Differential Transformer

Tianzhu Ye*[1], Li Dong*[2], Yuqing Xia*[2], Yutao Sun*[1]

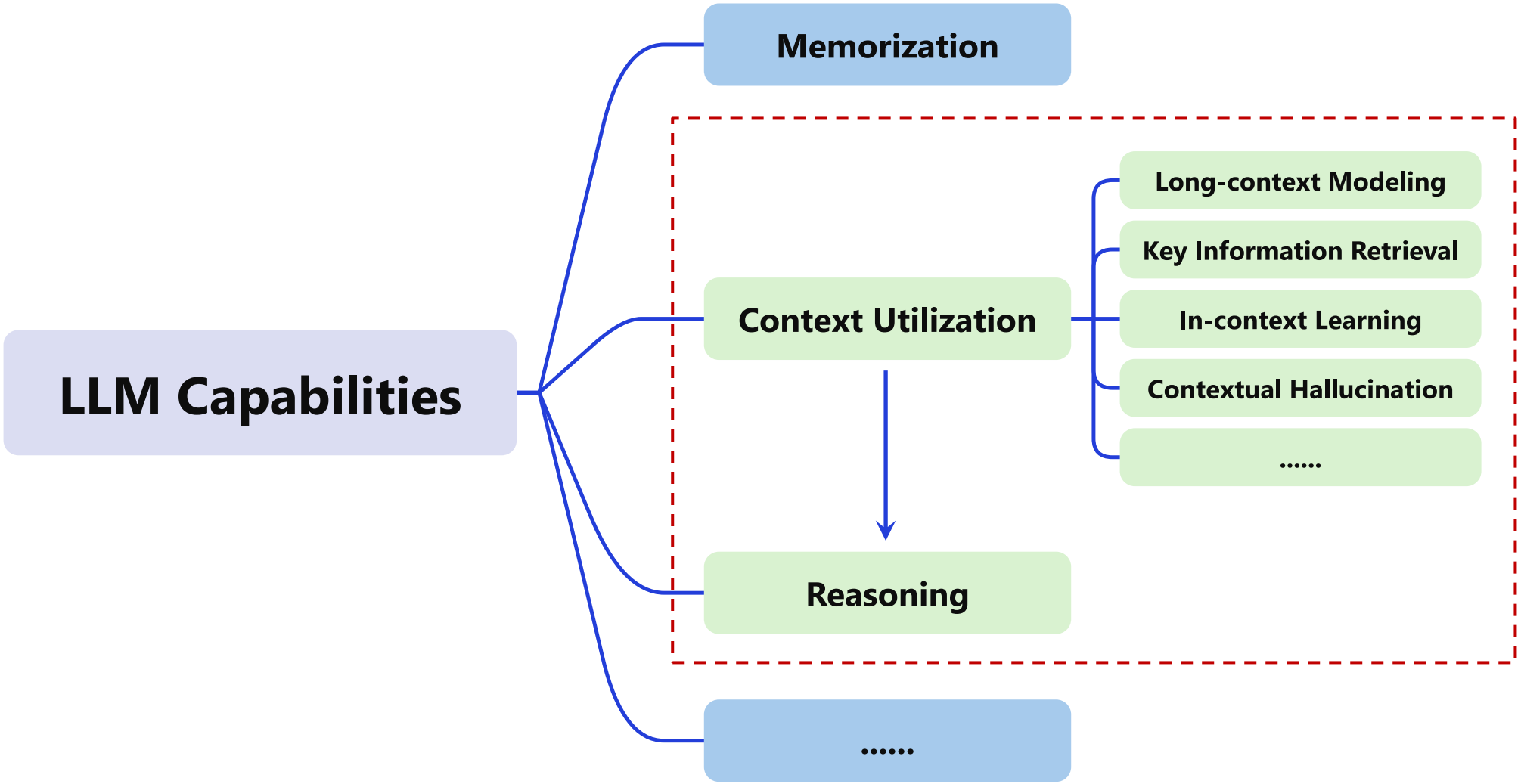
Yi Zhu[2], Gao Huang†[1], Furu Wei†[2]

Tsinghua University[1], Microsoft Research[2]

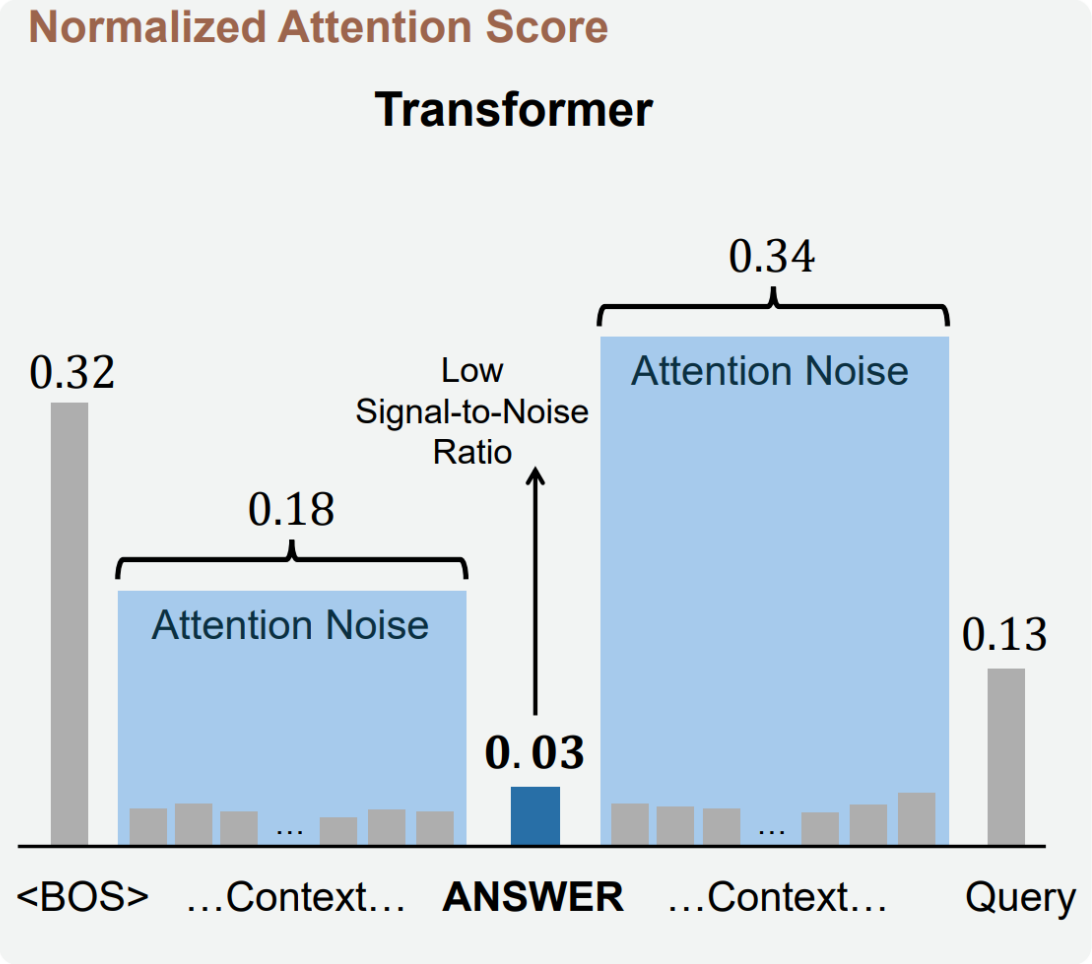
Presenter: Tianzhu Ye

ytz24@mails.tsinghua.edu.cn

LLM Capabilities



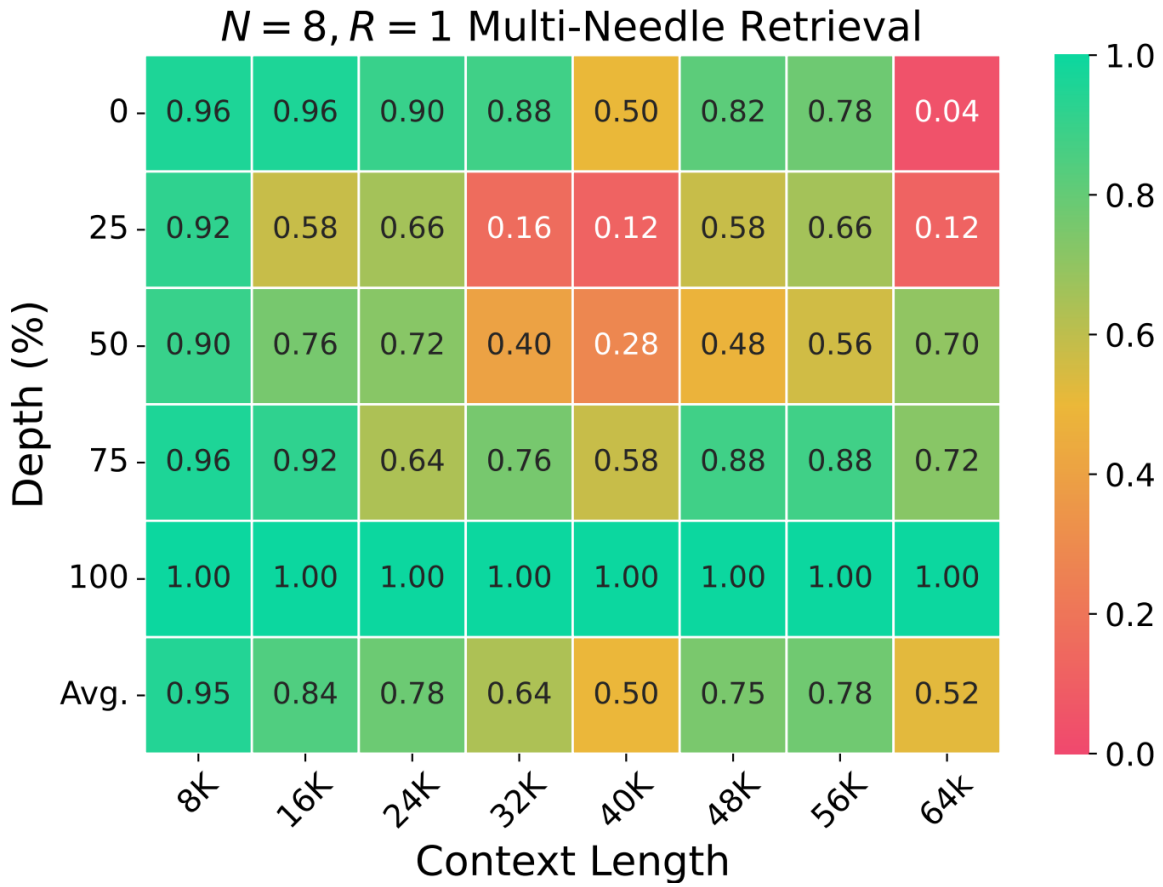
Motivation



- Key information embedded in a pile of documents
- Small attention to the answer
- Over-attend irrelevant context
- "*Attention Noise*"

Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157-173.

Motivation



- Leads to low accuracy in the retrieval task
- Caused by properties of Softmax function:

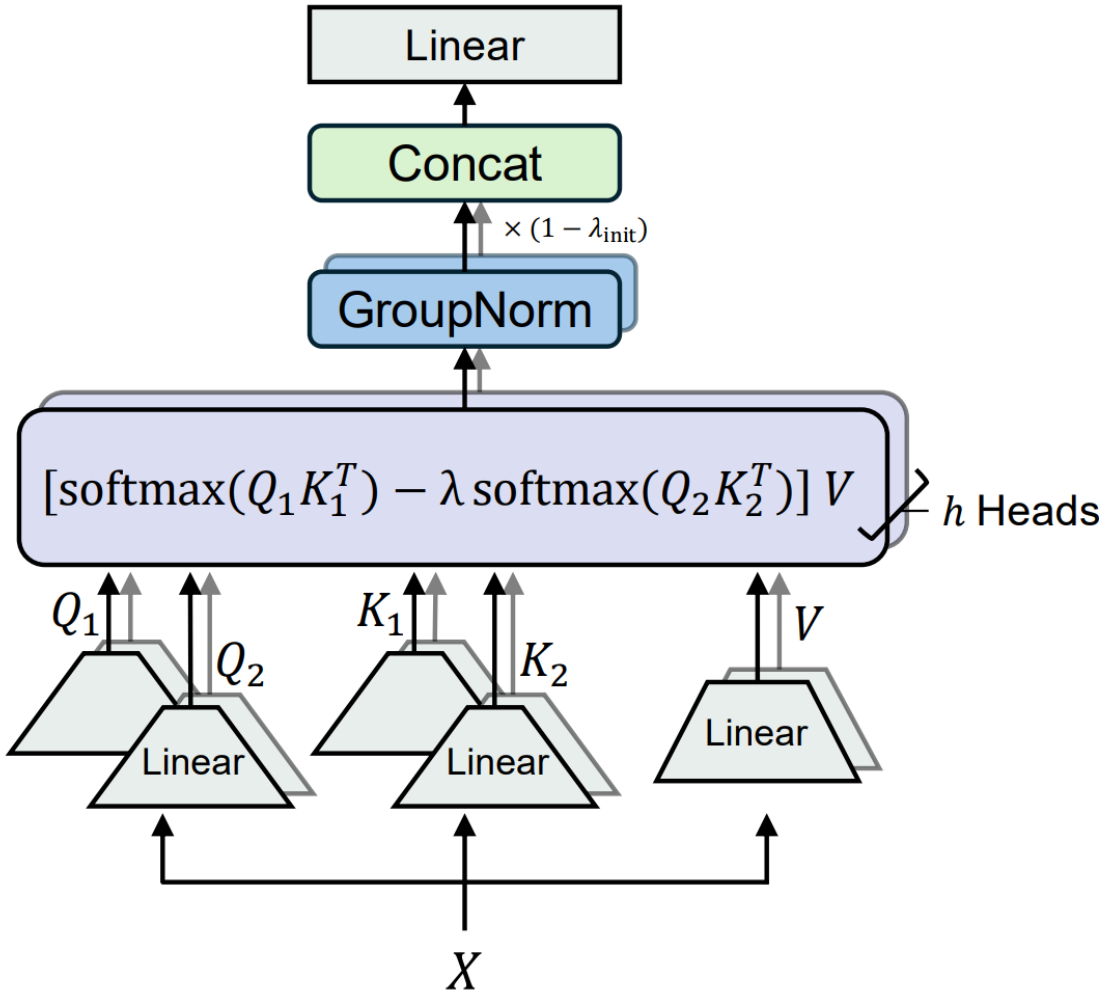
$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

I. All positive

II. Sparsity → Wide input → Instability

Veličković, Petar, et al. "softmax is not enough (for sharp out-of-distribution)." *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.
Barbero, Federico, et al. "Transformers need glasses! Information over-squashing in language tasks." *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Method



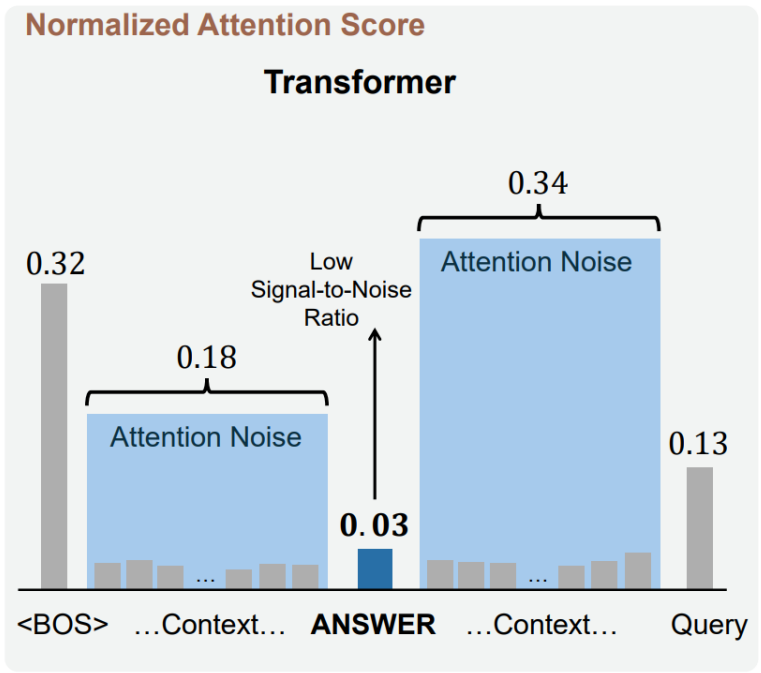
- **Differential Attention**
- Taking the difference between a pair of Softmax
- Break Softmax properties:
  I. Zero (and negative) weights

     [Sparse in forward]

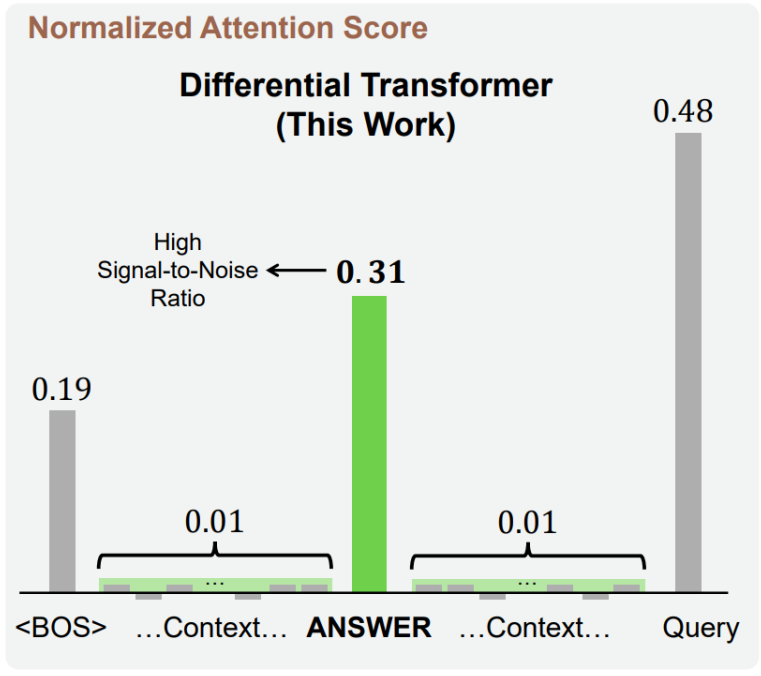
  II. No need for each Softmax to be sparse

     [Dense in backward]
- Keep gradient properties of each Softmax
- No human priors. Let the model decide

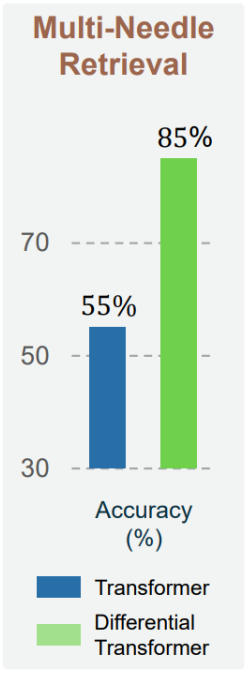
Key Information Retrieval



(a) Transformer



(b) DIFF



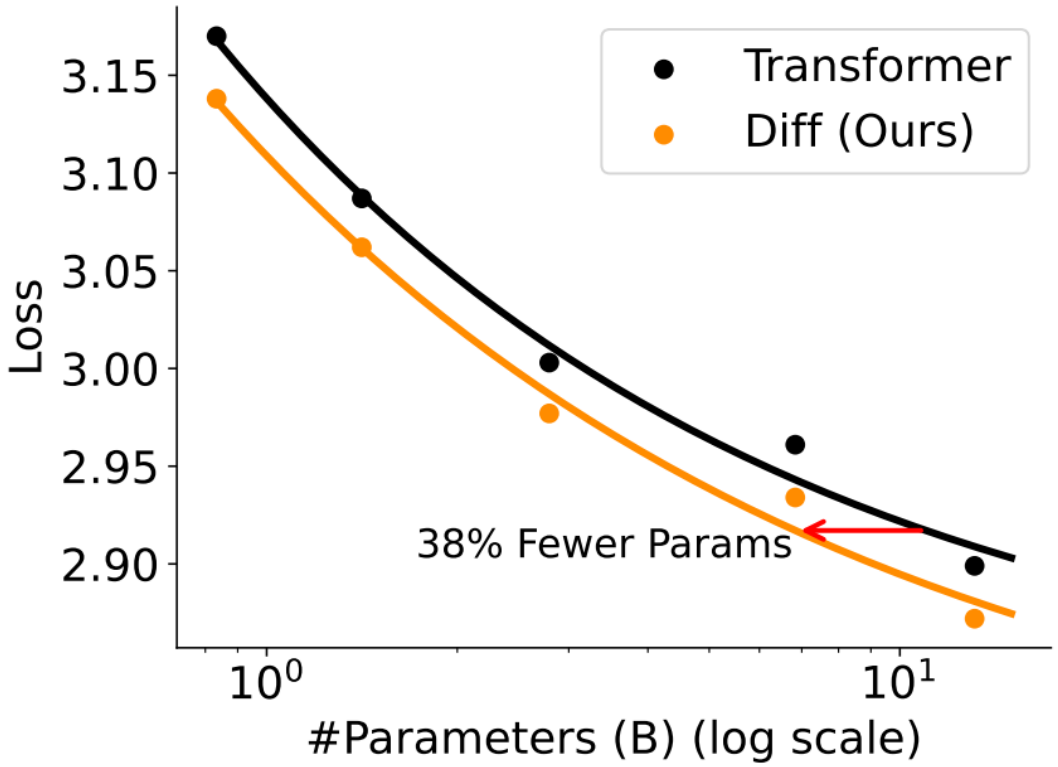
- Sparse attention pattern

Model	Attention to Answer ↑					Attention Noise ↓				
	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
Transformer	0.03	0.03	0.03	0.07	0.09	0.51	0.54	0.52	0.49	0.49
DIFF	**0.27**	**0.30**	**0.31**	**0.32**	**0.40**	**0.01**	**0.02**	**0.02**	**0.02**	**0.01**

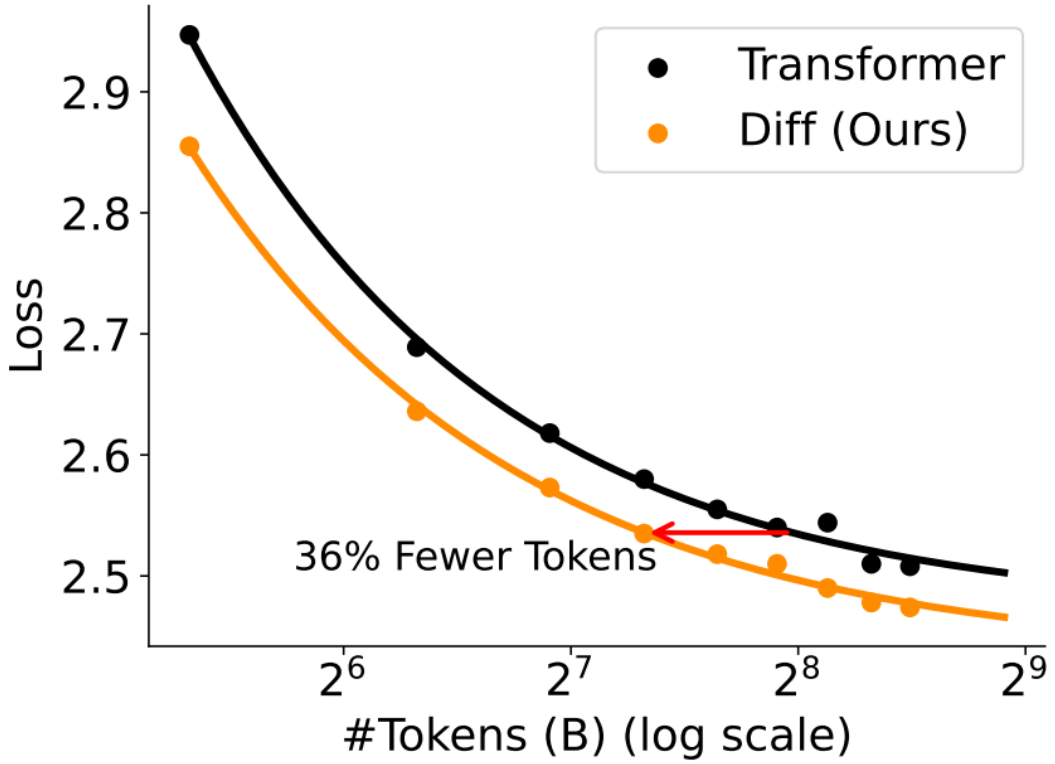
- High signal-to-noise ratio

Scaling Model Size and Training Tokens

- About **35%** fewer params or tokens compared to Transformer



(a) Scaling model size from 830M to 13B



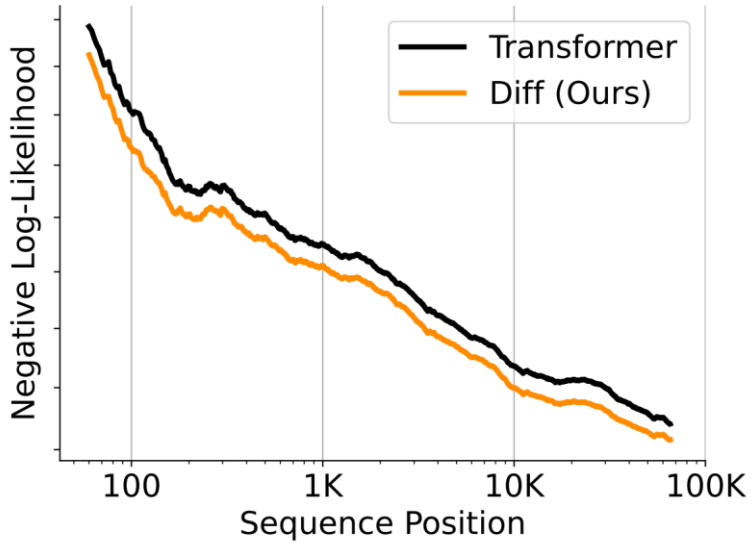
(b) Scaling training tokens of 3B model

Scaling Training Tokens and Sequence Length

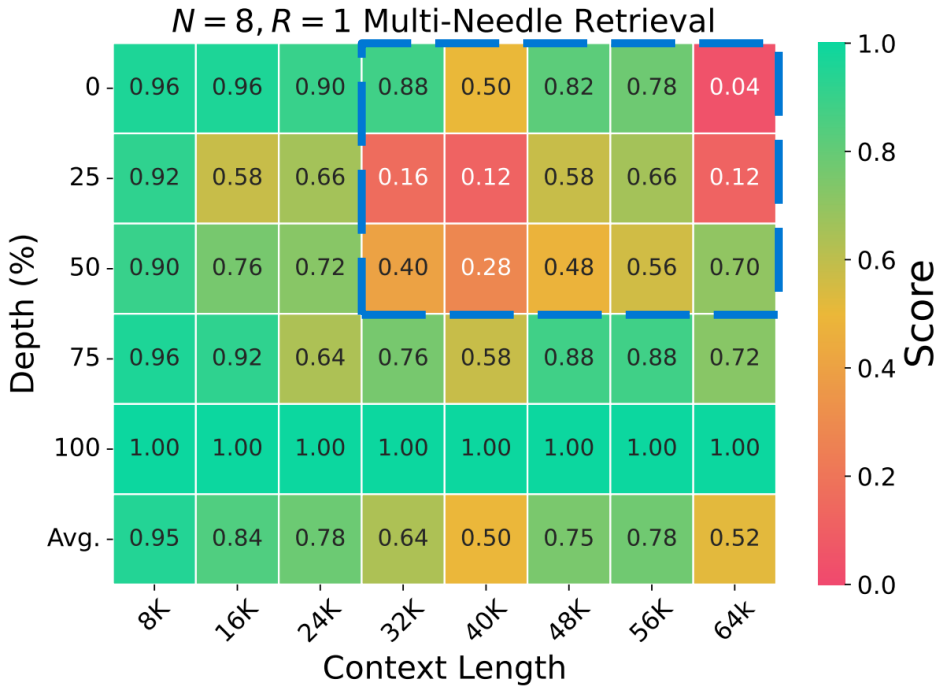
- Scaling training tokens to 1T and compare with well-trained Transformer-based models

Model	ARC-C	ARC-E	BoolQ	HellaSwag	OBQA	PIQA	WinoGrande	Avg
Training with 1T tokens								
OpenLLaMA-3B-v2 [13]	33.9	67.6	65.7	70.0	26.0	76.7	62.9	57.5
StableLM-base-alpha-3B-v2 [39]	32.4	67.3	64.6	68.6	26.4	76.0	62.1	56.8
StableLM-3B-4E1T [40]	—	66.6	—	—	—	**76.8**	63.2	—
DIFF-3B	**37.8**	**72.9**	**69.0**	**71.4**	**29.0**	**76.8**	**67.1**	**60.6**

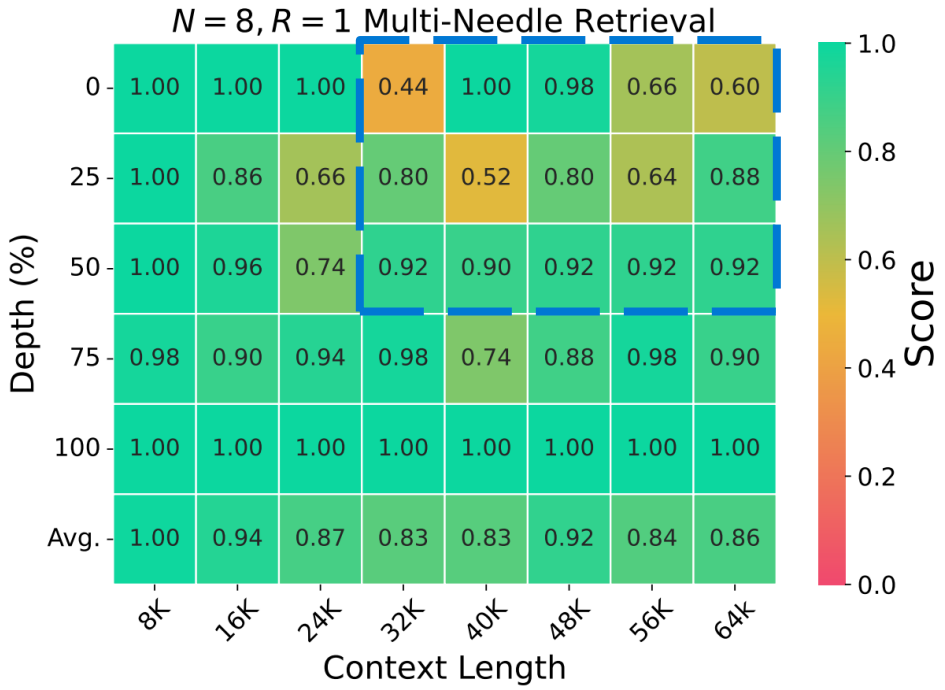
- Scaling sequence length to 64K
- Consistent lower NLL
- Leverages long context more effectively



Key Information Retrieval



(a) Transformer



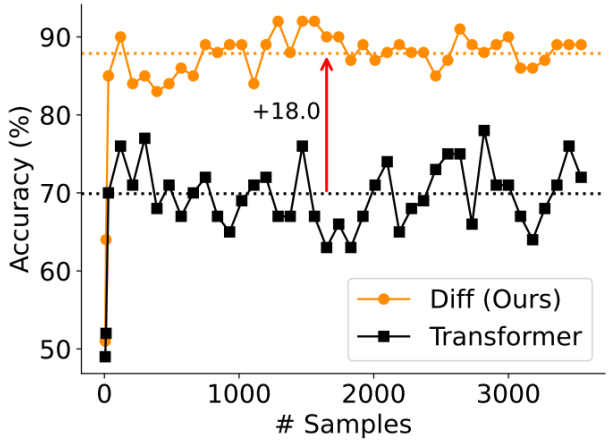
(b) DIFF

- 8K~64K length
- Particularly in the first half depth

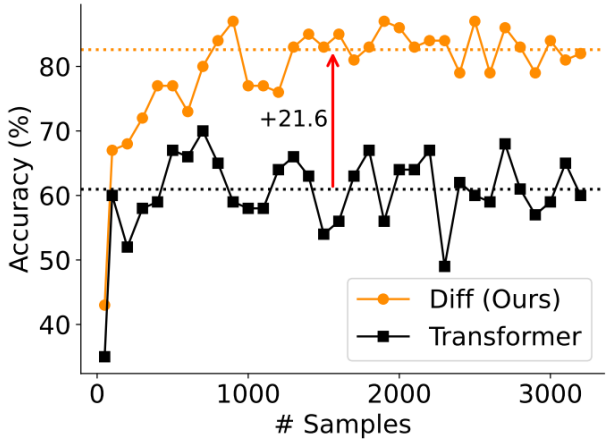
Model	$N=1$ $R=1$	$N=2$ $R=2$	$N=4$ $R=2$	$N=6$ $R=2$
Transformer	**1.00**	0.85	0.62	0.55
DIFF	**1.00**	**0.92**	**0.84**	**0.85**

- 4K length, averaged across depths
- Gains still exist in relatively short context

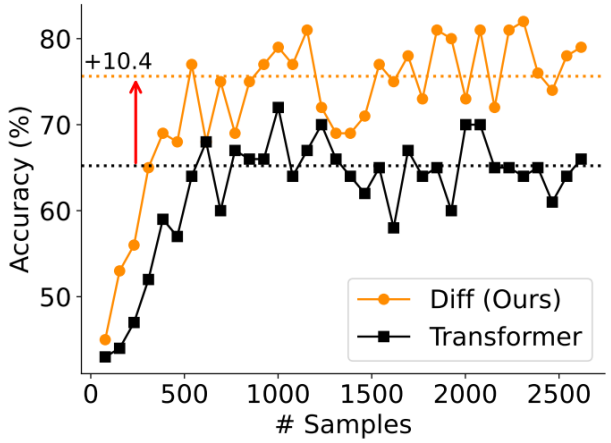
In-context Learning



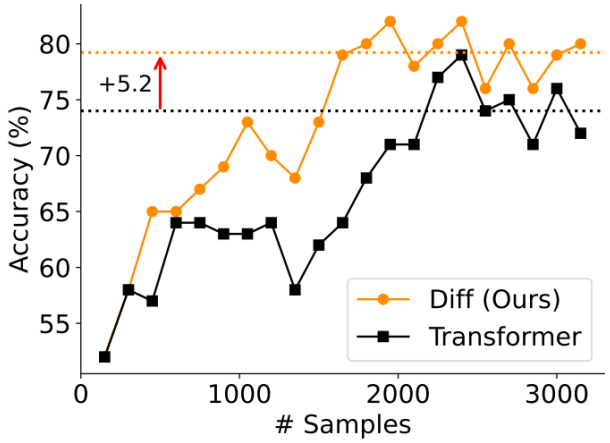
(a) TREC with 6 classes.



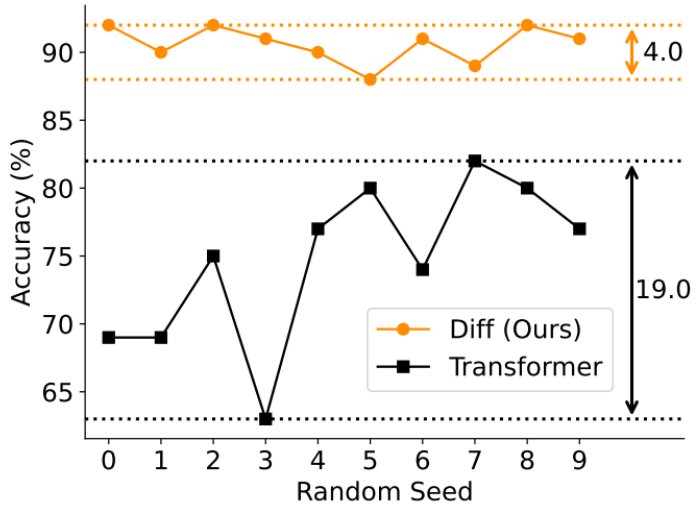
(b) TREC-fine with 50 classes.



(c) Banking-77 with 77 classes.

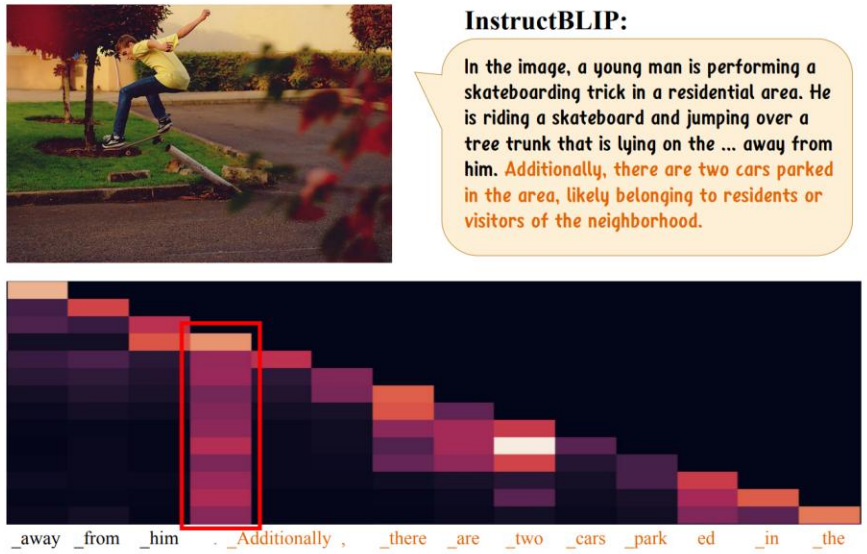


(d) Clinic-150 with 150 classes.



- Many-shot in-context learning
- Robustness to permutation

Contextual Hallucination Evaluation



- Previous research:

  hallucination may come from attention misallocation

- Contextual hallucination evaluation with

  text summarization and QA

Model	XSum	CNN/DM	MultiNews
Transformer	0.44	0.32	0.42
DIFF	**0.53**	**0.41**	**0.61**

(a) Summarization

Model	Qasper	HotpotQA	2WikiMQA
Transformer	0.28	0.36	0.29
DIFF	**0.39**	**0.46**	**0.36**

(b) QA

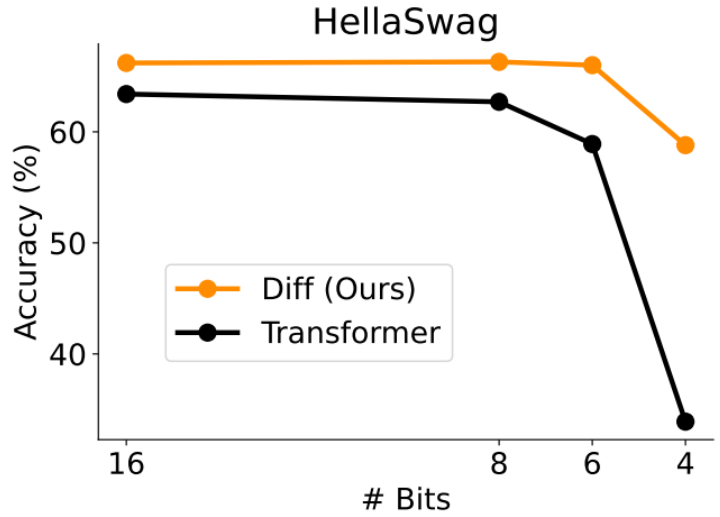
Huang, Qidong, et al. "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Activation Outliers Analysis

Model	Activation Type	Top-1	Top-2	Top-3	Top-10	Top-100	Median
Transformer	Attention Logits	318.0	308.2	304.9	284.7	251.5	5.4
DIFF	Attention Logits	38.8	38.8	37.3	32.0	27.4	3.3
Transformer	Hidden States	3608.6	3607.4	3603.6	3552.1	2448.2	0.6
DIFF	Hidden States	1688.2	1672.5	1672.1	1624.3	740.9	1.2

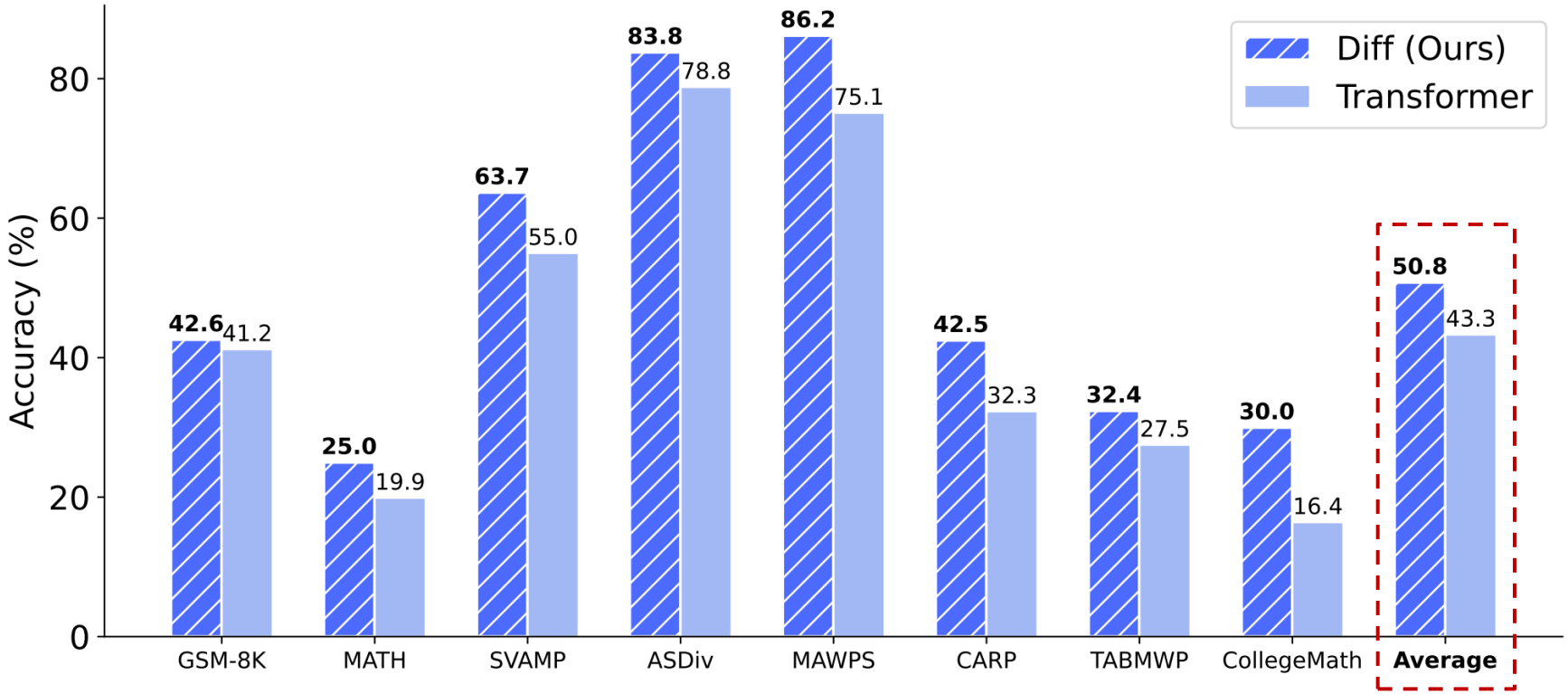
- Suppressing activation outliers
- Lower bits when quantizing attention logits
- Intuition: (observed in previous research as well)

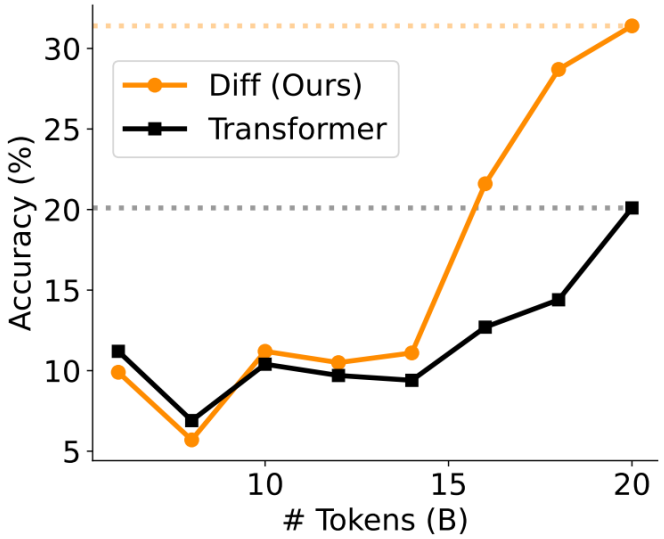
  Obtain zeros in attention scores & reduce over-attended tokens



Bondarenko, Yelysei, Markus Nagel, and Tijmen Blankevoort. "Quantizable transformers: Removing outliers by helping attention heads do nothing." *Advances in Neural Information Processing Systems* 36 (2023): 75067-75096.
Sun, Mingjie, et al. "Massive Activations in Large Language Models." *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Mathematical Reasoning Evaluation





- Two-stage fine-tuning
- Evaluation of o1-style mathematical reasoning

Indicates strong connection between context utilization and reasoning ability

Take-away Messages

- Cancel out attention noise by taking the difference between a pair of Softmax functions
- Enhance LLM's capability of utilizing context, including long-context modeling, key information retrieval, in-context learning, mathematical reasoning, mitigating contextual hallucination

Future Work

- Low-bit attention kernels
- Compress key-value caches with sparse attention patterns
- Explore applications in other modalities[1]
- More theoretical analysis[2]

1. Hammoud, Hasan Abed Al Kader, and Bernard Ghanem. "DiffCLIP: Differential Attention Meets CLIP." *arXiv preprint arXiv:2503.06626* (2025).
2. Naderi, Alireza, Thiziri Nait Saada, and Jared Tanner. "Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Transformers." *arXiv preprint arXiv:2410.07799* (2024).

- **Poster Session 4, Fri 25 Apr 3 p.m. — 5:30 p.m.**
- Code is available at https://aka.ms/Diff-Transformer
- Email: ytz24@mails.tsinghua.edu.cn
- Thank you!

Paper

Code

Contact

The presenter's attendance at this conference is generously supported by the Deng Feng Fund.