ACL2022 Findings **Meta**

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

ACL2023 Honorable Mentions **CMU&Princeton**

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

Ting-Han Fan
Princeton University

Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

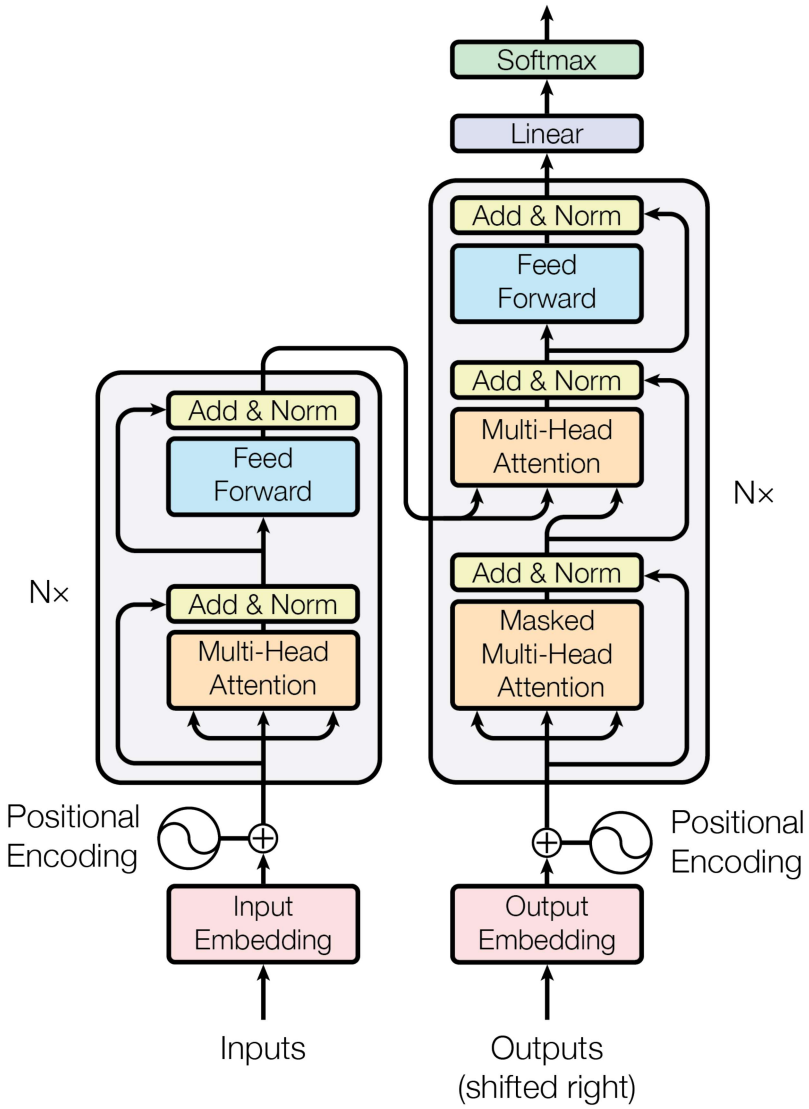
Peter J. Ramadge
Princeton University

Preprint **Mila**

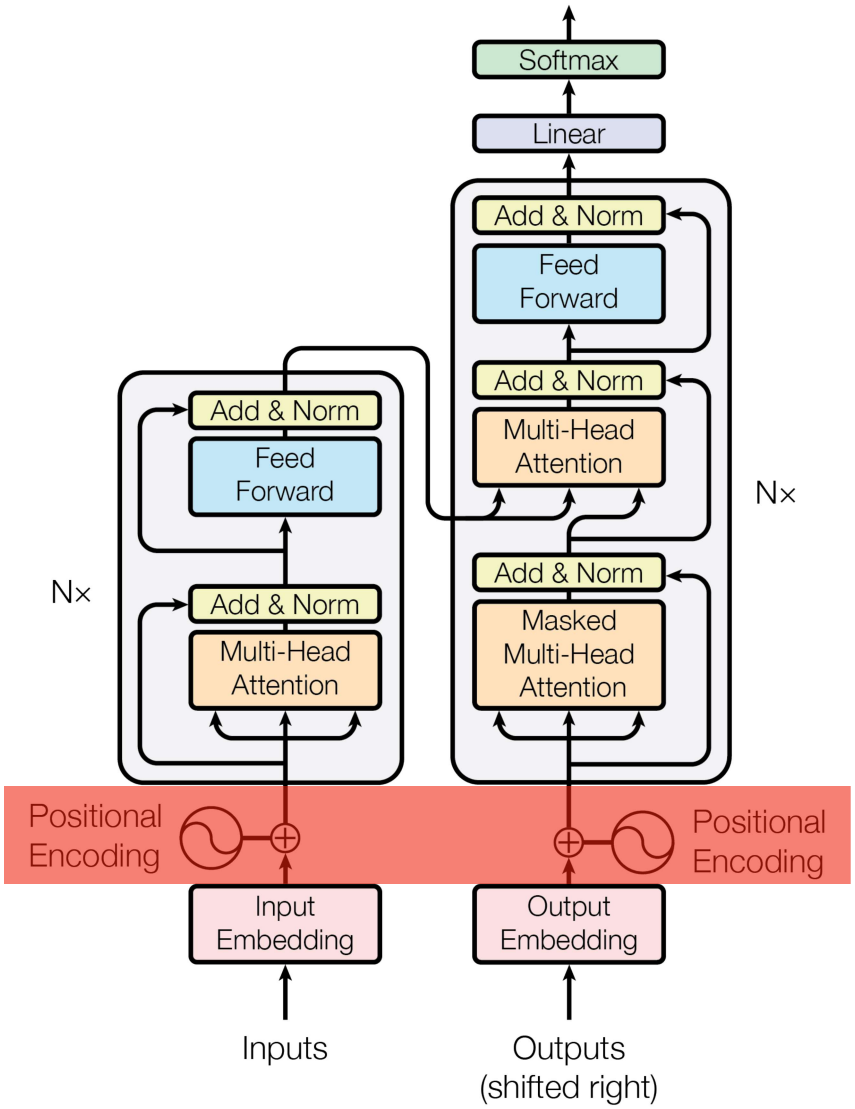
The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

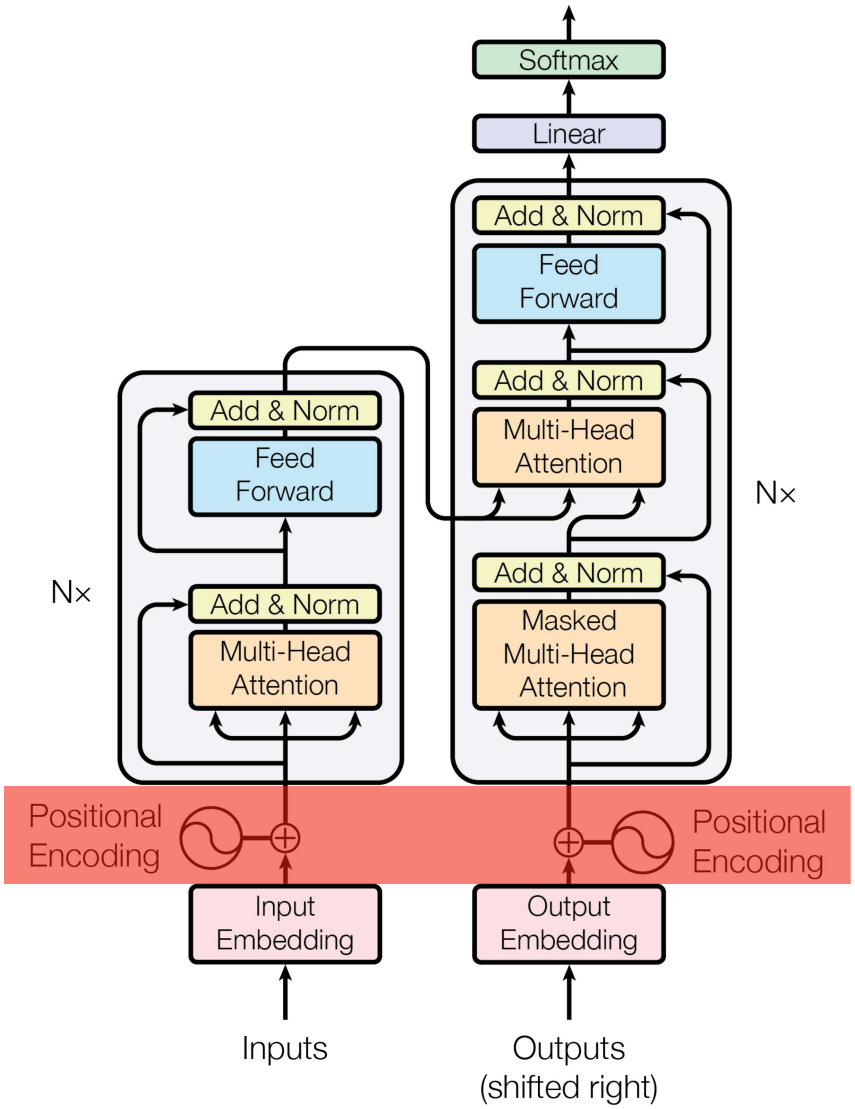
BG: PE in Transformer



BG: PE in Transformer



BG: PE in Transformer



Position Information in Transformers

纪 焘

Topic	Absolute	Reference Point: Absolute & Relative	Relative
Sequential	(Devlin et al., 2019) (Kitaev et al., 2020) (Liu et al., 2020b) (Press et al., 2020) (Wang et al., 2020) (Dehghani et al., 2019)	(Shaw et al., 2018) (Ke et al., 2020) (Dufter et al., 2020) (He et al., 2020)	(Dai et al., 2019) (Raffel et al., 2020) (Wu et al., 2020) (Huang et al., 2020) (Shen et al., 2018) (Neishi & Yoshinaga, 2019)
Sinusoidal	(Vaswani et al., 2017) (Li et al., 2019)		(Yan et al., 2019)
Graphs	(Shiv & Quirk, 2019)	(Wang et al., 2019)	(Zhu et al., 2019) (Cai & Lam, 2020) (Schmitt et al., 2021)
Decoder	(Takase & Okazaki, 2019) (Oka et al., 2020) (Bao et al., 2019)		
Crosslingual	(Artetxe et al., 2020) (Ding et al., 2020) (Liu et al., 2020a) (Liu et al., 2020c)		
Analysis	(Yang et al., 2019) (Wang & Chen, 2020)	(Rosendahl et al., 2019) (Wang et al., 2021)	

- Position Information in Transformers: An Overview

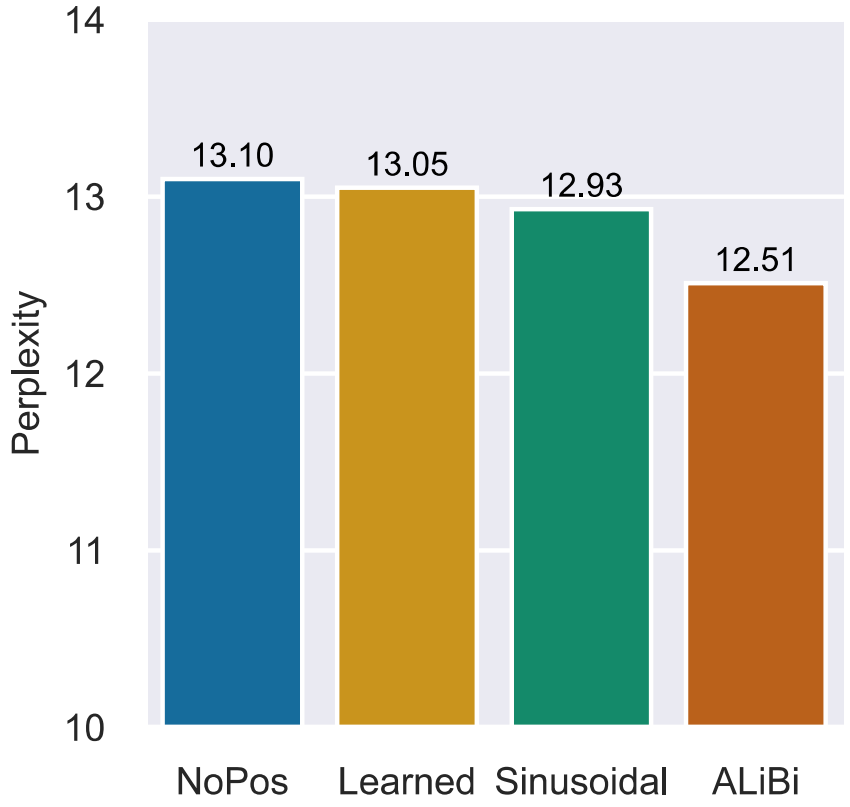
ACL2022 Findings

Transformer Language Models without Positional Encodings Still Learn Positional Information

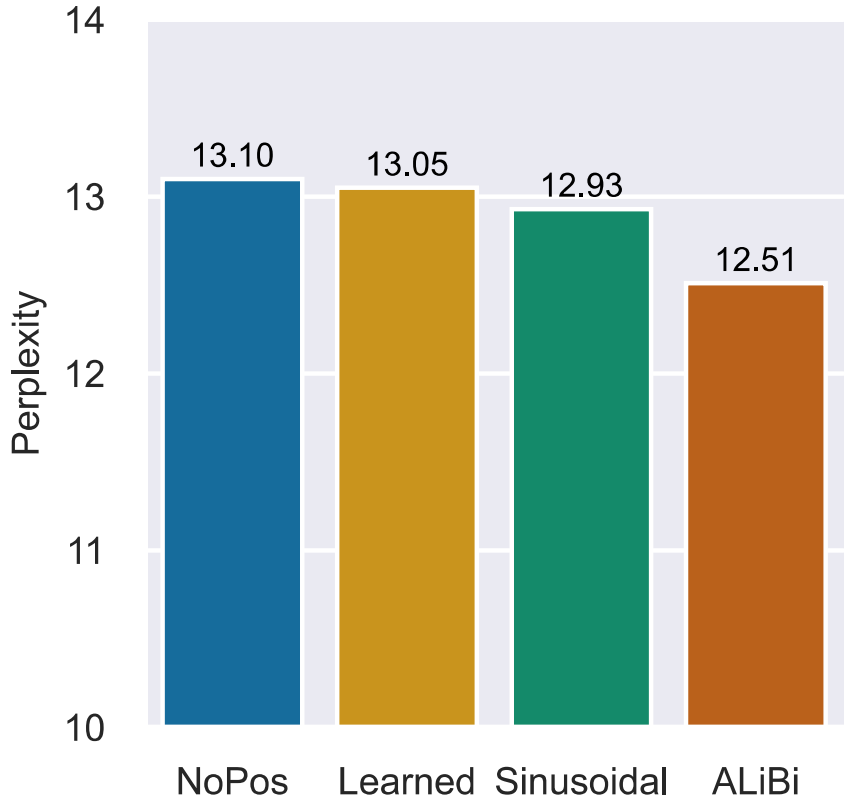
Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

1.3B GPT@ Pile


14
13
12
11
10
Perplexity
13.10
13.05
12.93
12.51
NoPos
Learned
Sinusoidal
ALiBi

1.3B GPT@ Pile



😯 NoPE 👌

1.3B GPT@ wiki103 & Pile

	WikiText-103	The Pile
NoPos	20.97	13.10
Learned	20.42	13.05
Sinusoidal	20.16	12.93
ALiBi	19.71	12.51

1.3B GPT@ wiki103 & Pile

	WikiText-103	The Pile
NoPos	20.97	13.10
Learned	20.42	13.05
Sinusoidal	20.16	12.93
ALiBi	19.71	12.51

2 datasets → NoPE 👌

GPT@ Pile

Model Size	125M	350M	760M	1.3B
NoPos	22.15	16.87	14.29	13.10
Learned	22.04	16.84	14.21	13.05
Sinusoidal	21.49	16.58	14.04	12.93
ALiBi	19.94	15.66	13.53	12.51

Seq Length	256	512	1024	2048
NoPos	14.98	13.82	13.10	12.87
Learned	14.94	13.77	13.05	12.72
Sinusoidal	14.84	13.66	12.93	12.62
ALiBi	14.65	13.37	12.51	12.06

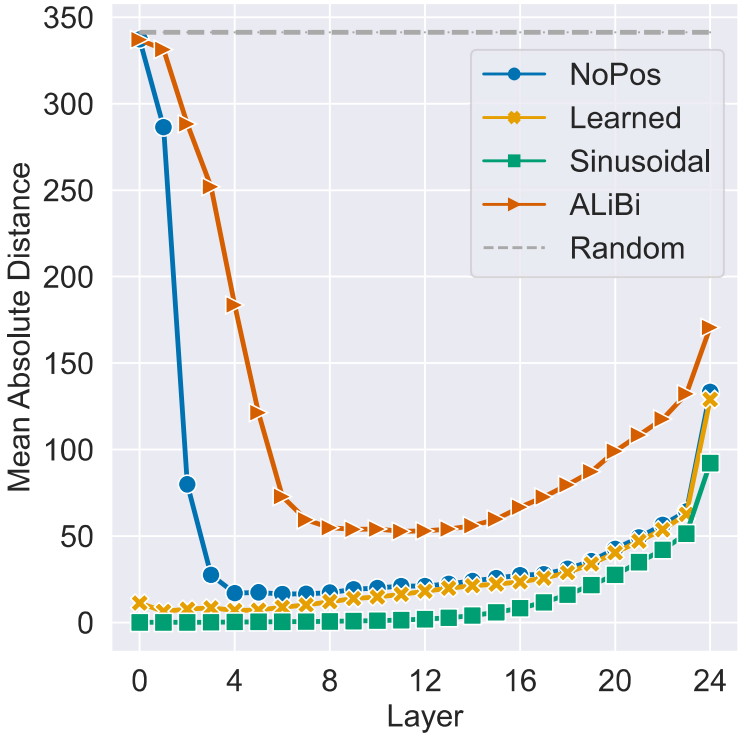
GPT@ Pile

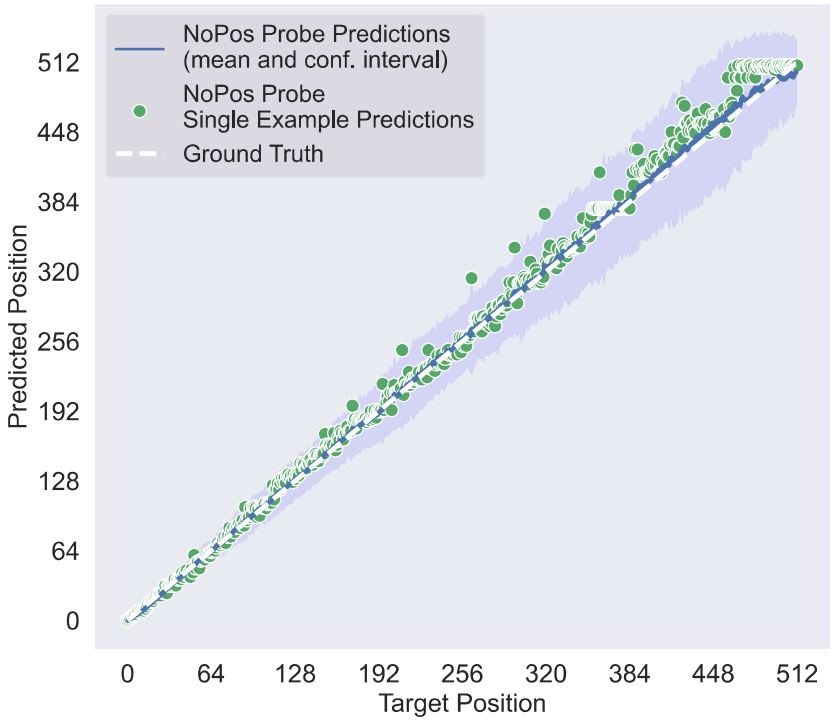
Model Size	125M	350M	760M	1.3B
NoPos	22.15	16.87	14.29	13.10
Learned	22.04	16.84	14.21	13.05
Sinusoidal	21.49	16.58	14.04	12.93
ALiBi	19.94	15.66	13.53	12.51

Seq Length	256	512	1024	2048
NoPos	14.98	13.82	13.10	12.87
Learned	14.94	13.77	13.05	12.72
Sinusoidal	14.84	13.66	12.93	12.62
ALiBi	14.65	13.37	12.51	12.06

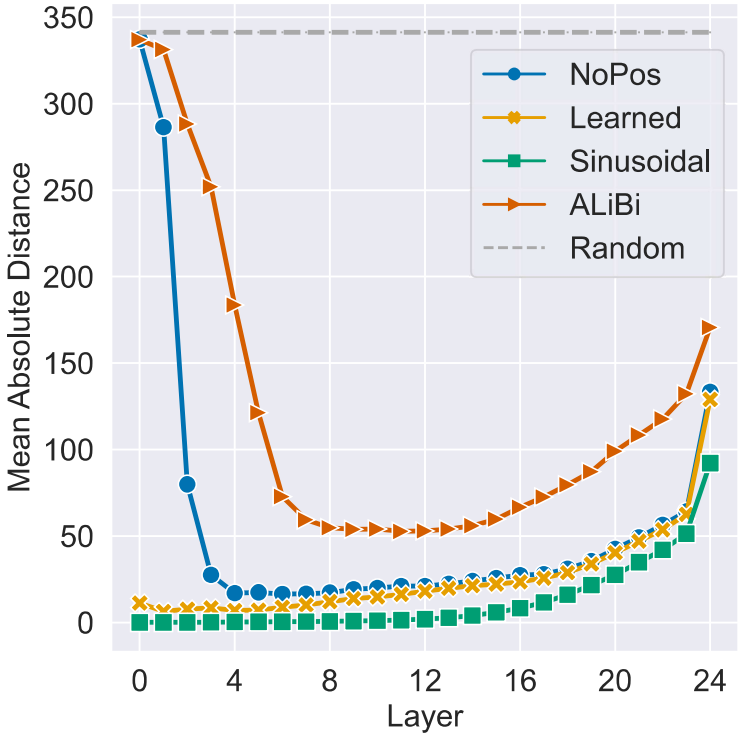
Various size👈🏻 & 👉🏻length → NoPE 👌

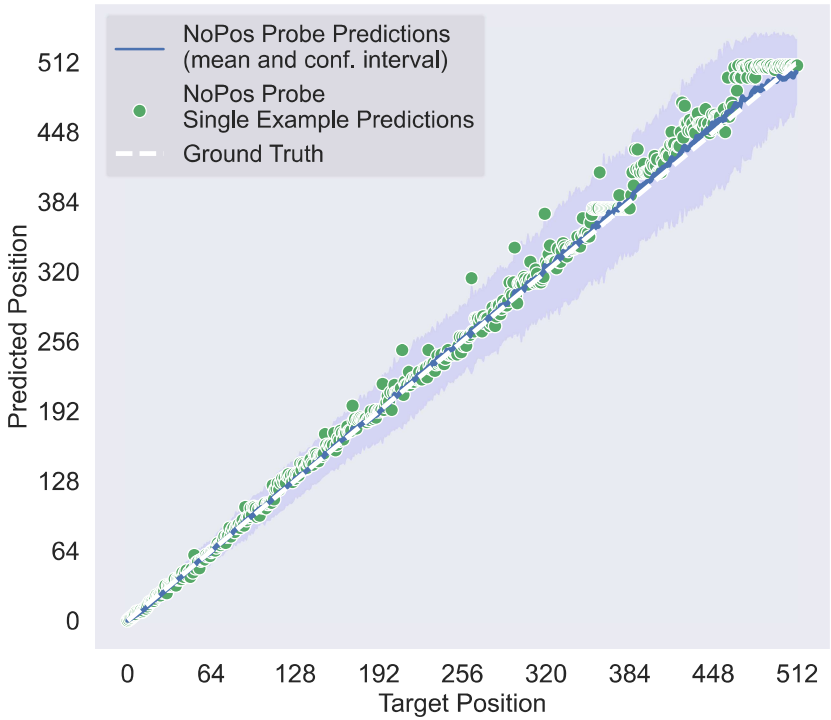
POS Probing


Mean Absolute Distance
Layer
NoPos
Learned
Sinusoidal
ALiBi
Random
0
50
100
150
200
250
300
350
0
4
8
12
16
20
24


Predicted Position
Target Position
NoPos Probe Predictions
(mean and conf. interval)
NoPos Probe
Single Example Predictions
Ground Truth
0
64
128
192
256
320
384
448
512

POS Probing





NoPos: POS info in hidden too

How about BERT-style

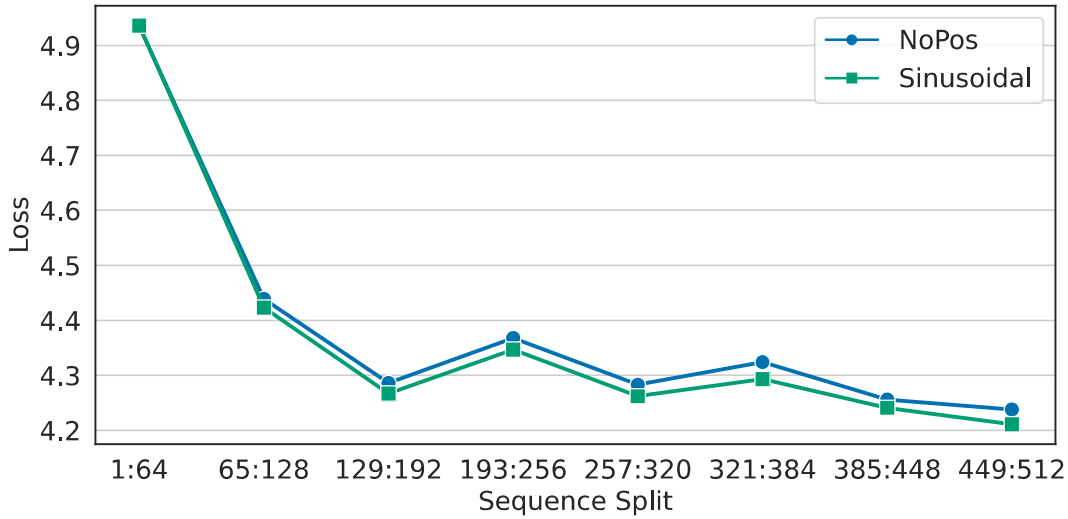
	MLM Perplexity
NoPos	147.18
Learned	4.06
Sinusoidal	4.07
ALiBi	4.00

How about BERT-style

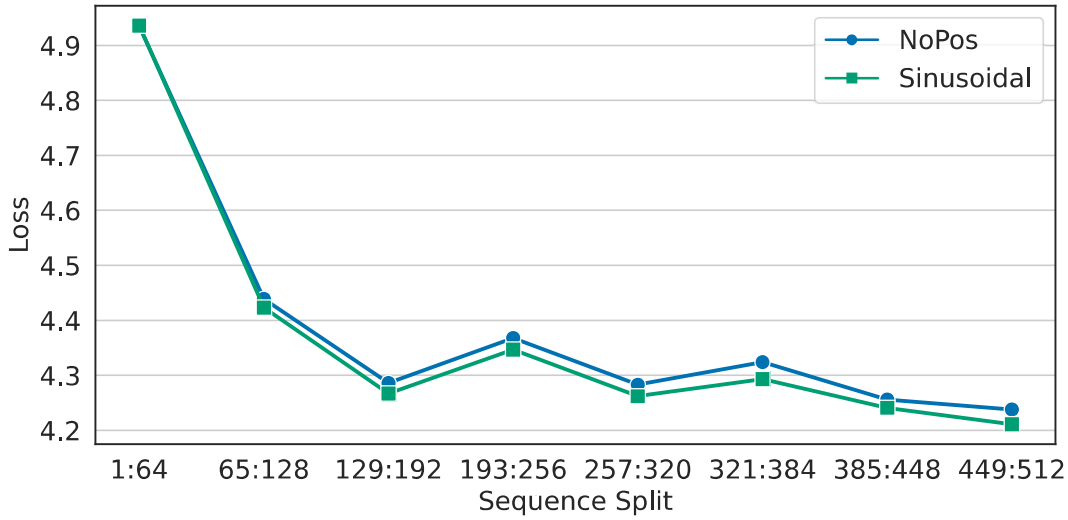
	MLM Perplexity
NoPos	147.18
Learned	4.06
Sinusoidal	4.07
ALiBi	4.00

GPT-style ✅ BERT-style ❎ T5-style ❓ (❎)

Sequence split

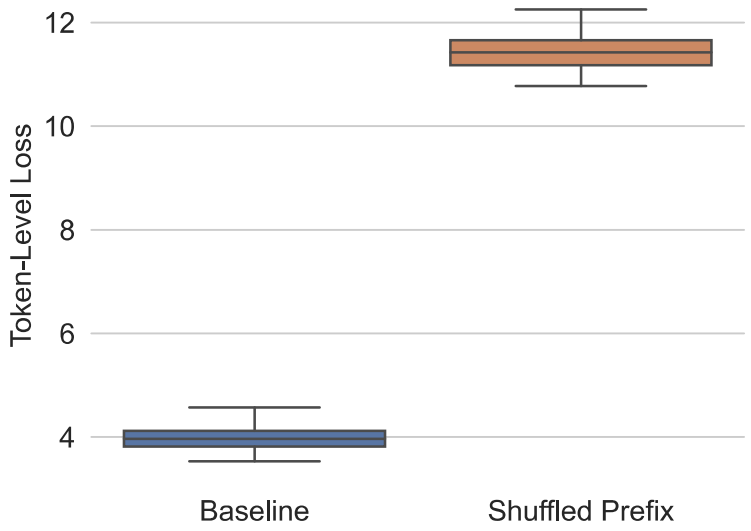


Sequence split



Slightly Worse

Shuffled Prefix



B Word Order Analysis

Is positional information necessary for language modeling, or does the order of the input tokens not matter? To answer this, we conduct the following experiment: instead of computing the loss on the complete sequence, we pick a specific token in the sequence. The next token prediction is conditioned on the previous tokens in the sequence, and so we shuffle the order of the tokens in the prefix and compute the loss only for that specific token. We repeat the experiment with the original, un-shuffled prefix sequence as the baseline and compare the results.

ACL2022 Findings

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

ACL2022 Findings

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

😲 NoPE 👌 But why? 🤔

ACL2023 Honorable Mentions

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

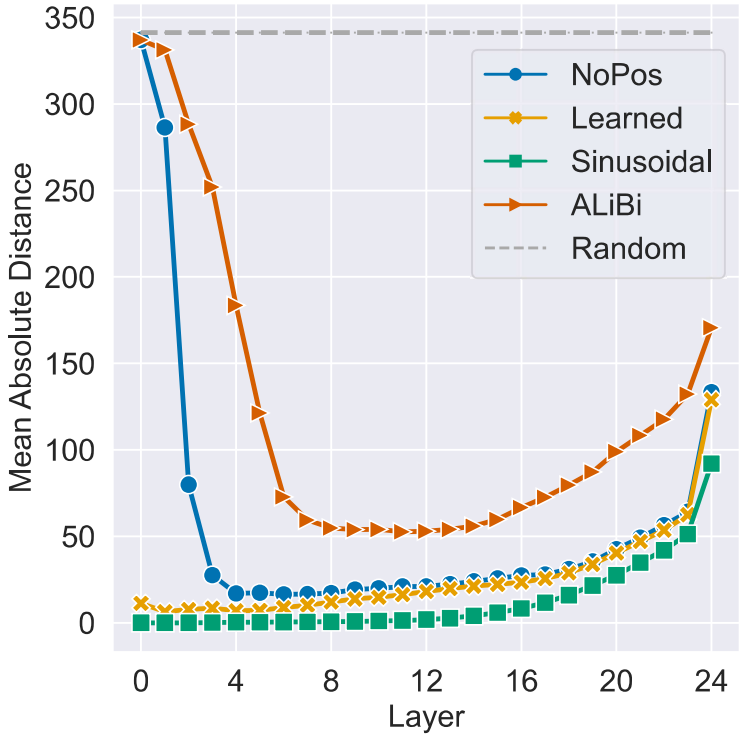
Ting-Han Fan
Princeton University

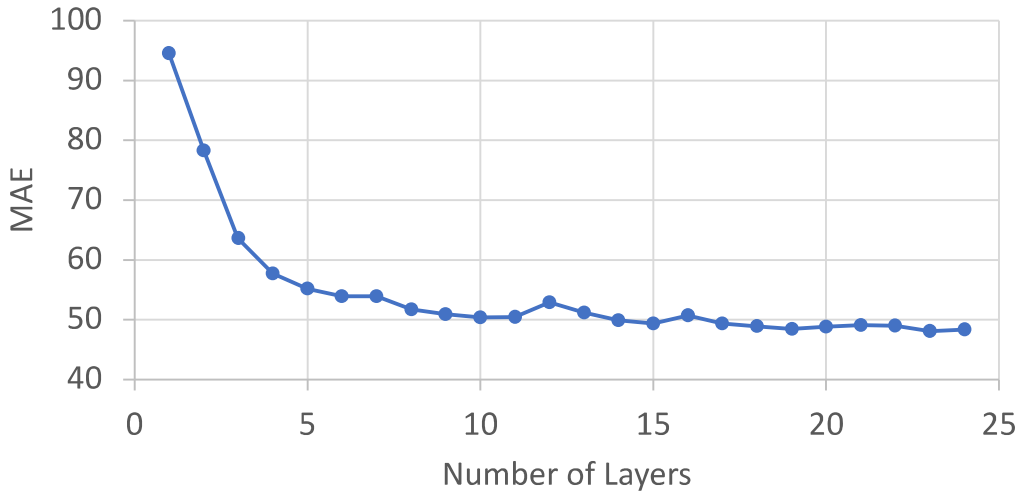
Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

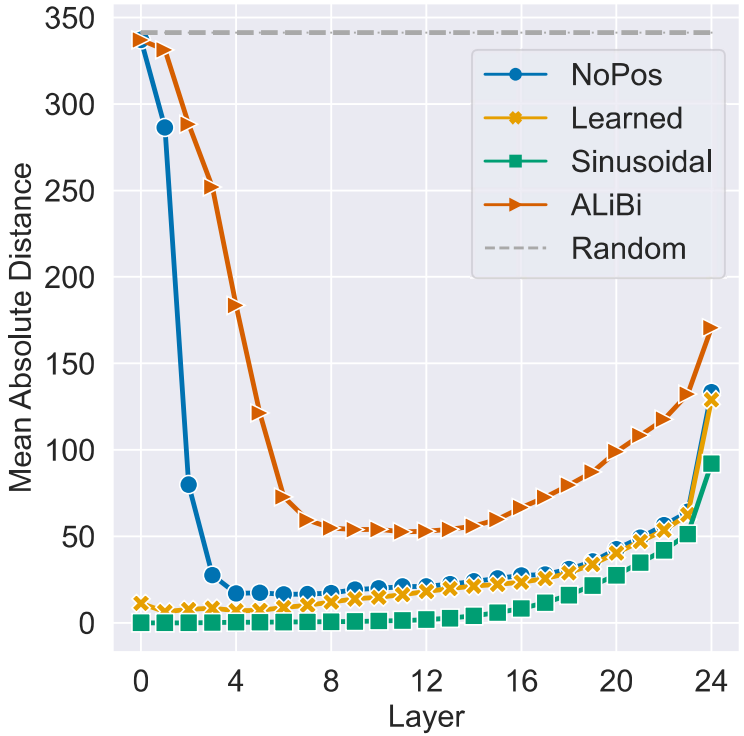
Peter J. Ramadge
Princeton University

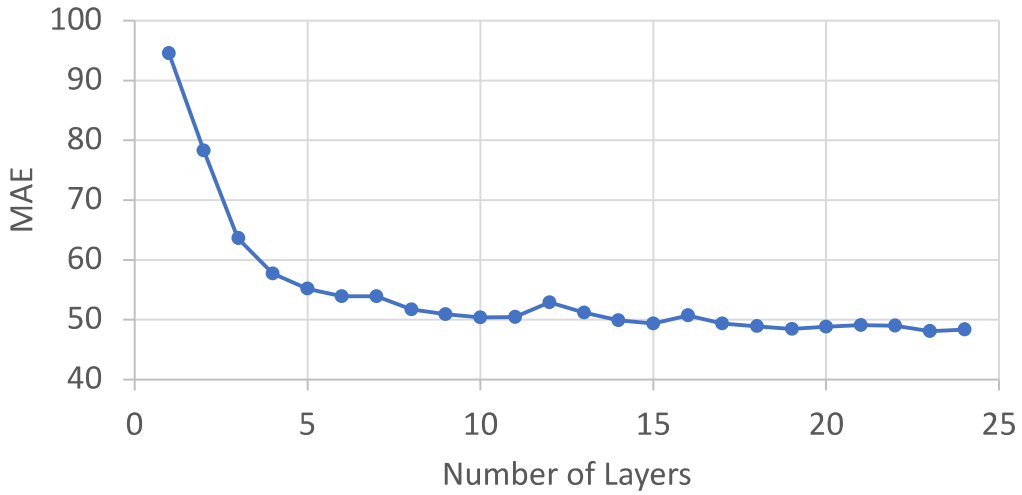
Probing


Mean Absolute Distance
350
300
250
200
150
100
50
0
0
4
8
12
16
20
24
Layer
NoPos
Learned
Sinusoidal
ALiBi
Random


MAE
100
90
80
70
60
50
40
0
5
10
15
20
25
Number of Layers

Probing





NoPos: POS info in hidden too

Proof

🎯 vector variance ↔️ position

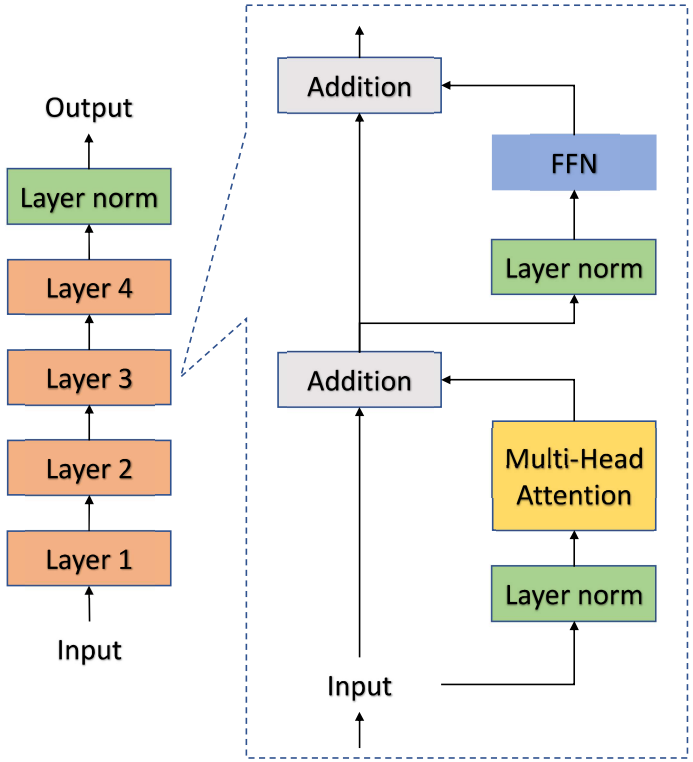
Proof

🎯 vector variance ↔️ position

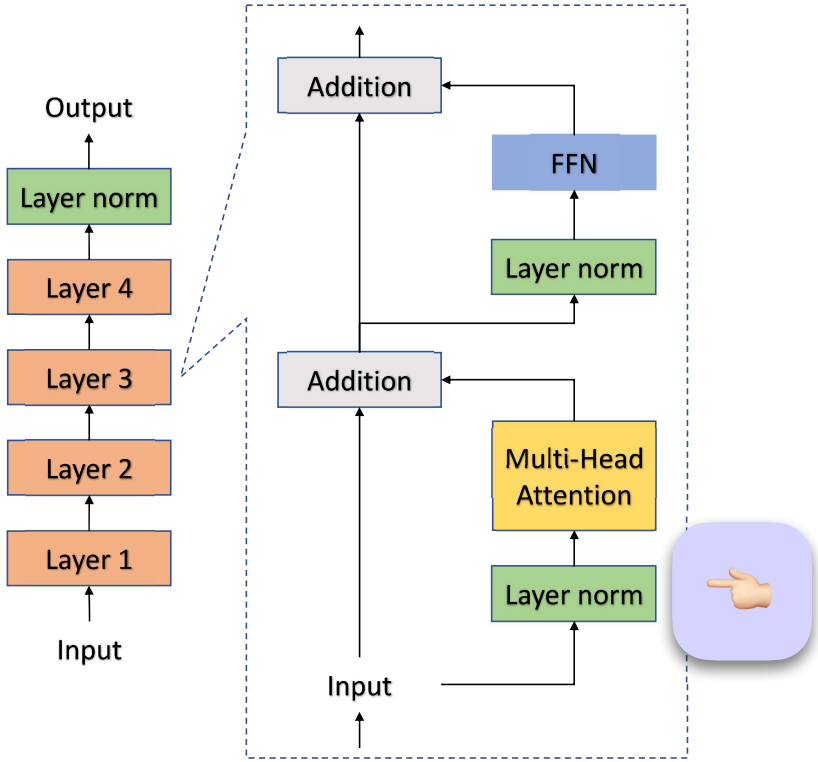
are: hidden dimension $d = 768$, number of attention heads $H = 12$, head dimension $d/H = 64$, sequence length $L = 512$, standard deviation for initialization $\sigma = 0.02$. All proofs of lemmas are deferred to Appendix A.

Given a sequence of randomly sampled input embeddings $\{\boldsymbol{x}_m\}_{m=1}^{L}$, where each element of $\boldsymbol{x}_m \in \mathbb{R}^d$ is sampled i.i.d from $N(0, \sigma^2)$, a TLM

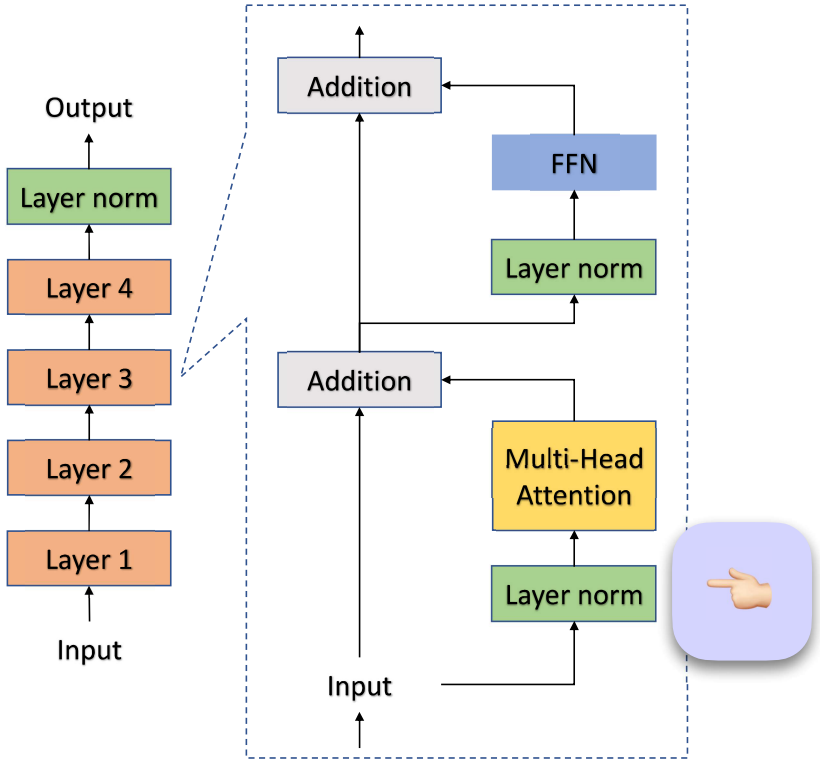
Layer Norm


Output
Layer norm
Layer 4
Layer 3
Layer 2
Layer 1
Input
Addition
FFN
Layer norm
Addition
Multi-Head Attention
Layer norm
Input

Layer Norm


Output
Layer norm
Layer 4
Layer 3
Layer 2
Layer 1
Input
Addition
FFN
Layer norm
Addition
Multi-Head Attention
Layer norm
Input

Layer Norm

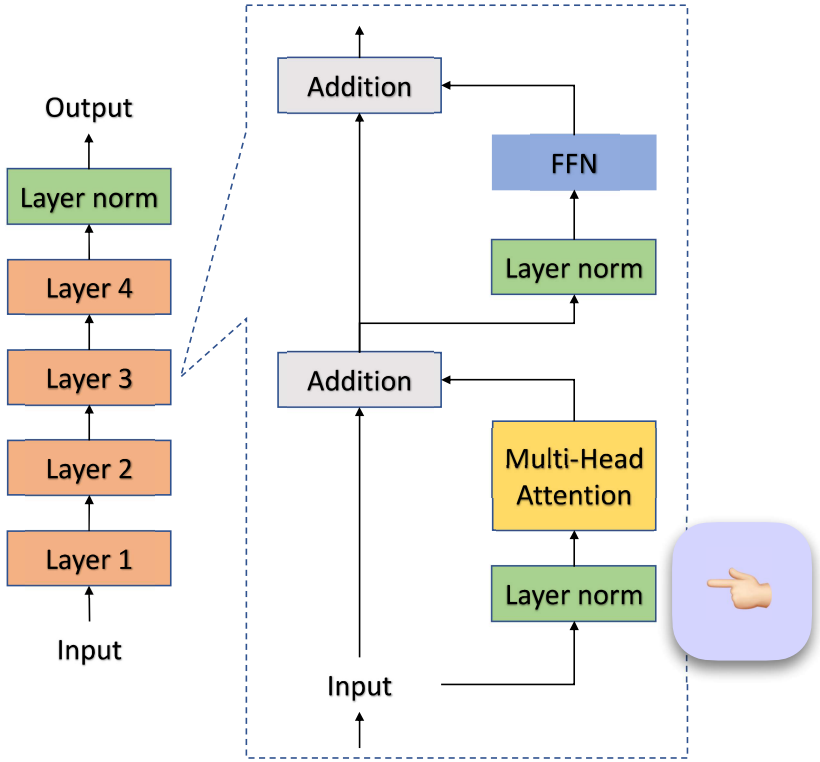


$$\overline{\boldsymbol{x}}_{m,:} = \frac{\sum_{i=1}^{d} \boldsymbol{x}_{mi}}{d},\ S(\boldsymbol{x}_{m,:}) = \frac{\sum_{i=1}^{d} (\boldsymbol{x}_{mi} - \overline{\boldsymbol{x}}_{m,:})^2}{d}$$

$$\boldsymbol{e}_{mi} = \frac{\boldsymbol{x}_{mi} - \overline{\boldsymbol{x}}_{m,:}}{\sqrt{S(\boldsymbol{x}_{m,:})}} * \gamma + \beta$$

$$\overset{(*)}{\approx} \frac{\boldsymbol{x}_{mi} - \mathbb{E}[\boldsymbol{x}_{mi}]}{\sqrt{\mathbb{V}[\boldsymbol{x}_{mi}]}} \sim N(0,1),$$

Layer Norm



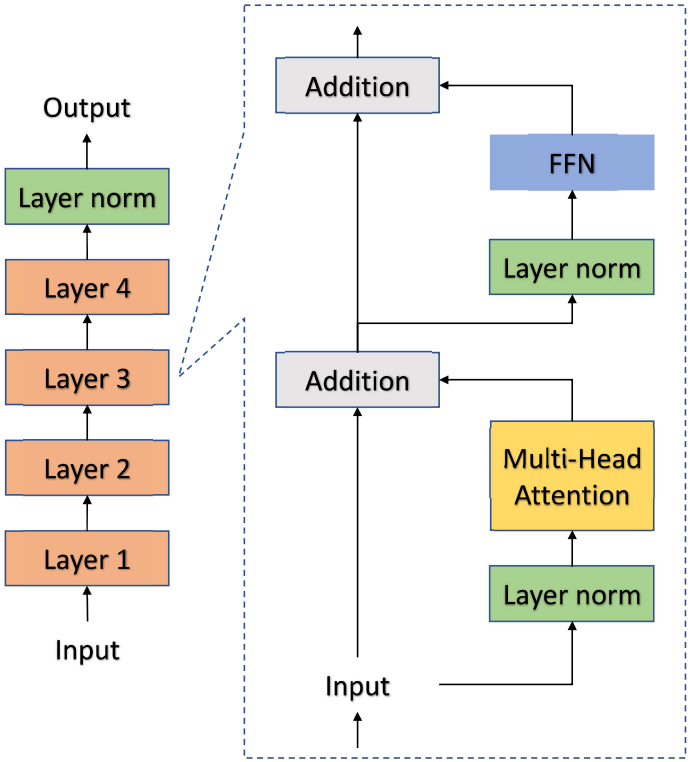
$$\overline{\boldsymbol{x}}_{m,:} = \frac{\sum_{i=1}^{d} \boldsymbol{x}_{mi}}{d},\ S(\boldsymbol{x}_{m,:}) = \frac{\sum_{i=1}^{d} (\boldsymbol{x}_{mi} - \overline{\boldsymbol{x}}_{m,:})^2}{d}$$

$$\boldsymbol{e}_{mi} = \frac{\boldsymbol{x}_{mi} - \overline{\boldsymbol{x}}_{m,:}}{\sqrt{S(\boldsymbol{x}_{m,:})}} * \gamma + \beta$$

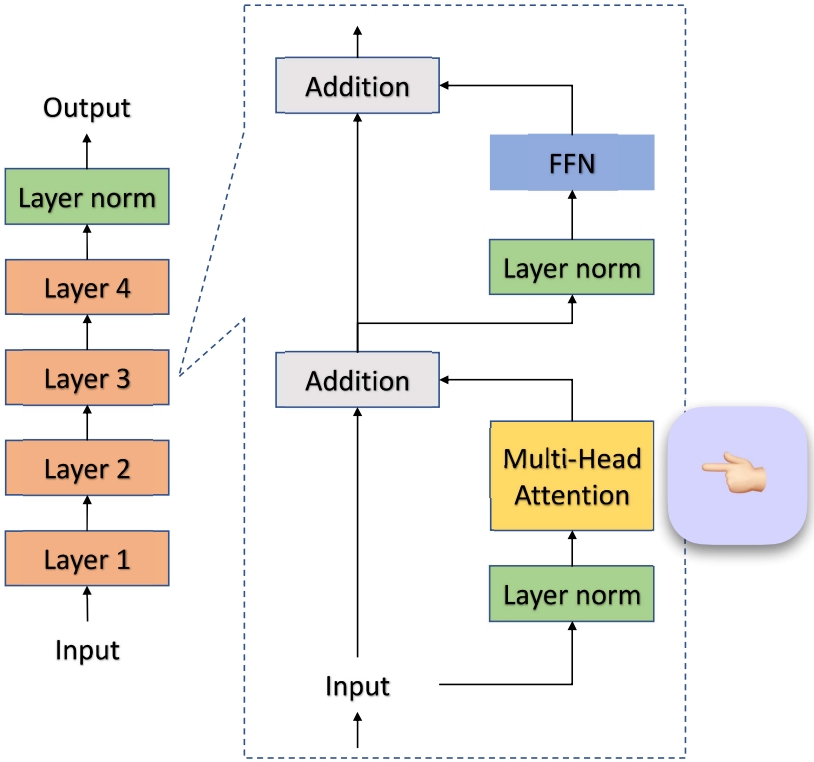
$$\overset{(*)}{\approx} \frac{\boldsymbol{x}_{mi} - \mathbb{E}[\boldsymbol{x}_{mi}]}{\sqrt{\mathbb{V}[\boldsymbol{x}_{mi}]}} \sim N(0,1),$$

$\gamma = 1$ and $\beta = 0$

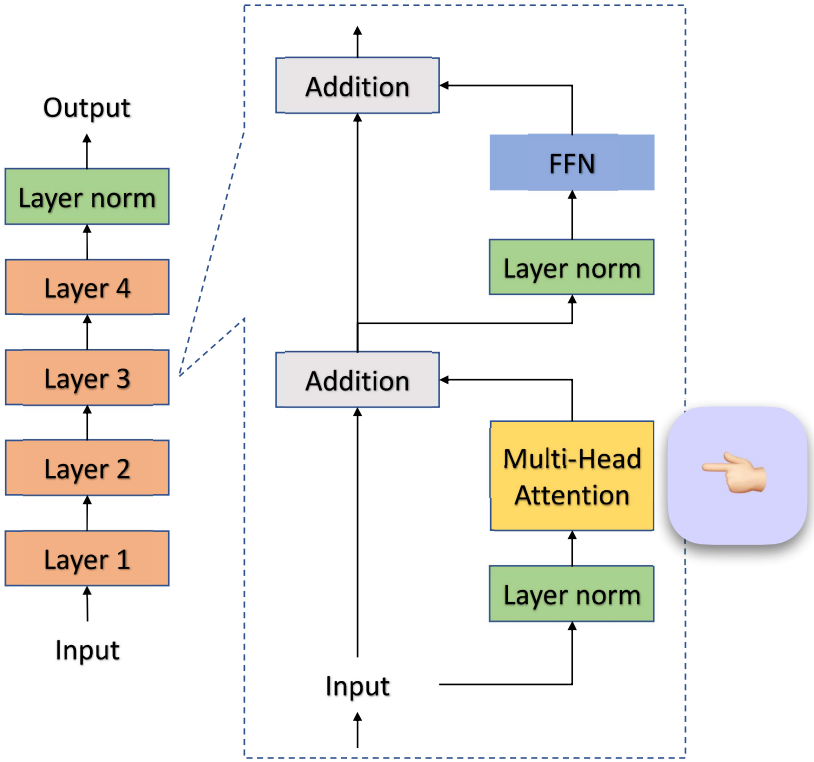
MHA


Output
Layer norm
Layer 4
Layer 3
Layer 2
Layer 1
Input
Addition
FFN
Layer norm
Addition
Multi-Head Attention
Layer norm
Input

MHA

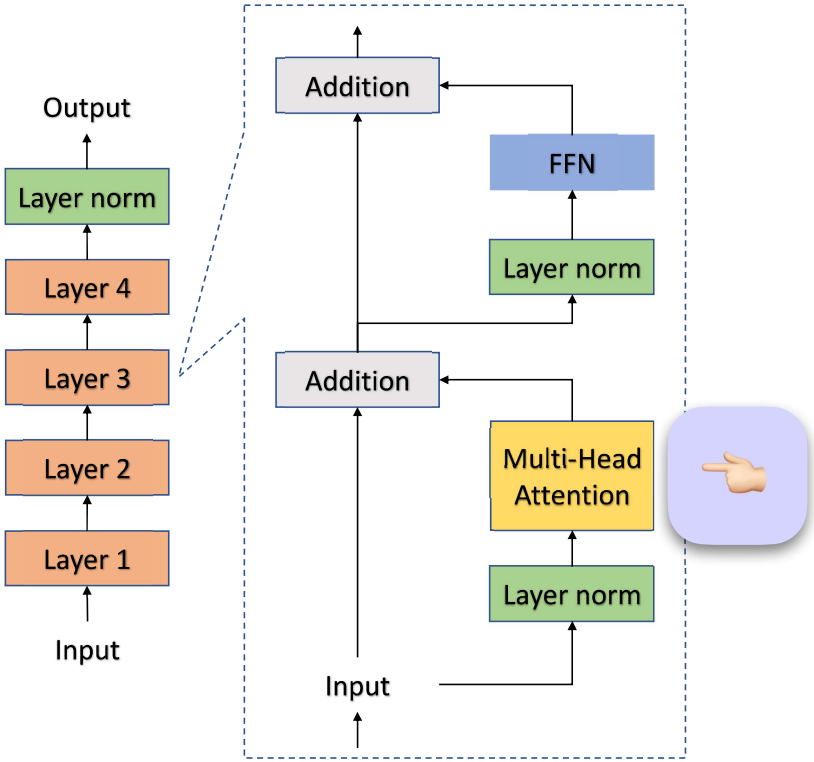

Output
Layer norm
Layer 4
Layer 3
Layer 2
Layer 1
Input
Addition
FFN
Layer norm
Addition
Multi-Head Attention
Layer norm
Input

MHA



Lemma 1. *$\boldsymbol{q}_m$, $\boldsymbol{k}_m$, and $\boldsymbol{v}_m$ have zero mean and $(d\sigma^2) \cdot I$ covariance matrix.*

MHA

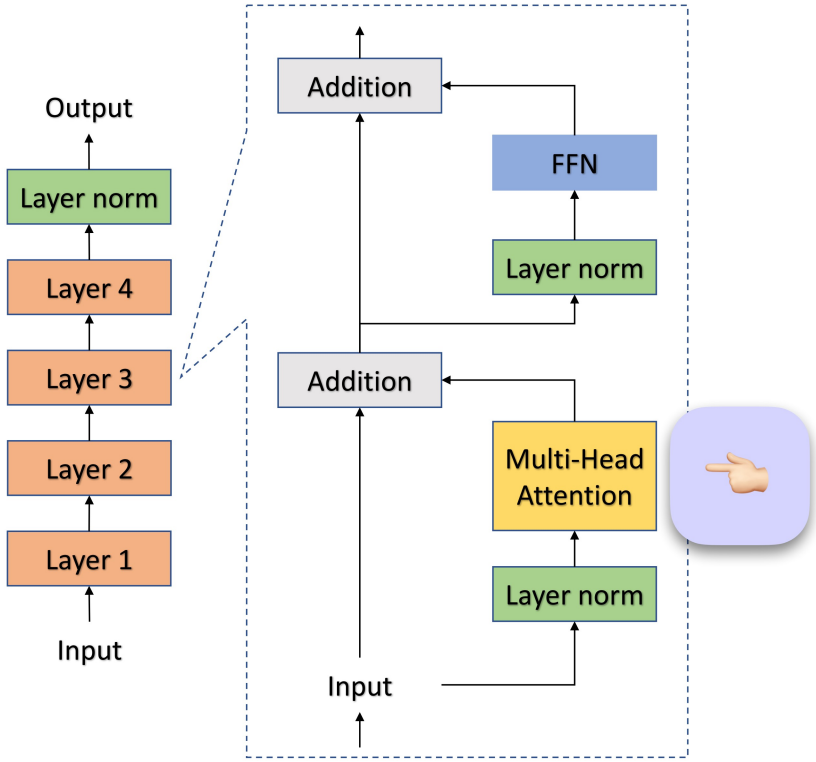


Lemma 1. *$\boldsymbol{q}_m$, $\boldsymbol{k}_m$, and $\boldsymbol{v}_m$ have zero mean and $(d\sigma^2) \cdot I$ covariance matrix.*

$$\boldsymbol{q}_m = \boldsymbol{W}_q \boldsymbol{e}_m, \quad \boldsymbol{k}_m = \boldsymbol{W}_k \boldsymbol{e}_m, \quad \boldsymbol{v}_m = \boldsymbol{W}_v \boldsymbol{e}_m,$$

$$\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{\frac{d}{H} \times d} \quad \text{sampled i.i.d from } N(0, \sigma^2)$$

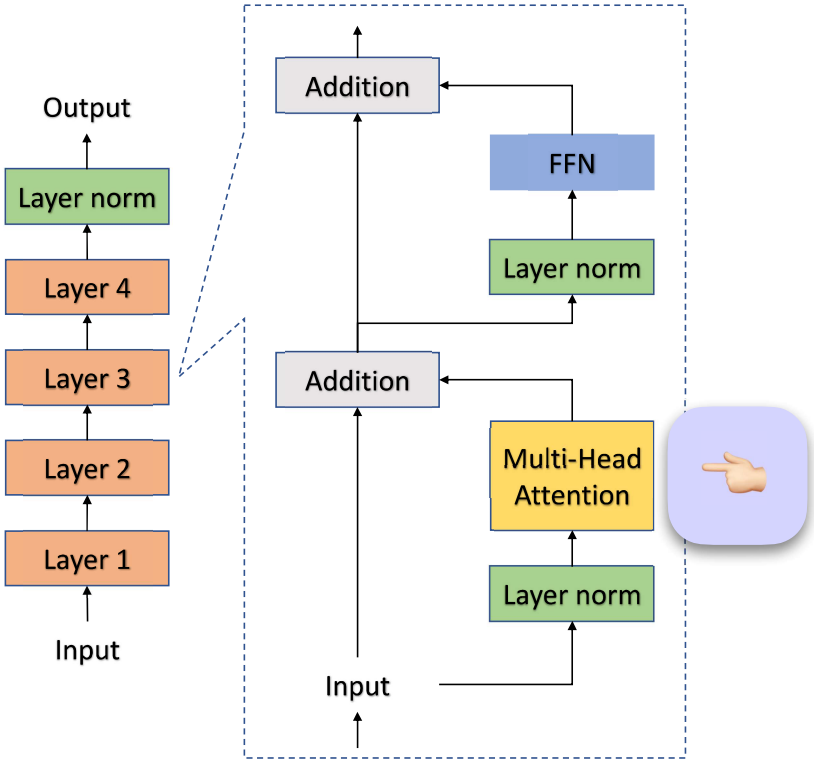
MHA: Proof



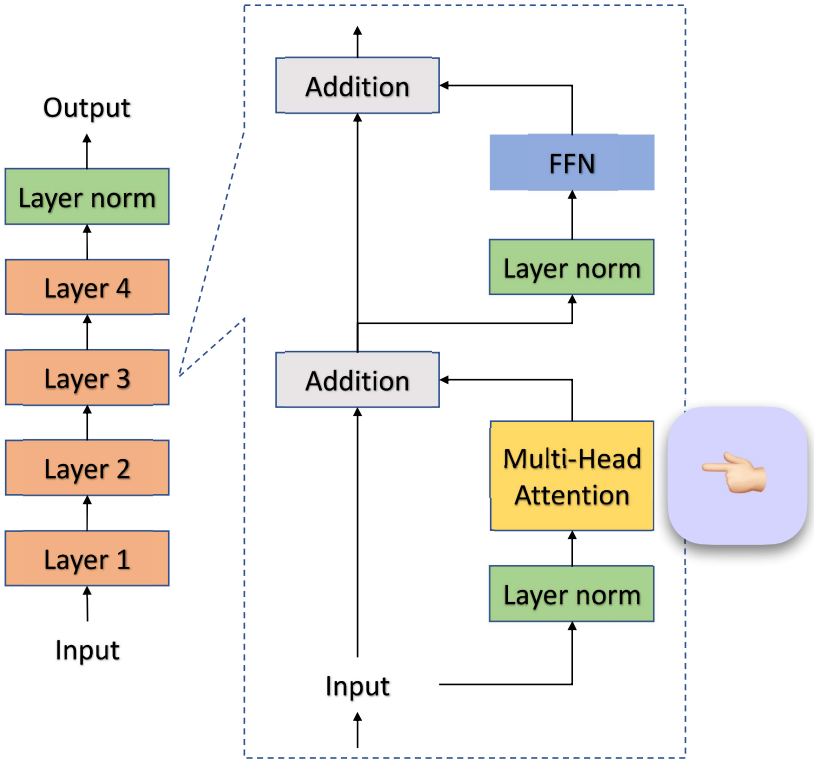
Lemma 1. *$\boldsymbol{q}_m$, $\boldsymbol{k}_m$, and $\boldsymbol{v}_m$ have zero mean and $(d\sigma^2) \cdot I$ covariance matrix.*

$$
\begin{aligned}
\operatorname{cov}(\boldsymbol{v}_m, \boldsymbol{v}_n) &= \mathbb{E}[\boldsymbol{v}_m \boldsymbol{v}_n^\top] \\
&= \mathbb{E}[\boldsymbol{W}_v \boldsymbol{e}_m \boldsymbol{e}_n^\top \boldsymbol{W}_v^\top] \\
&= \mathbb{E}\left[\begin{bmatrix} \boldsymbol{r}_1^\top \boldsymbol{e}_m \\ \vdots \\ \boldsymbol{r}_{\frac{d}{H}}^\top \boldsymbol{e}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_n^\top \boldsymbol{r}_1 & \dots & \boldsymbol{e}_n^\top \boldsymbol{r}_{\frac{d}{H}} \end{bmatrix}\right] \\
&= \left[\mathbb{E}[\boldsymbol{r}_i^\top \boldsymbol{e}_m \boldsymbol{e}_n^\top \boldsymbol{r}_j]\right]_{i,j=1}^{\frac{d}{H}} \\
&= \left[\mathbb{E}[\operatorname{Tr}(\boldsymbol{r}_j \boldsymbol{r}_i^\top \boldsymbol{e}_m \boldsymbol{e}_n^\top)]\right]_{i,j=1}^{\frac{d}{H}} \\
&= \left[\operatorname{Tr}(\mathbb{E}[\boldsymbol{r}_j \boldsymbol{r}_i^\top]\mathbb{E}[\boldsymbol{e}_m \boldsymbol{e}_n^\top])\right]_{i,j=1}^{\frac{d}{H}} \\
&\overset{(*)}{=} \left[\operatorname{Tr}((\mathbb{1}_{i=j}\sigma^2) \cdot I_d \cdot \mathbb{1}_{m=n} \cdot I_d)\right]_{i,j=1}^{\frac{d}{H}} \\
&= \left[\mathbb{1}_{i=j}\mathbb{1}_{m=n} d\sigma^2\right]_{i,j=1}^{\frac{d}{H}} \\
&= (\mathbb{1}_{m=n} d\sigma^2) \cdot I_{d/H}
\end{aligned}
$$

MHA

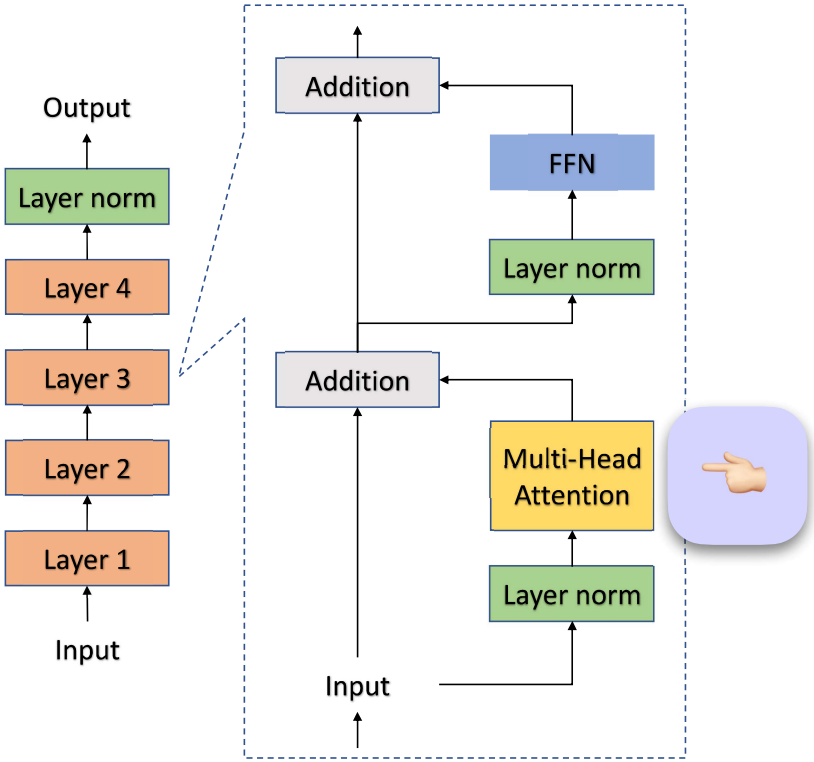

Output
Layer norm
Layer 4
Layer 3
Layer 2
Layer 1
Input
Addition
FFN
Layer norm
Addition
Multi-Head Attention
Layer norm
Input

MHA



Lemma 2. *l_{mn} has zero mean and $\frac{d^3\sigma^4}{H^2}$ variance. $l_{mn}/\sqrt{d/H}$ has $\frac{d^2\sigma^4}{H}$ variance.*

MHA

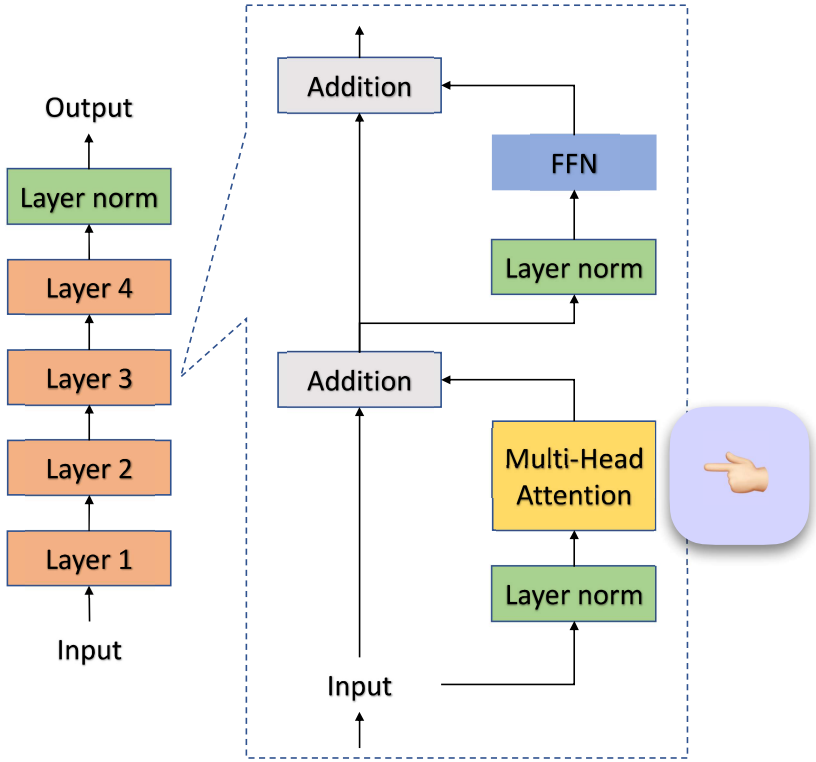


Lemma 2. *l_{mn} has zero mean and $\frac{d^3\sigma^4}{H^2}$ variance. $l_{mn}/\sqrt{d/H}$ has $\frac{d^2\sigma^4}{H}$ variance.*

$$l_{mn} = \begin{cases} \langle \boldsymbol{q}_m, \boldsymbol{k}_n \rangle, & \text{if } m \geq n \\ -\inf, & \text{otherwise} \end{cases}$$

$$a_{mn} = \frac{\exp\left(l_{mn}/\sqrt{d/H}\right)}{\sum_{i=1}^{L} \exp\left(l_{mi}/\sqrt{d/H}\right)}$$

MHA



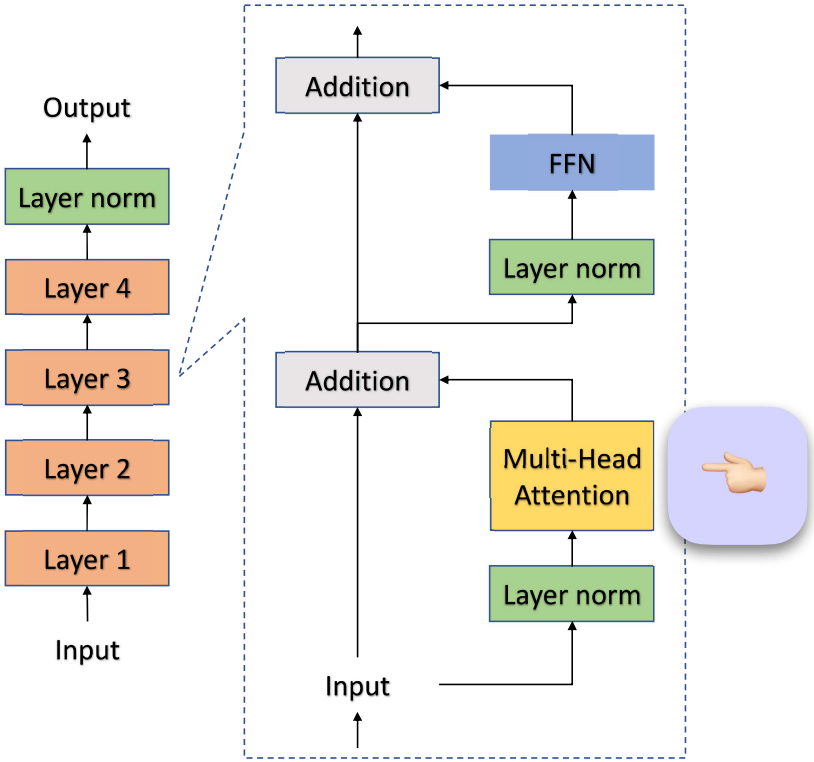
Lemma 2. *l_{mn} has zero mean and $\frac{d^3\sigma^4}{H^2}$ variance. $l_{mn}/\sqrt{d/H}$ has $\frac{d^2\sigma^4}{H}$ variance.*

$$l_{mn} = \begin{cases} \langle \boldsymbol{q}_m, \boldsymbol{k}_n \rangle, & \text{if } m \geq n \\ -\inf, & \text{otherwise} \end{cases}$$

$$a_{mn} = \frac{\exp\left(l_{mn}/\sqrt{d/H}\right)}{\sum_{i=1}^{L} \exp\left(l_{mi}/\sqrt{d/H}\right)}$$

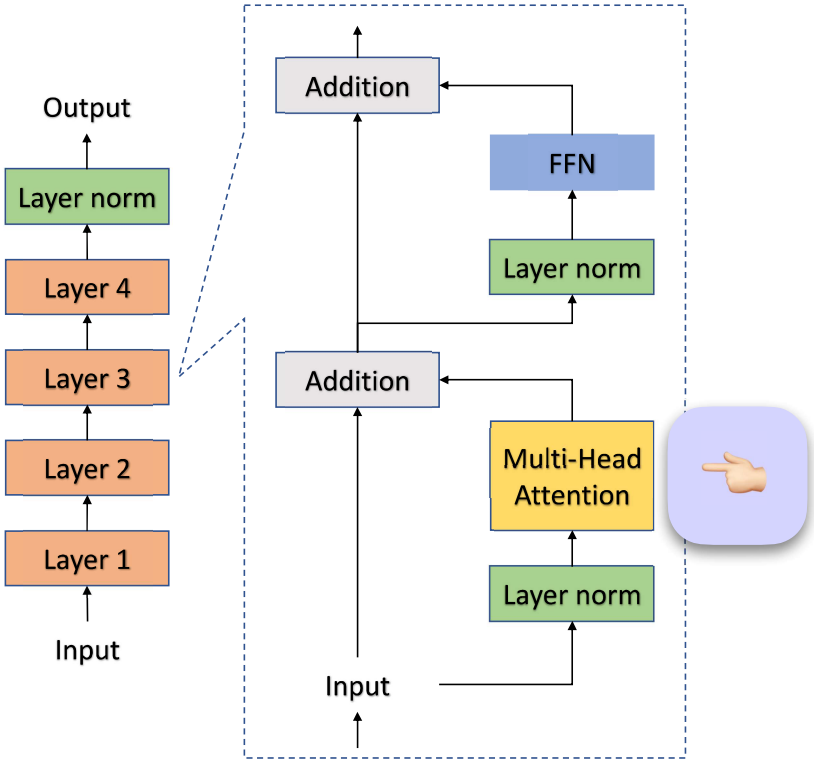
Property 1. *When $\sigma^4 \ll \frac{H}{d^2}$, $l_{m,:}$ has small variance, making the attention weights $a_{m,:}$ almost evenly distributed among all positions.*[2]

MHA



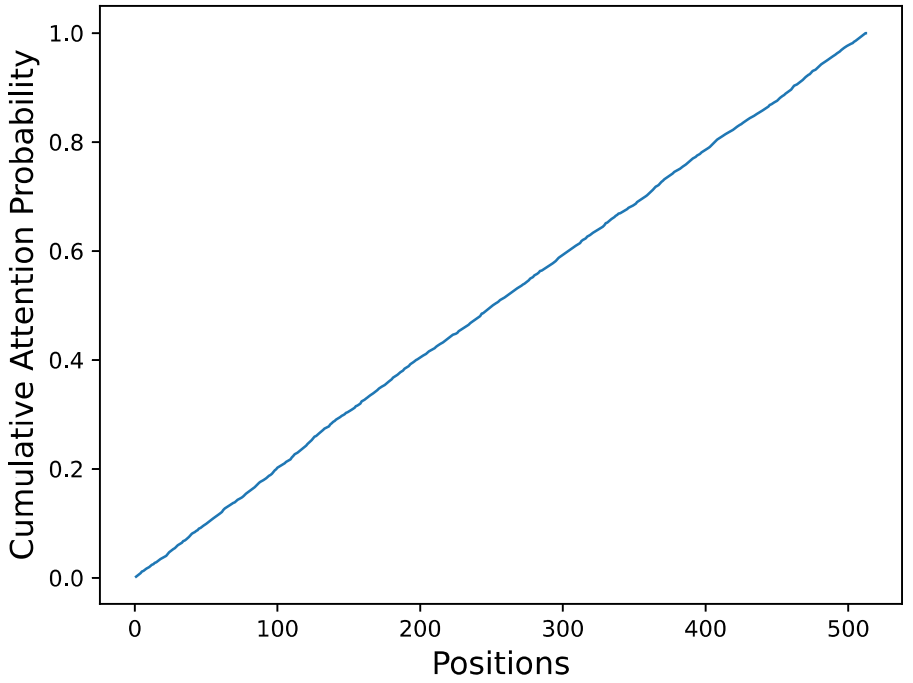
Property 1. *When* $\sigma^4 \ll \frac{H}{d^2}$, $l_{m,:}$ *has small variance, making the attention weights* $a_{m,:}$ *almost evenly distributed among all positions.*[2]

MHA

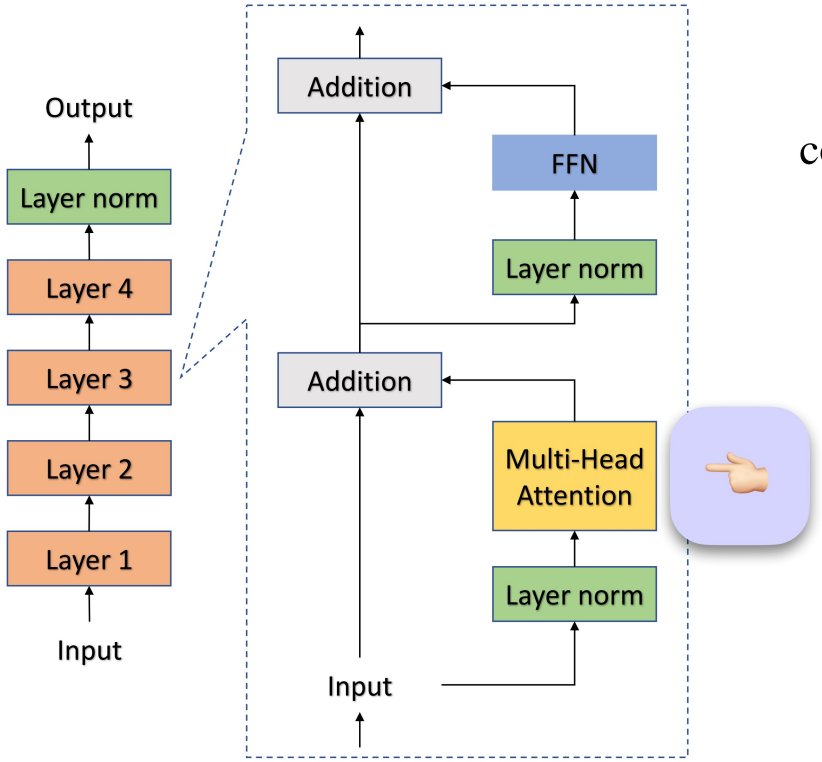


Property 1. *When* $\sigma^4 \ll \frac{H}{d^2}$, $l_{m,:}$ *has small variance, making the attention weights* $a_{m,:}$ *almost evenly distributed among all positions.*[2]

$$\frac{768^2 \cdot 0.02^4}{12} \approx 0.0079$$



MHA: Proof

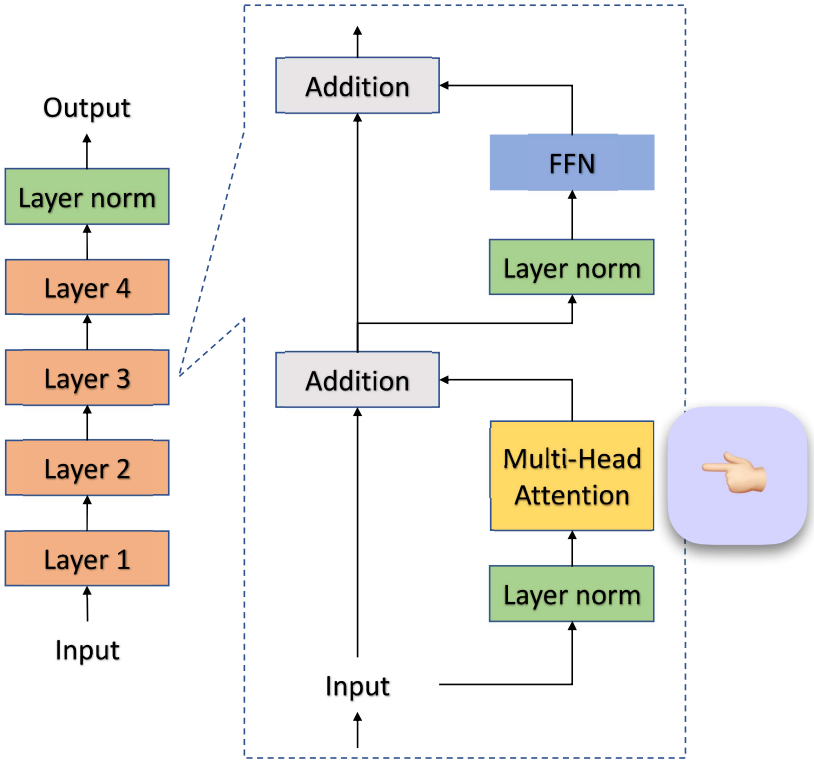


Lemma 2. *l_{mn} has zero mean and $\frac{d^3\sigma^4}{H^2}$ variance. $l_{mn}/\sqrt{d/H}$ has $\frac{d^2\sigma^4}{H}$ variance.*

$$
\begin{aligned}
&\mathrm{cov}(l_{mn}, l_{mp}) \\
&= \mathbb{E}[(\boldsymbol{e}_m^\top \boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{e}_n)(\boldsymbol{e}_m^\top \boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{e}_p)^\top] \\
&= \mathbb{E}[\mathrm{Tr}(\boldsymbol{e}_m^\top \boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{e}_n \boldsymbol{e}_p^\top \boldsymbol{W}_k^\top \boldsymbol{W}_q \boldsymbol{e}_m)] \\
&= \mathbb{E}[\mathrm{Tr}(\boldsymbol{e}_m \boldsymbol{e}_m^\top \boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{e}_n \boldsymbol{e}_p^\top \boldsymbol{W}_k^\top \boldsymbol{W}_q)] \\
&= \mathrm{Tr}(\mathbb{E}[\boldsymbol{e}_m \boldsymbol{e}_m^\top] \mathbb{E}[\boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{e}_n \boldsymbol{e}_p^\top \boldsymbol{W}_k^\top \boldsymbol{W}_q]) \\
&= \mathbb{E}[\mathrm{Tr}(\boldsymbol{e}_n \boldsymbol{e}_p^\top \boldsymbol{W}_k^\top \boldsymbol{W}_q \boldsymbol{W}_q^\top \boldsymbol{W}_k)] \\
&= \mathrm{Tr}(\mathbb{E}[\boldsymbol{e}_n \boldsymbol{e}_p^\top] \mathbb{E}[\boldsymbol{W}_k^\top \boldsymbol{W}_q \boldsymbol{W}_q^\top \boldsymbol{W}_k])] \\
&= (\mathbb{1}_{n=p}) \mathrm{Tr}(\mathbb{E}[\boldsymbol{W}_q \boldsymbol{W}_q^\top] \mathbb{E}[\boldsymbol{W}_k \boldsymbol{W}_k^\top]) \\
&\overset{(*)}{=} (\mathbb{1}_{n=p}) \mathrm{Tr}((\frac{d}{H}\sigma^2 \cdot I)(\frac{d}{H}\sigma^2 \cdot I)) \\
&= (\mathbb{1}_{n=p}) \frac{d^3\sigma^4}{H^2}
\end{aligned}
$$

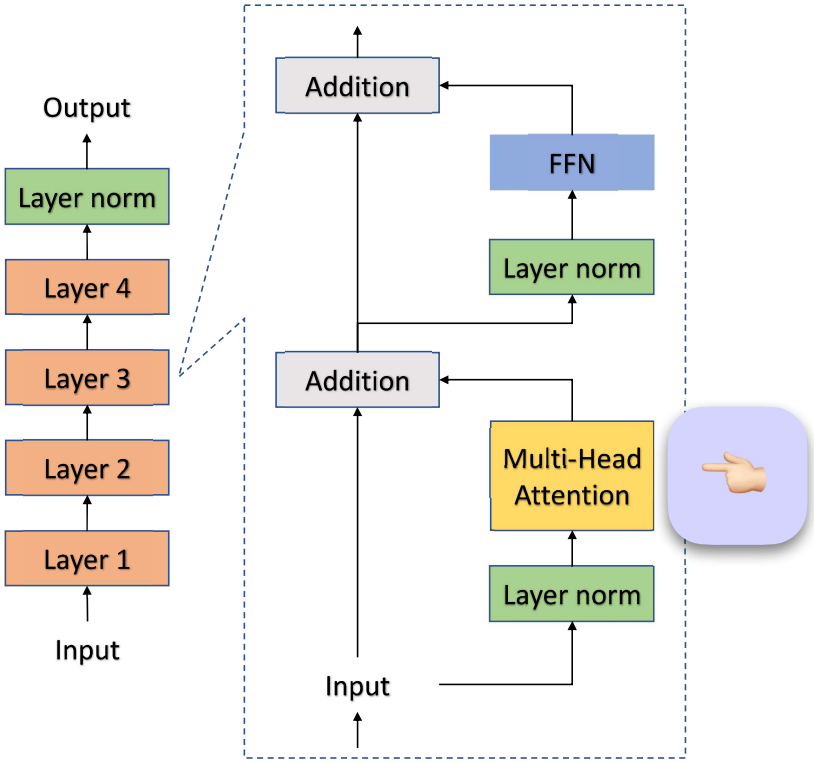
$$
\begin{aligned}
\mathbb{E}[\boldsymbol{W}_q \boldsymbol{W}_q^\top] &= \mathbb{E}\left[\begin{bmatrix} \boldsymbol{r}_1^\top \\ \vdots \\ \boldsymbol{r}_{\frac{d}{H}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_1 & \cdots & \boldsymbol{r}_{\frac{d}{H}} \end{bmatrix}\right] \\
&= \left[\mathbb{E}[\boldsymbol{r}_i^\top \boldsymbol{r}_j]\right]_{i,j=1}^{\frac{d}{H}} = \frac{d}{H}\sigma^2 \cdot I
\end{aligned}
$$

MHA



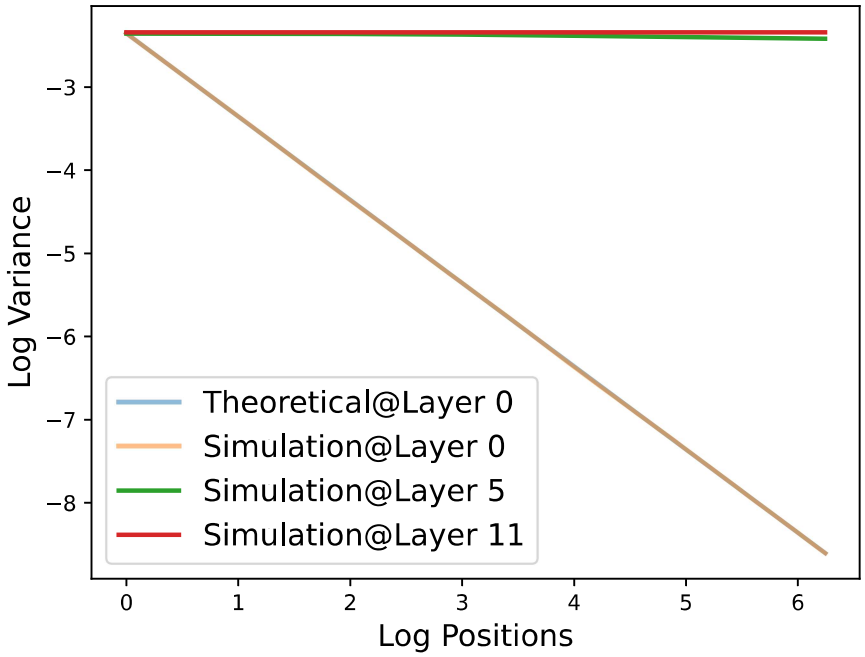
Lemma 3. *o_m has zero mean and $\frac{d^2\sigma^4}{m}I$ covariance matrix.*

MHA

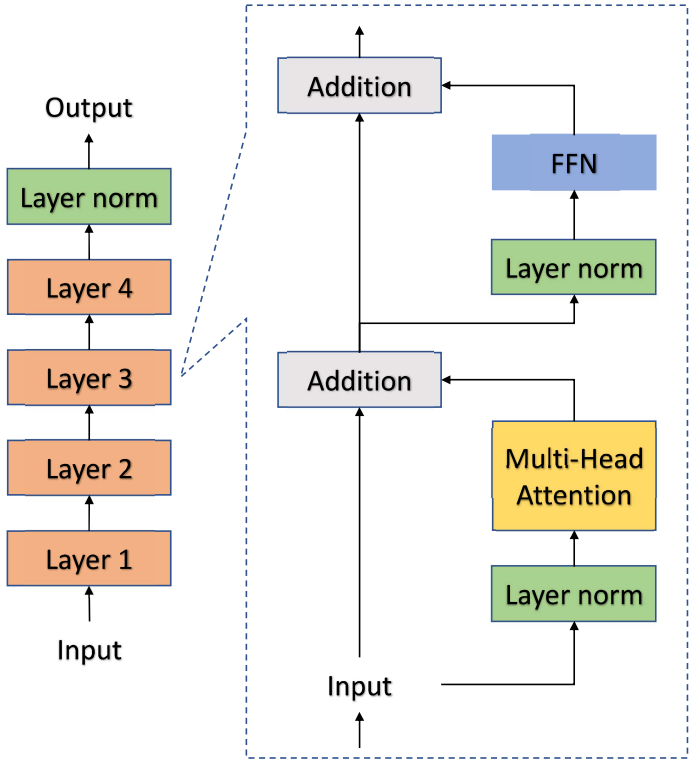


Lemma 3. *o_m has zero mean and $\frac{d^2\sigma^4}{m}I$ covariance matrix.*

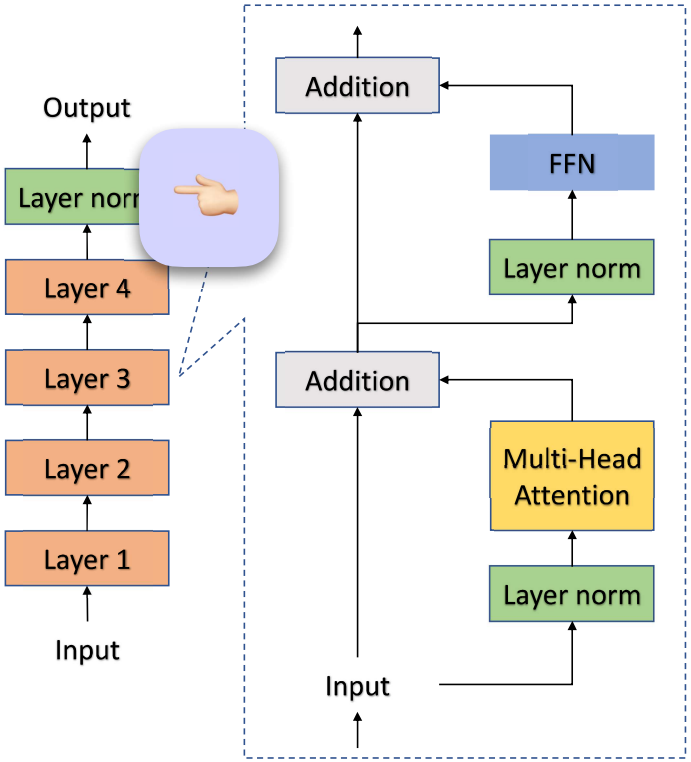
$$\boldsymbol{o}_m = \boldsymbol{W}_o \left(\oplus_{h=1}^{H} \sum_{n=1}^{L} a_{mn}^{(h)} \boldsymbol{v}_n^{(h)} \right),$$



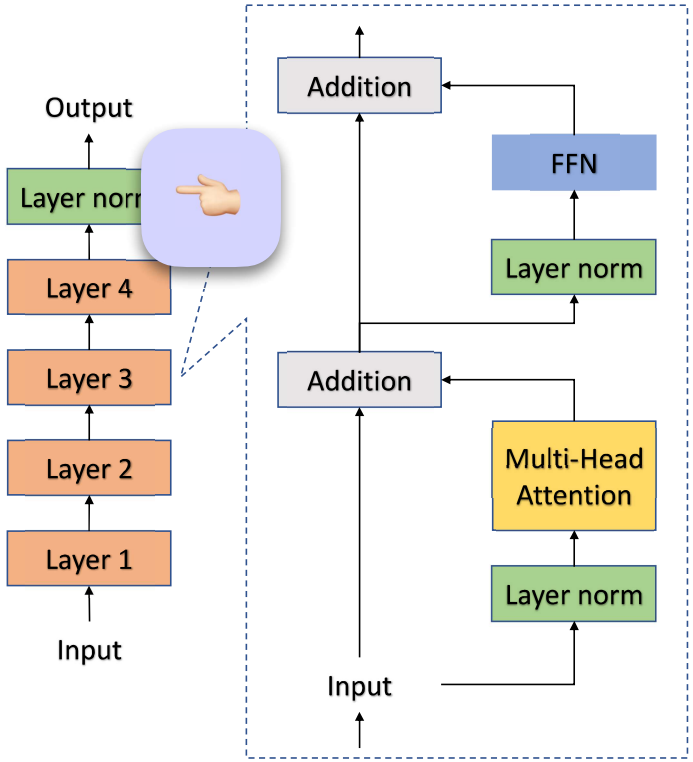
Final Layer Norm



Final Layer Norm



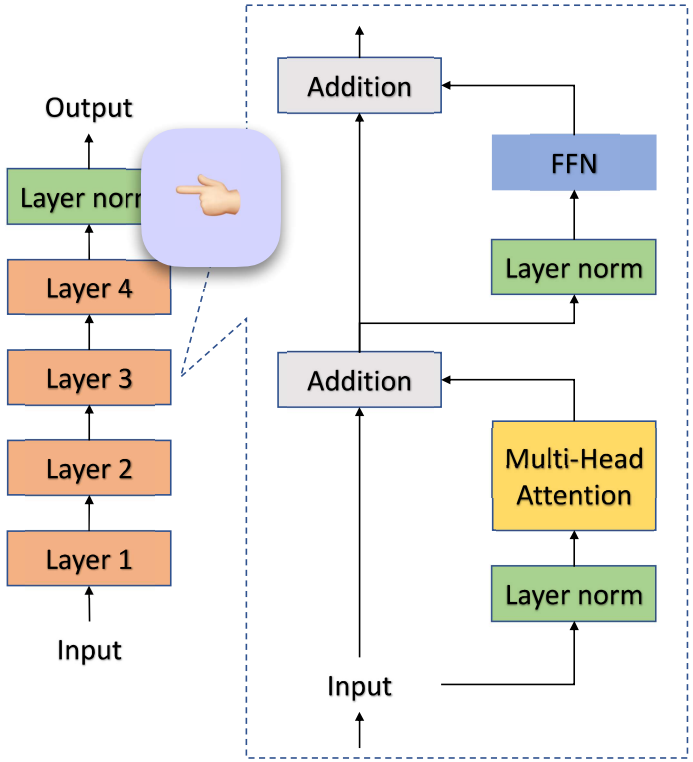
Final Layer Norm



Lemma 4. *The variance of the j-th dimension of $\boldsymbol{y}_m$ is:*

$$\frac{m\sigma^2 + \sum_i (\boldsymbol{W}_{o,j:}\boldsymbol{W}_{v,:i})^2}{m\sigma^2 + d^2\sigma^4},$$

Final Layer Norm



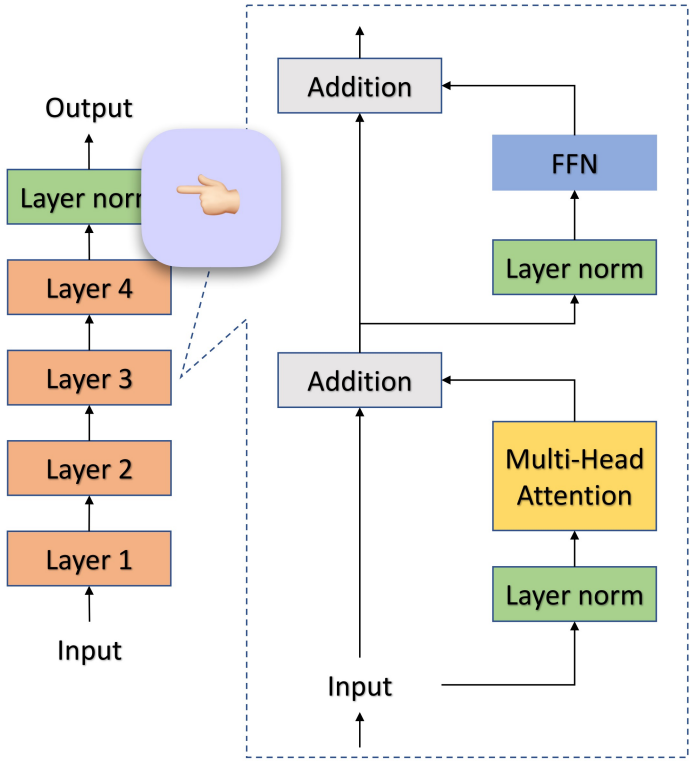
Lemma 4. *The variance of the j-th dimension of $\boldsymbol{y}_m$ is:*

$$\frac{m\sigma^2 + \sum_i (\boldsymbol{W}_{o,j:}\boldsymbol{W}_{v,:i})^2}{m\sigma^2 + d^2\sigma^4},$$

$$\boldsymbol{y}'_{mi} \approx \frac{\boldsymbol{y}_{mi} - \mathbb{E}[\boldsymbol{y}_{mi}]}{\sqrt{\mathbb{V}[\boldsymbol{y}_{mi}]}} \approx \frac{\boldsymbol{x}_{mi} + \boldsymbol{W}_o\boldsymbol{W}_v\frac{\sum_n^m \boldsymbol{e}_{ni}}{m}}{\sqrt{\sigma^2 + \frac{d^2\sigma^4}{m}}},$$

$$\mathbb{E}[\boldsymbol{y}_{mi}] = 0,\ \mathbb{V}[\boldsymbol{y}_{mi}] = \mathbb{V}[\boldsymbol{x}_{mi}] + \mathbb{V}[\boldsymbol{o}_{mi}]$$
$$= \sigma^2 + \frac{d^2\sigma^4}{m}$$

Final Layer Norm: Proof



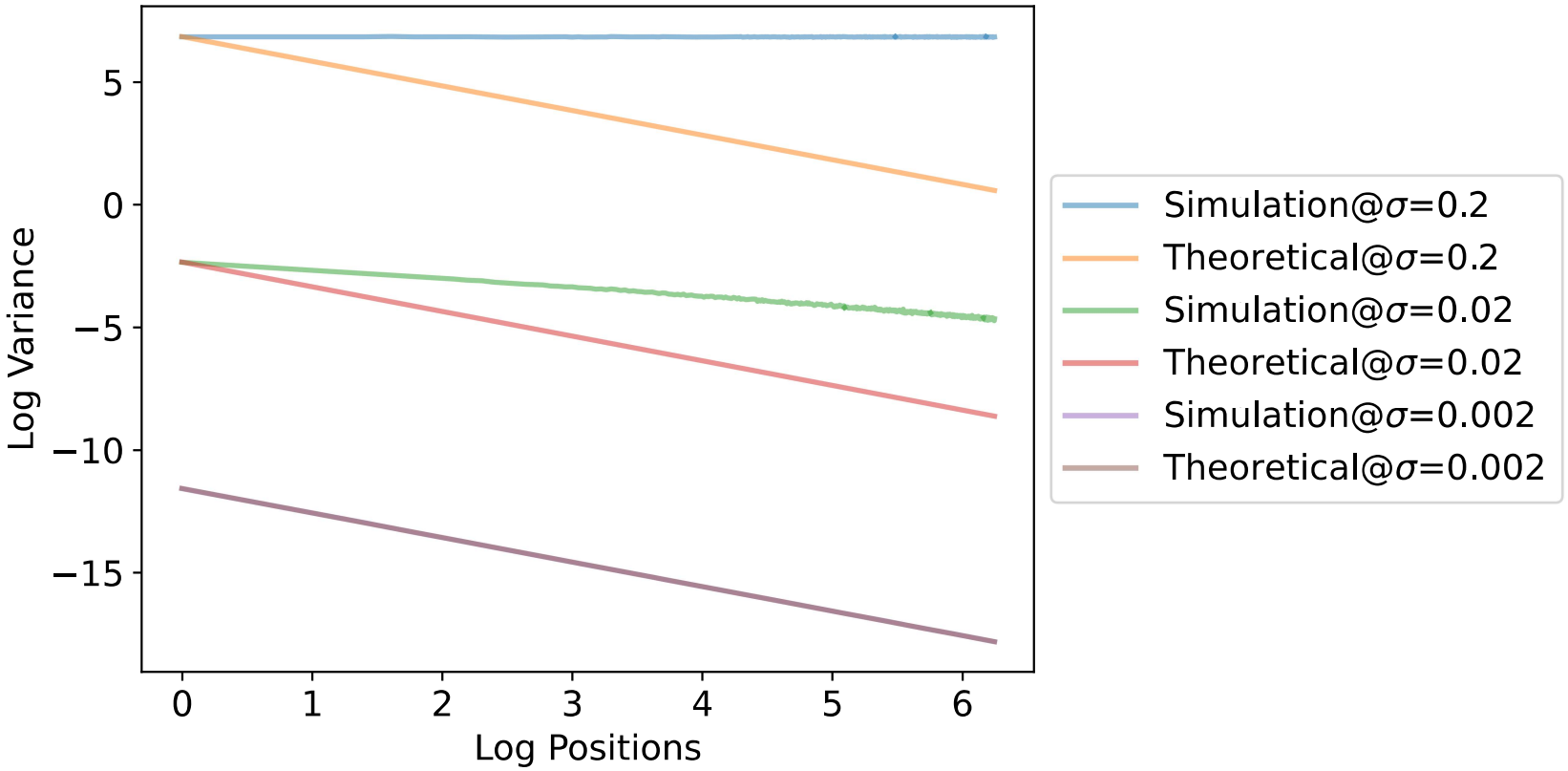
Lemma 4. *The variance of the j-th dimension of $\boldsymbol{y}_m$ is:*

$$\frac{m\sigma^2 + \sum_i (\boldsymbol{W}_{o,j:}\boldsymbol{W}_{v,:i})^2}{m\sigma^2 + d^2\sigma^4},$$

$$\begin{aligned}
&\mathrm{cov}(\boldsymbol{o}_m, \boldsymbol{o}_m) \\
&\overset{\text{Property 1}}{\approx} \mathbb{E}\left[\boldsymbol{W}_o \frac{\sum_{i=1}^m \boldsymbol{v}_i}{m} \frac{\sum_{j=1}^m \boldsymbol{v}_j^\top}{m} \boldsymbol{W}_o^\top\right] \\
&= \frac{1}{m^2}\sum_{i,j=1}^m \mathbb{E}[\boldsymbol{W}_o \boldsymbol{v}_i \boldsymbol{v}_j^\top \boldsymbol{W}_o^\top] \\
&= \frac{1}{m^2}\sum_{i,j=1}^m \mathbb{E}\left[\begin{bmatrix} \boldsymbol{r}_1^\top \boldsymbol{v}_i \\ \vdots \\ \boldsymbol{r}_d^\top \boldsymbol{v}_i \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_j^\top \boldsymbol{r}_1 & \dots & \boldsymbol{v}_j^\top \boldsymbol{r}_d \end{bmatrix}\right] \\
&= \frac{1}{m^2}\sum_{i,j=1}^m \left[\mathbb{E}[\boldsymbol{r}_k^\top \boldsymbol{v}_i \boldsymbol{v}_j^\top \boldsymbol{r}_l]\right]_{k,l=1}^d \\
&= \frac{1}{m^2}\sum_{i,j=1}^m \left[\mathbb{E}[\mathrm{Tr}(\boldsymbol{r}_l \boldsymbol{r}_k^\top \boldsymbol{v}_i \boldsymbol{v}_j^\top)]\right]_{k,l=1}^d \\
&= \frac{1}{m^2}\sum_{i,j=1}^m \left[\mathrm{Tr}(\mathbb{E}[\boldsymbol{r}_l \boldsymbol{r}_k^\top]\mathbb{E}[\boldsymbol{v}_i \boldsymbol{v}_j^\top])\right]_{k,l=1}^d \\
&\overset{(*)}{=} \frac{1}{m^2}\sum_{i,j=1}^m \left[\mathrm{Tr}((\mathbb{1}_{k=l}\sigma^2)\cdot I \cdot (\mathbb{1}_{i=j} d\sigma^2)\cdot I)\right]_{k,l=1}^d \\
&= \frac{d^2\sigma^4}{m} I
\end{aligned}$$

Relaxing

? What if Property 1 does not hold?



Relaxing

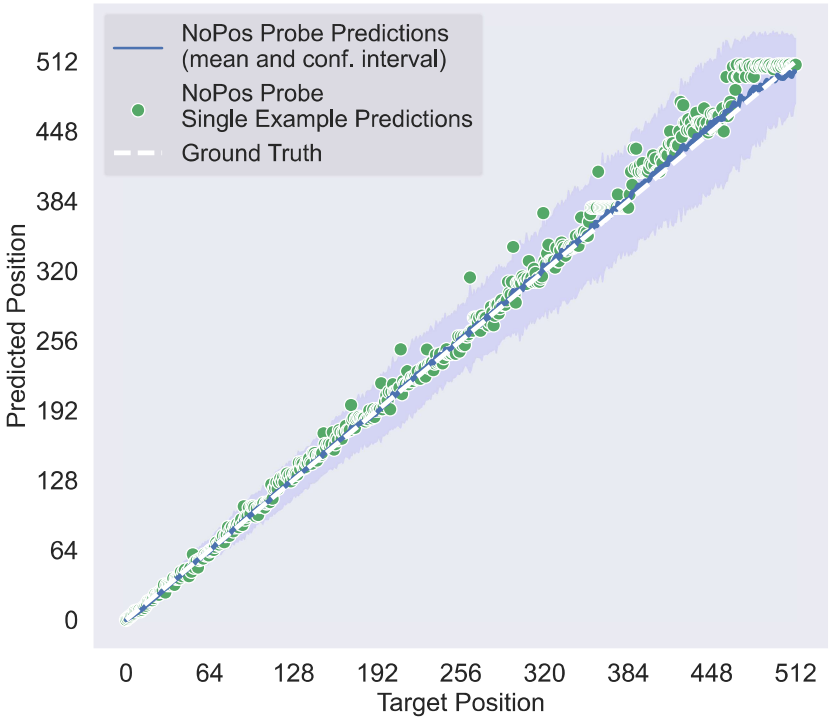
? Other Initialization Schemes

$(\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v, \boldsymbol{W}_o)$

i.i.d. samples from a distribution

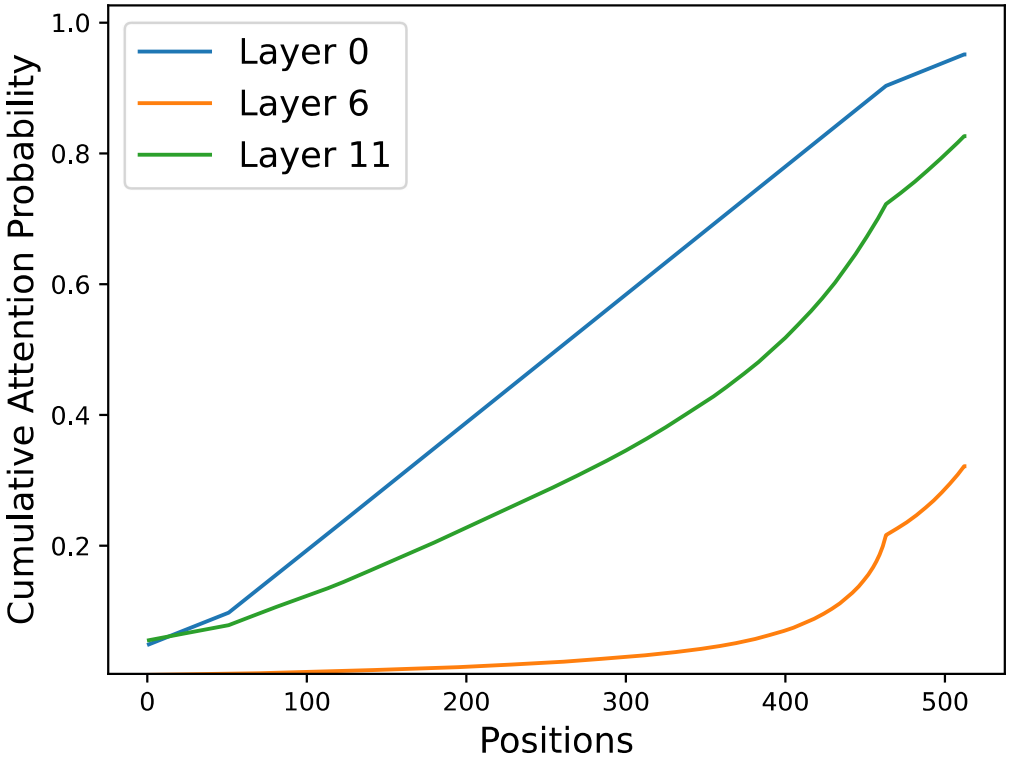
with zero mean and finite variance.

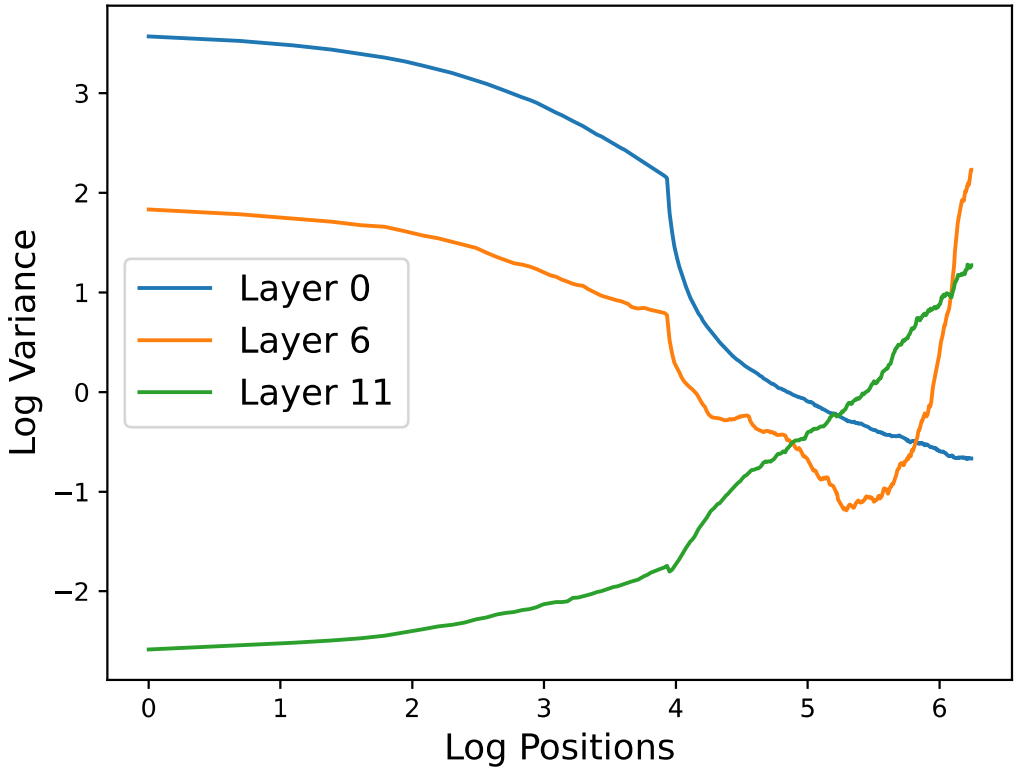
Discussions



Lemma 3. *o_m has zero mean and $\frac{d^2\sigma^4}{m}I$ covariance matrix.*

Discussions: Pretrained





Discussions

? Why does BERT fail to converge NoPE

Lemma 3. *o_m has zero mean and $\frac{d^2\sigma^4}{L}I$ covariance matrix.*

ACL2023 Honorable Mentions

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

Ting-Han Fan
Princeton University

Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

Peter J. Ramadge
Princeton University

ACL2023 Honorable Mentions

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

Ting-Han Fan
Princeton University

Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

Peter J. Ramadge
Princeton University

😲 NoPos 👌 why → variance

ACL2023 Honorable Mentions

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

Ting-Han Fan
Princeton University

Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

Peter J. Ramadge
Princeton University

😲 NoPos 👌 why → variance

initialization🙅‍♀️ pretrained🤗; Downstream tasks🤔

ACL2023 Honorable Mentions

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

Ting-Han Fan
Princeton University

Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

Peter J. Ramadge
Princeton University

😲 NoPos 👌 why → variance

initialization🙅‍♀️ pretrained🤗; Downstream tasks🤔

<S> $\{\boldsymbol{x}_m\}_{m=1}^{L}$

苏剑林 发表于 August 11th, 2023

在我的实验中，NoPE没有任何长度外推性，甚至连RoPE都不如，所以我无法评价

回复评论

Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

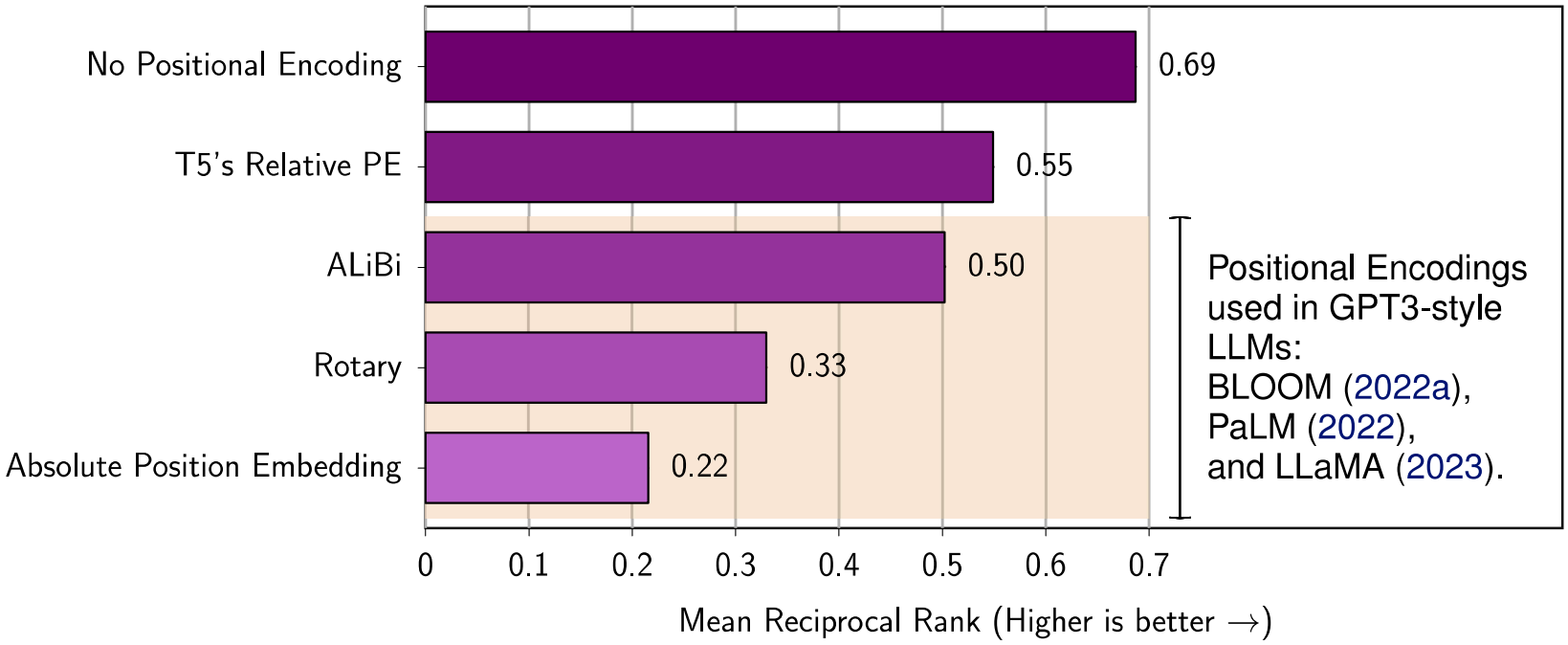
Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

😯 NoPE 👌 But why? 🤔

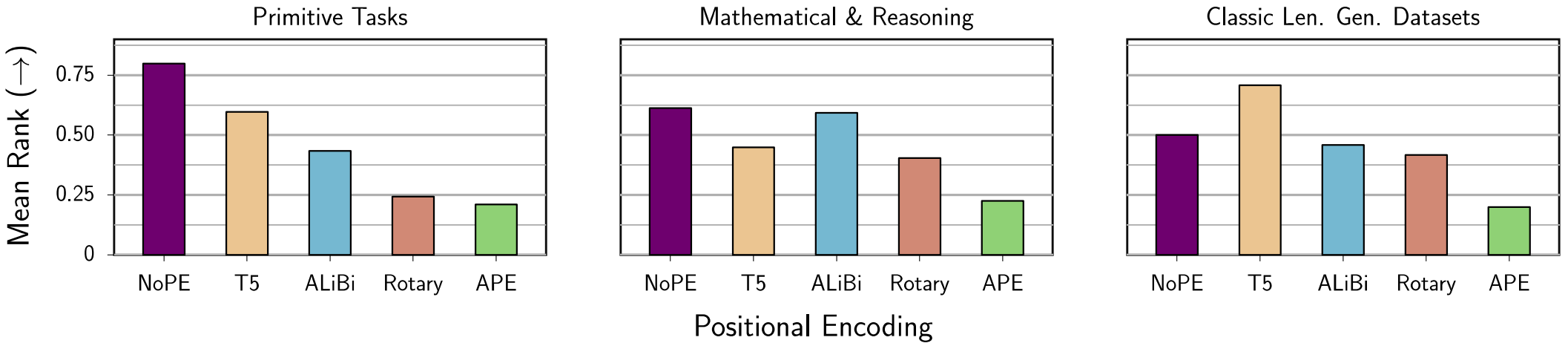
NoPE vs PE


No Positional Encoding
T5's Relative PE
ALiBi
Rotary
Absolute Position Embedding
0.69
0.55
0.50
0.33
0.22
Positional Encodings used in GPT3-style LLMs: BLOOM (2022a), PaLM (2022), and LLaMA (2023).
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Mean Reciprocal Rank (Higher is better →)

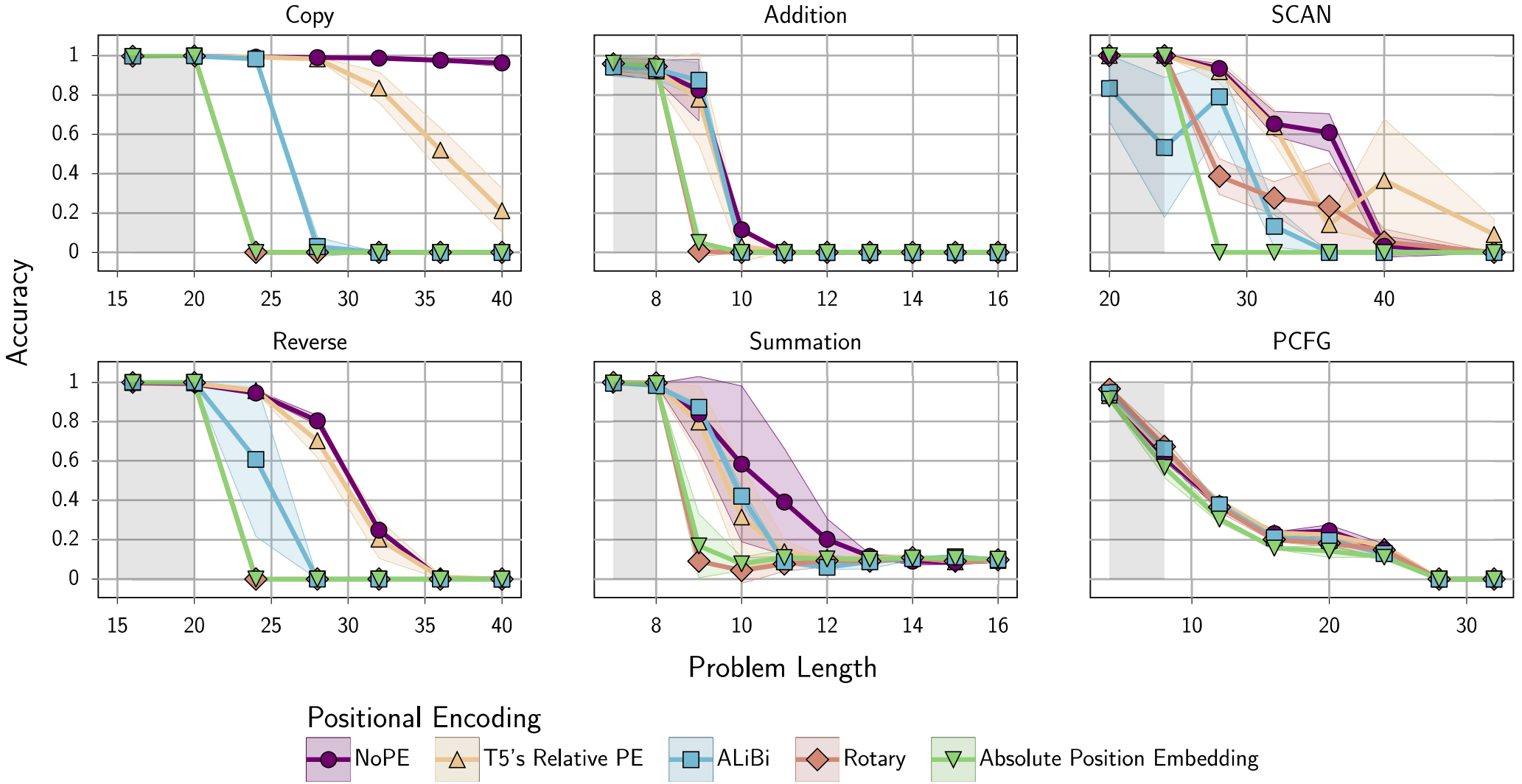
Length Generalization

Task	Input Example	Output Example
	Primitive Tasks	
Copy	`Copy the following words: <w1> <w2> <w3> <w4> <w5>`	`<w1> <w2> <w3> <w4> <w5>`
Reverse	`Reverse the following words: <w1> <w2> <w3> <w4> <w5>`	`<w5> <w4> <w3> <w2> <w1>`
	Mathematical and Algroithmic Tasks	
Addition	`Compute: 5 3 7 2 6 + 1 9 1 7 ?`	`The answer is 5 5 6 4 3.`
Polynomial Eval.	`Evaluate x = 3 in ( 3 x ** 0 + 1 x ** 1 + 1 x ** 2 ) % 10 ?`	`The answer is 5.`
Sorting	`Sort the following numbers: 3 1 4 1 5 ?`	`The answer is 1 1 3 4 5.`
Summation	`Compute: ( 1 + 2 + 3 + 4 + 7 ) % 10 ?`	`The answer is 7.`
Parity	`Is the number of 1's even in [ 1 0 0 1 1] ?`	`The answer is No.`
LEGO	`If a = -1; b = -a; c = +b; d = +c. Then what is c?`	`The answer is +1.`
	Classical Length Generalization Datasets	
SCAN	`jump twice and run left`	`JUMP JUMP TURN_LEFT RUN`
PCFG	`shift prepend K10 R1 K12 , E12 F16`	`F16 K10 R1 K12 E12`

Length Generalization



Length Generalization

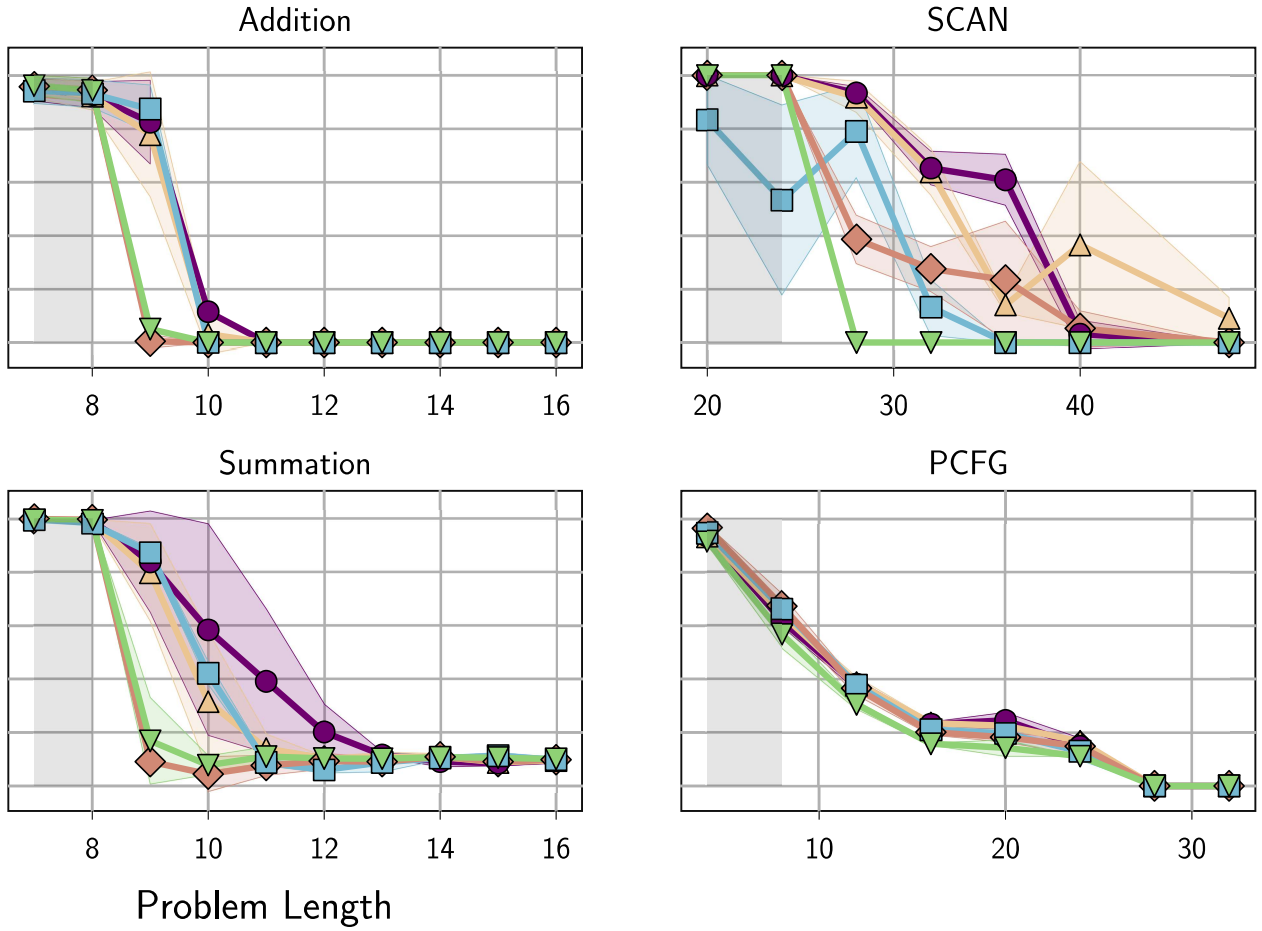


Length Generalization

Length Generalization

👇

PE

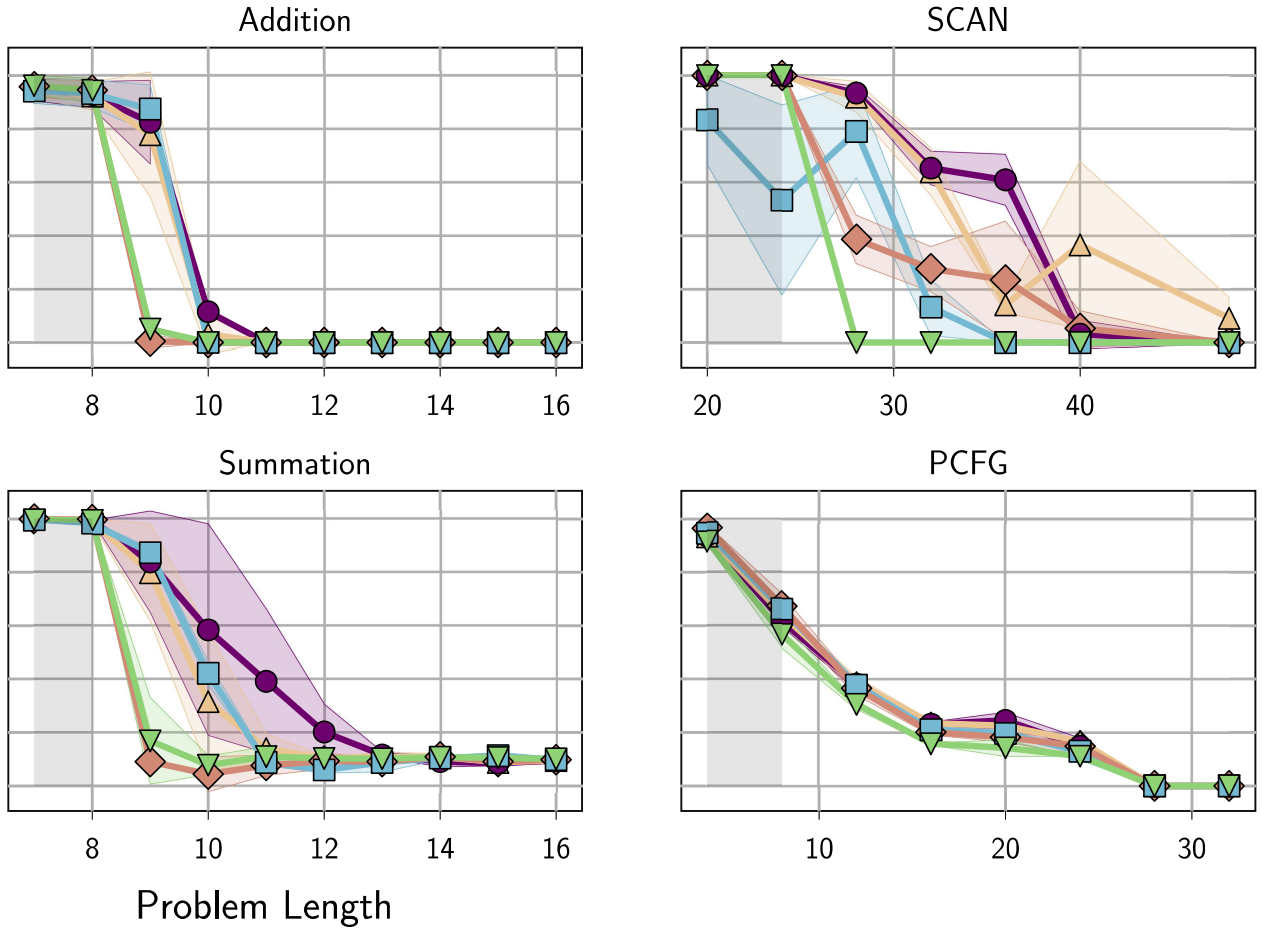


Length Generalization

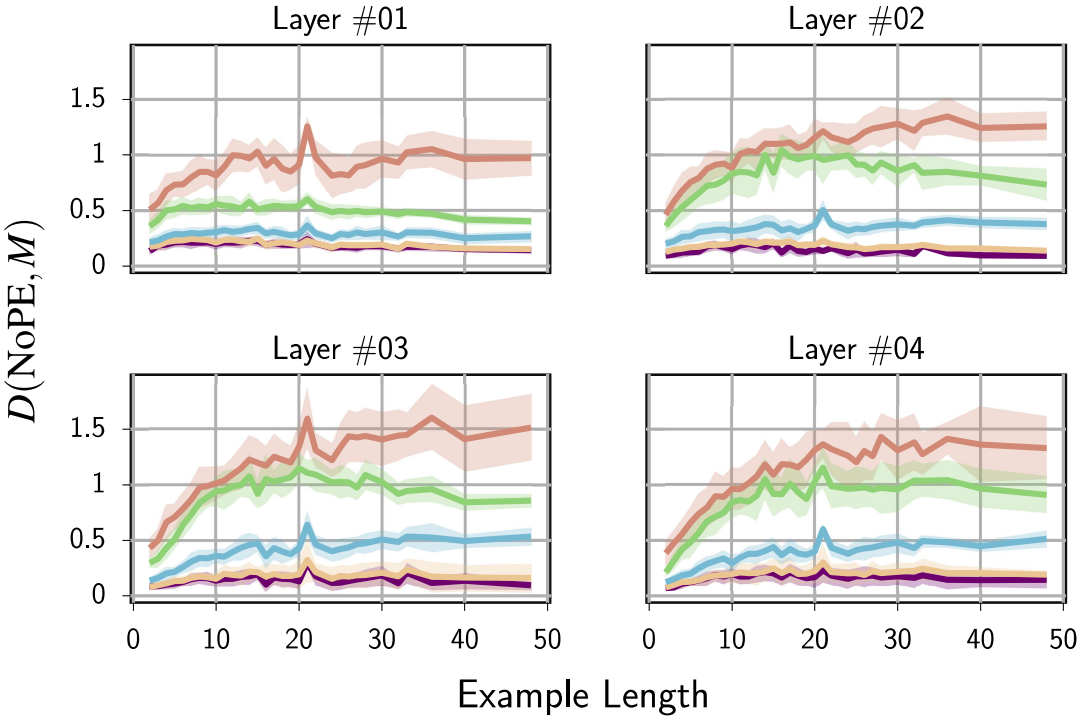
Length Generalization

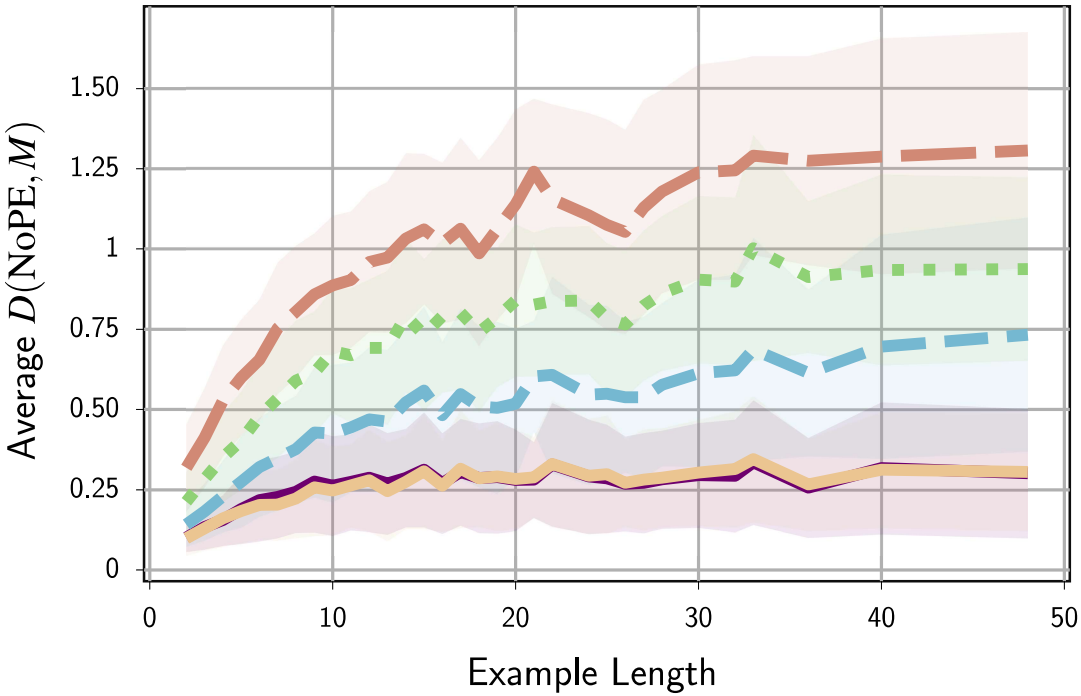
👇 👇 😈

PE



NoPE 🤝 T5-rel





Positional Encoding (M)

NoPE' | T5's Relative PE | ALiBi | Rotary | Absolute Position Embedding

$$D_{\mathrm{AT}}(\mathrm{P}, \mathrm{Q}) = \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{JSD}}(\mathrm{P}_t || \mathrm{Q}_t)$$

$$D^{(l)}(A, B) = \min_{(\mathrm{P}, \mathrm{Q}) \in A_l \times B_l} D_{\mathrm{AT}}(\mathrm{P}, \mathrm{Q})$$

NoPE still has abs-PE

input sequence $\boldsymbol{x} = [\texttt{<bos>}, x_1, \ldots, x_T]$

Theorem 1 (**Absolute Encoding**). *Let $\boldsymbol{x}$ be an input sequence of length $T+1$ to the model. Then, the first layer of f_θ can recover absolute positions $[1, \ldots, T+1]$ in the hidden state $\boldsymbol{H}^{(1)}$. That is, there exist $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, $\boldsymbol{W}_O$, $\boldsymbol{W}_1$, and $\boldsymbol{W}_2$ such that the self-attention and feedforward operations in the first layer compute absolute positions and write it to the next hidden state.*

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_E = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ e_{4,1} & e_{4,2} & e_{4,3} & \dots & e_{4,V} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{d,1} & e_{d,2} & e_{d,2} & \dots & e_{d,V} \end{bmatrix}_{d \times V}$$

$$\boldsymbol{H}^{(0)} = \boldsymbol{W}_E \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ e_{4,1} & e_{4,2} & e_{4,3} & \dots & e_{4,V} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{d,1} & e_{d,2} & e_{d,2} & \dots & e_{d,V} \end{bmatrix}_{d \times (T+1)}$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_K = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}_{h \times d}$$

$$\boldsymbol{k}_i = \boldsymbol{W}_K \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{k}_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{k}_2 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \dots \quad \boldsymbol{k}_t = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_K = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}_{h \times d}$$

$$\boldsymbol{k}_i = \boldsymbol{W}_K \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{k}_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{k}_2 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \dots \quad \boldsymbol{k}_t = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$\boldsymbol{q}_t = \boldsymbol{W}_Q \boldsymbol{h}_t^{(0)}$:

$$\boldsymbol{q}_t = [q_1, q_2, q_3, \dots, q_h]^\intercal$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_K = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}_{h \times d}$$

$$\boldsymbol{k}_i = \boldsymbol{W}_K \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{k}_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{k}_2 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \dots \quad \boldsymbol{k}_t = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$\boldsymbol{q}_t = \boldsymbol{W}_Q \boldsymbol{h}_t^{(0)}$:

$$\boldsymbol{q}_t = [q_1, q_2, q_3, \dots, q_h]^\intercal$$

$$\begin{aligned} \boldsymbol{\alpha} &= [\langle \boldsymbol{q}_t, \boldsymbol{k}_1 \rangle, \langle \boldsymbol{q}_t, \boldsymbol{k}_2 \rangle, \dots, \langle \boldsymbol{q}_t, \boldsymbol{k}_t \rangle]^\intercal \\ &= [\alpha^*, \alpha^*, \dots, \alpha^*]^\intercal \end{aligned}$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_K = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}_{h \times d}$$

$$\boldsymbol{k}_i = \boldsymbol{W}_K \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{k}_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{k}_2 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \dots \quad \boldsymbol{k}_t = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$\boldsymbol{q}_t = \boldsymbol{W}_Q \boldsymbol{h}_t^{(0)}$:

$$\boldsymbol{q}_t = [q_1, q_2, q_3, \dots, q_h]^\intercal$$

$$\begin{aligned} \boldsymbol{\alpha} &= [\langle \boldsymbol{q}_t, \boldsymbol{k}_1 \rangle, \langle \boldsymbol{q}_t, \boldsymbol{k}_2 \rangle, \dots, \langle \boldsymbol{q}_t, \boldsymbol{k}_t \rangle]^\intercal \\ &= [\alpha^*, \alpha^*, \dots, \alpha^*]^\intercal \end{aligned}$$

$$\hat{\boldsymbol{\alpha}} = \text{softmax}(\boldsymbol{\alpha}) = \left[\frac{1}{t}, \frac{1}{t}, \dots, \frac{1}{t}\right]^\intercal$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_V = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}_{h \times d}$$

$$\boldsymbol{v}_i = \boldsymbol{W}_V \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \boldsymbol{v}_t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_V = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}_{h \times d}$$

$$\boldsymbol{v}_i = \boldsymbol{W}_V \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \boldsymbol{v}_t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\boldsymbol{W}_O = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}_{d \times h}$$

NoPE still has abs-PE: Proof

$$\boldsymbol{W}_V = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}_{h \times d} \qquad \boldsymbol{v}_i = \boldsymbol{W}_V \boldsymbol{h}_i^{(0)}$$

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \boldsymbol{v}_t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\boldsymbol{W}_O = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}_{d \times h}$$

$$\boldsymbol{o}_t = \boldsymbol{W}_O \left(\sum_{i \le t} \hat{\boldsymbol{\alpha}}_i \boldsymbol{v}_i \right) = \boldsymbol{W}_O \left(\frac{1}{t} \sum_{i \le t} \boldsymbol{v}_i \right) = \boldsymbol{W}_O \begin{pmatrix} 1/t \\ 0 \\ \vdots \\ 0 \end{pmatrix}_h = \begin{pmatrix} 0 \\ 0 \\ 1/t \\ 0 \\ \vdots \\ 0 \end{pmatrix}_d$$

NoPE still has rel-PE

Theorem 2 (Relative Encoding). *Suppose that the hidden state $\boldsymbol{H}^{(1)}$ contains absolute positional information, as stated in Theorem 1, and assume that it is not overwritten by any subsequent layers. Then, the self-attention in all subsequent layers can implement a relative positional encoding: there exists a parameterization of f_θ such that, for $l \geq 2$, the attention dot product between query $\boldsymbol{q}_t$ and key $\boldsymbol{k}_i$ at positions t and i ($t \geq i$) can be expressed as:*

$$\langle \boldsymbol{q}_t, \boldsymbol{k}_i \rangle = f_{\mathrm{cnt}}(\boldsymbol{q}_t, \boldsymbol{k}_i) + f_{\mathrm{rel}}(t - i) \tag{1}$$

where f_{cnt} is a function of their content, and f_{rel} is a function of their relative distance.

NoPE still has rel-PE: Proof

$$\boldsymbol{H}^{(l)} = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 2 & 3 & 4 & \dots & T+1 \\ h_{4,1} & h_{4,2} & h_{4,3} & h_{4,4} & \dots & h_{4,T+1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{d,1} & h_{d,2} & h_{d,3} & h_{d,4} & \dots & h_{d,T+1} \end{bmatrix}_{d \times (T+1)}$$

$$\boldsymbol{W}_Q = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 0 & \dots & 0 \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} & \dots & w_{3,d} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{h,1} & w_{h,2} & w_{h,3} & w_{h,4} & \dots & w_{h,d} \end{bmatrix}_{h \times d}$$

$$\boldsymbol{q}_t = \boldsymbol{W}_Q \boldsymbol{h}_t^{(l)}:$$

$$\boldsymbol{q}_t = [1, -t, q_3, \dots, q_h]^\intercal$$

NoPE still has rel-PE: Proof

$$\boldsymbol{W}_V = \begin{bmatrix} 0 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 \\ w'_{3,1} & w'_{3,2} & w'_{3,3} & w'_{3,4} & \dots & w'_{3,d} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w'_{h,1} & w'_{h,2} & w'_{h,3} & w'_{h,4} & \dots & w'_{h,d} \end{bmatrix}_{h \times d}$$

$$\boldsymbol{k}_1 = \begin{pmatrix} 1 \\ 1 \\ k_{3,1} \\ \vdots \\ k_{h,1} \end{pmatrix} \quad \boldsymbol{k}_2 = \begin{pmatrix} 2 \\ 1 \\ k_{3,2} \\ \vdots \\ k_{h,2} \end{pmatrix} \quad \boldsymbol{k}_3 = \begin{pmatrix} 3 \\ 1 \\ k_{3,3} \\ \vdots \\ k_{h,3} \end{pmatrix} \quad \dots \quad \boldsymbol{k}_t = \begin{pmatrix} t \\ 1 \\ k_{3,t} \\ \vdots \\ k_{h,t} \end{pmatrix}$$

$$\boldsymbol{k}_i = [i, 1, k_{3,i}, \dots, k_{h,i}]^\intercal$$

NoPE still has rel-PE: Proof

$$\boldsymbol{q}_t = [1, -t, q_3, \ldots, q_h]^\intercal \qquad \boldsymbol{k}_i = [i, 1, k_{3,i}, \ldots, k_{h,i}]^\intercal$$

$$\begin{aligned}
\langle \boldsymbol{q}_t, \boldsymbol{k}_i \rangle =& 1.i + (-t).1 + q_3.k_{3,i} + \cdots + q_h.k_{h,i} \\
=& i - t + \sum_{j=3}^{h} q_j.k_{j,i} \\
=& \left(\sum_{j=3}^{h} q_j.k_{j,i} \right) - (t - i) \\
=& f_{\text{cnt}}(\boldsymbol{q}_t, \boldsymbol{k}_i) + f_{\text{rel}}(t - i)
\end{aligned}$$

CoT

Input (x)

```
Compute 5 3 7 2 6 + 1 9 1 7 =
```

Output (s; y)

```
<scratch>
```

$\mathcal{I}$ `For digits 6 and 7,`

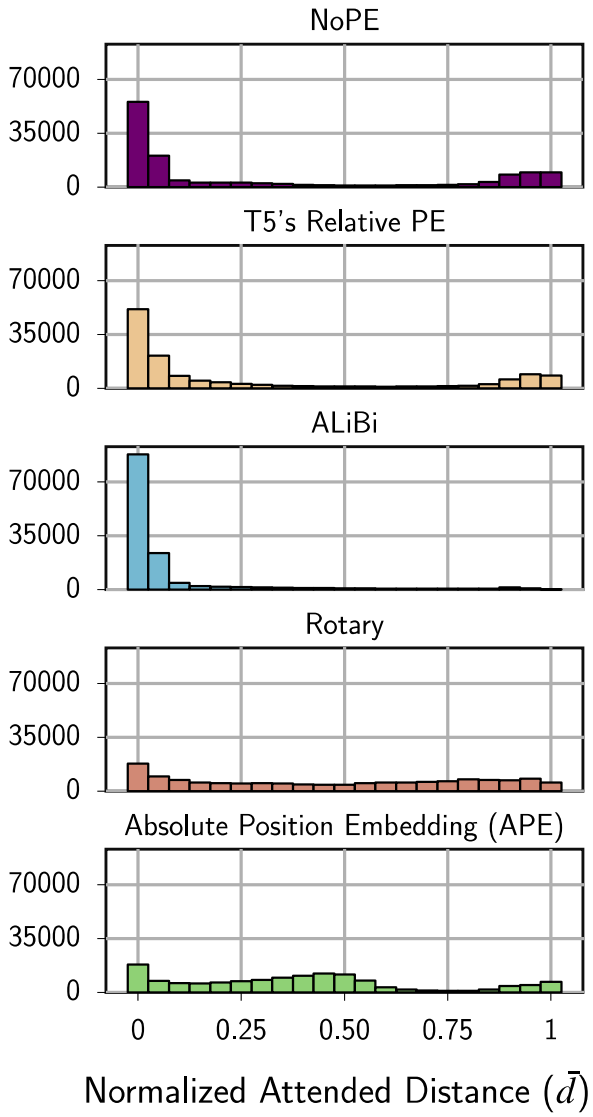
$\mathcal{C}$ `We have ( 6 + 7 + carry ) % 10 = 13 % 10 = 13 % 10`

$\mathcal{O}$ `Which is equal to 3 .`

$\mathcal{V}$ `We update carry to 13 // 10 = 1.`

$\mathcal{R}$ `So, the remaining input is 5 3 7 2 + 1 9 1`

```
...
</scratch>
The answer is 5 5 6 4 3.
```



ACL2022 Findings Meta

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

ACL2023 Honorable Mentions CMU&Princeton

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†]
Carnegie Mellon University

Ting-Han Fan
Princeton University

Li-Wei Chen
Carnegie Mellon University

Alexander I. Rudnicky
Carnegie Mellon University

Peter J. Ramadge
Princeton University

Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

ACL2022 Findings Meta

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

NoPE 👌

ACL2023 Honorable Mentions CMU&Princeton

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†] Carnegie Mellon University

Ting-Han Fan Princeton University

Li-Wei Chen Carnegie Mellon University

Alexander I. Rudnicky Carnegie Mellon University

Peter J. Ramadge Princeton University

NoPE 👌

Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

NoPE 👌

ACL2022 Findings Meta

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

NoPE 👌

ACL2023 Honorable Mentions CMU&Princeton

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†] Carnegie Mellon University

Ting-Han Fan Princeton University

Li-Wei Chen Carnegie Mellon University

Alexander I. Rudnicky Carnegie Mellon University

Peter J. Ramadge Princeton University

NoPE 👌 why? 🤔

Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

NoPE 👌 why? 🤔

ACL2022 Findings Meta

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

NoPE 👌

ACL2023 Honorable Mentions CMU&Princeton

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†] Carnegie Mellon University

Ting-Han Fan Princeton University

Li-Wei Chen Carnegie Mellon University

Alexander I. Rudnicky Carnegie Mellon University

Peter J. Ramadge Princeton University

NoPE 👌 why? 🤔

Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

NoPE 👌 why? 🤔 🌊Task

ACL2022 Findings Meta

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv[τ] **Ori Ram**[τ] **Ofir Press**[ω] **Peter Izsak**[ι] **Omer Levy**[τμ]

[τ]Tel Aviv University [ω]University of Washington [ι]Intel Labs [μ]Meta AI

NoPE 👌

ACL2023 Honorable Mentions CMU&Princeton

Latent Positional Information is in the Self-Attention Variance of Transformer Language Models Without Positional Embeddings

Ta-Chung Chi[†] Carnegie Mellon University

Ting-Han Fan Princeton University

Li-Wei Chen Carnegie Mellon University

Alexander I. Rudnicky Carnegie Mellon University

Peter J. Ramadge Princeton University

NoPE 👌 why? 🤔

Preprint Mila

The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad[1,2], **Inkit Padhi**[3]
Karthikeyan Natesan Ramamurthy[3], **Payel Das**[3], **Siva Reddy**[1,2,4]
[1]Mila - Québec AI Institute; [2]McGill University;
[3]IBM Research; [4]Facebook CIFAR AI Chair

NoPE 👌 why? 🤔 🌊Task

👇 👇 👇