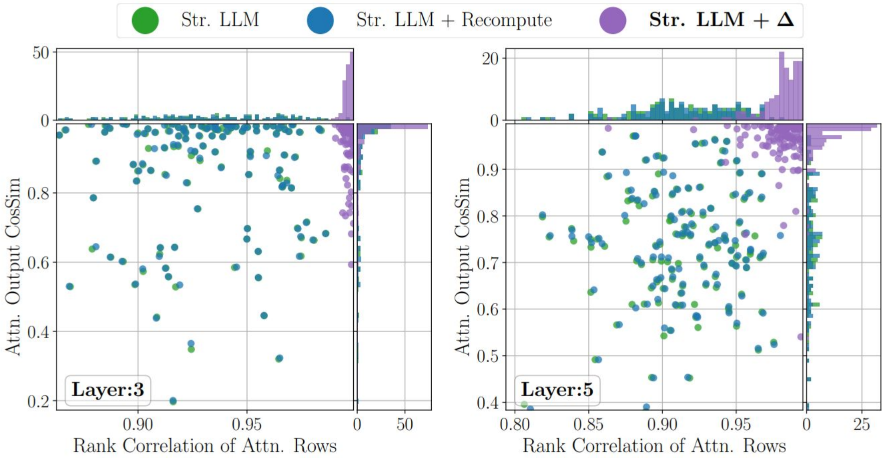




[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

Ablation: Recomputing Tokens is not Enough

Densely recomputing some tokens without making the full delta correction is not enough to fix the shift in attention outputs.



Densely recomputing tokens without the delta correction loses 12%pp at 131K

Model	131K	65K	32K	…	Avg.
Str. LLM	27.45	18.59	30.25	…	44.25
Str. LLM + Recompute	52.67	72.71	78.39	…	79.99
Str. LLM + Δ	64.40	75.22	81.27	…	**83.06**

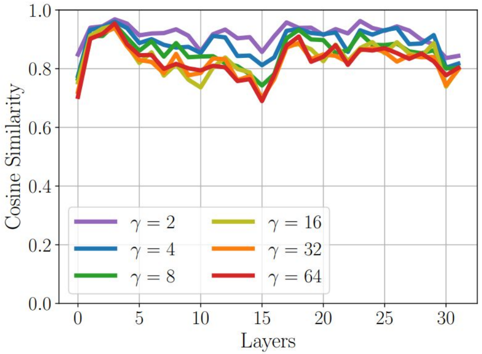
At 131K, the full delta correction still shows an approximately 12%pp accuracy improvement over 'recompute'.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

InfiniteBench and Delta Justification

Left: InfiniteBench Results show a similar pattern to RULER, Delta Attention shows a significant average improvement over all baselines.

Model	Method	Ctx Len.	En.MC	En.QA	En.QAR	En.Sum	Passkey	Number	KV	Math.F	Avg.
Llama 3.1 8B Instruct	Flash Attention	126K	64.19	35.89	44.69	31.59	99.13	99.83	92.40	24.86	61.57
	HiP	126K	54.15	31.49	38.12	31.06	75.08	96.10	30.60	18.86	46.93
	HiP + Δ	126K	61.14	33.70	43.54	31.30	100.0	97.97	69.60	25.71	**57.87**
	Str. LLM	126K	27.95	07.25	14.67	20.57	02.71	01.36	01.20	25.14	12.51
	Str. LLM + Δ	126K	56.33	24.93	33.35	26.95	96.27	68.81	00.40	25.43	**41.66**
Llama 4 Scout 109B	Flash Attention	384K	82.10	44.34	48.82	35.30	100.0	100.0	99.20	43.14	69.11
	HiP	384K	74.67	43.19	48.29	34.28	100.0	99.83	99.40	41.14	67.60
	HiP + Δ	384K	78.60	42.84	48.14	34.06	100.0	99.66	97.20	44.29	**68.10**
	Str. LLM	384K	49.78	15.23	26.11	31.50	52.88	08.31	03.40	40.57	28.47
	Str. LLM + Δ	384K	73.80	37.82	43.03	30.62	94.75	91.36	46.60	40.86	**57.35**



(b) $\cos([\mathbf{A}^{\Delta}\mathbf{V}]_i, [\mathbf{A}^{\Delta}\mathbf{V}]_{i+\nu})$

Right: Computing the delta correction on all rows (no skipping) we find a high cosine similarity within a gamma window, justifying reusing the delta within those rows.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

References

[1] Acharya, S., Jia, F., & Ginsburg, B. (2024). Star attention: Efficient llm inference over long sequences. arXiv preprint arXiv:2411.17116.

[2] Willette, J., Lee, H., & Hwang, S. J. (2025). Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. arXiv preprint arXiv:2505.11254.

[3] Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2023). Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453.

[4] Jiang, H., Li, Y., Zhang, C., Wu, Q., Luo, X., Ahn, S., ... & Qiu, L. (2024). Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. Advances in Neural Information Processing Systems, 37, 52481-52515.

[5] Lee, H., Park, G., Lee, Y., Kim, J., Jeong, W., Jeon, M., & Hwang, S. J. (2024). HiP attention: Sparse sub-quadratic attention with hierarchical attention pruning. arXiv e-prints, arXiv-2406.