Positional Encodings and Length Generalization for Generative Transformers

2024-11-22

Li Peng-Hsuan 李朋軒

Agenda

Positional Encodings

→ Transformer, absolute PE, relative PE

No Positional Encodings

→ Generative transformer, NoPE

Length Generalization

→ Out-of-distribution, extrapolation, interpolation

Agenda

Positional Encodings

→ Transformer, absolute PE, relative PE

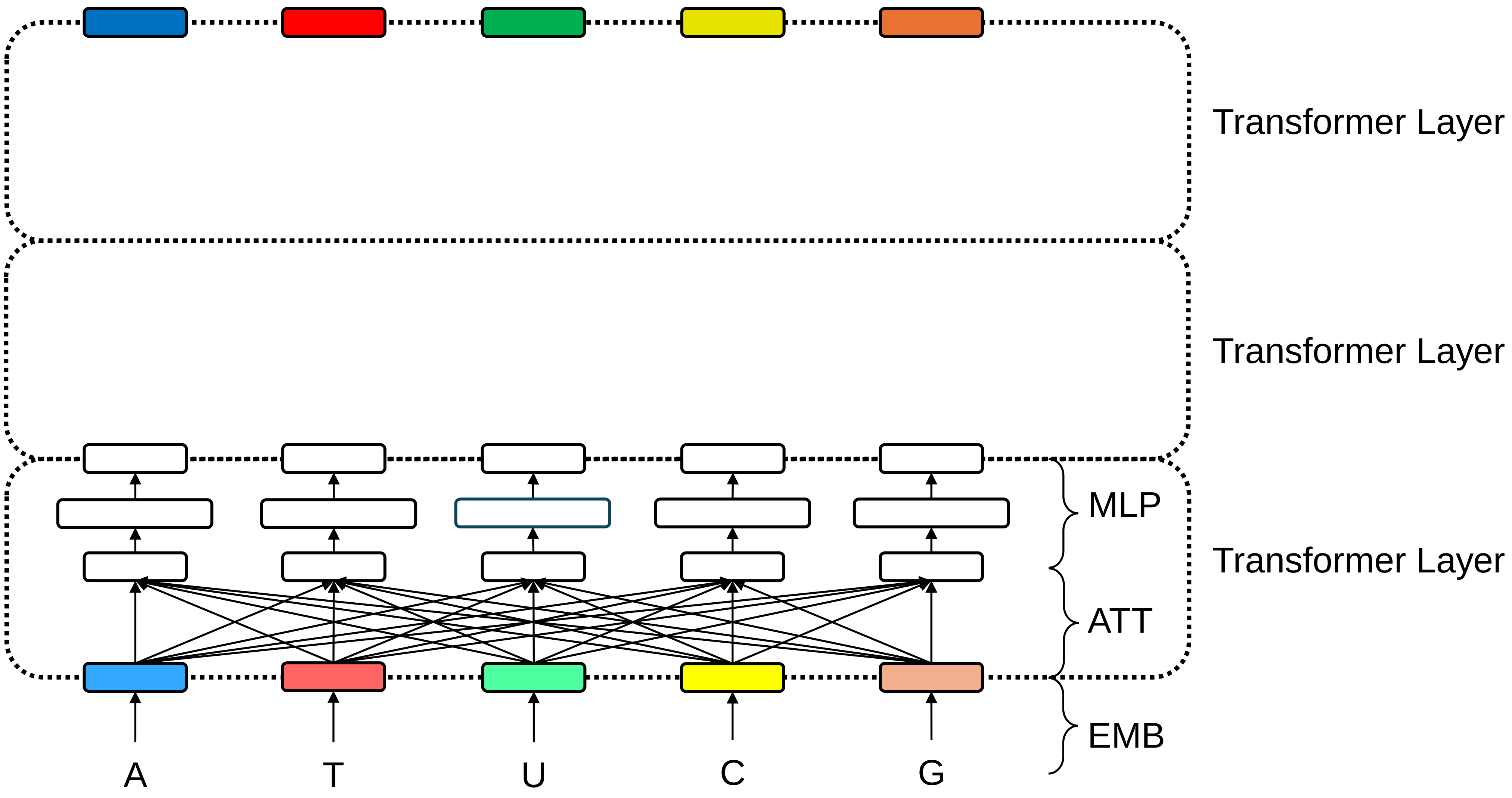
No Positional Encodings

→ Generative transformer, NoPE

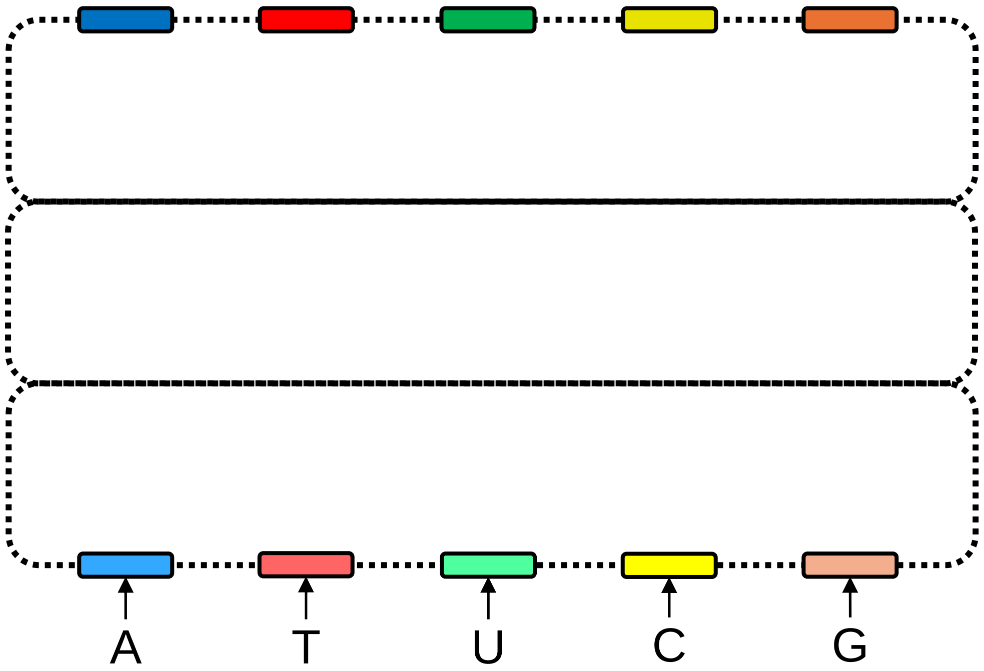
Length Generalization

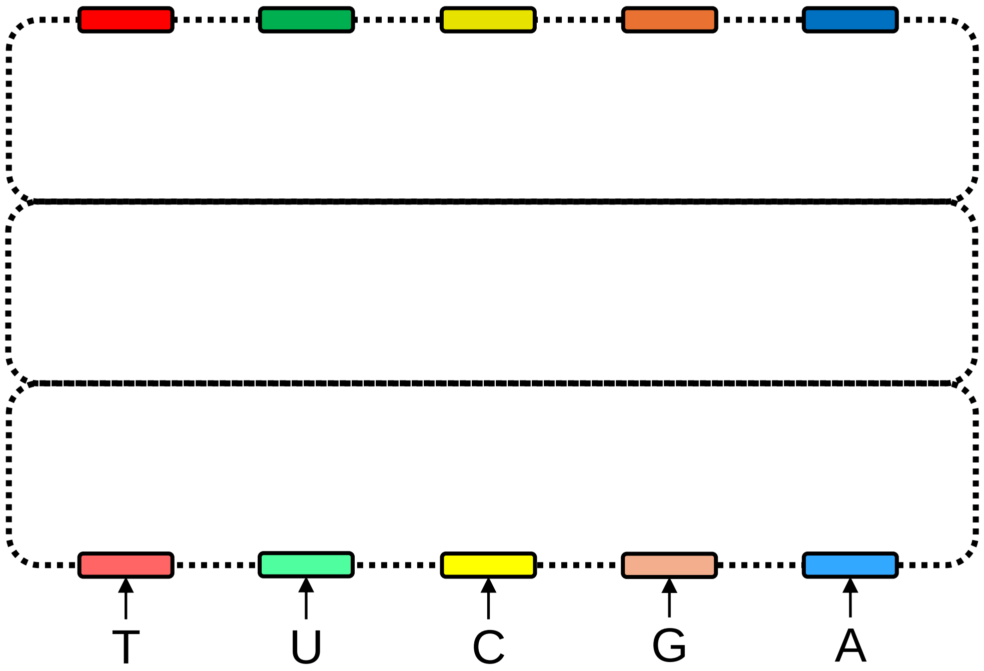
→ Out-of-distribution, extrapolation, interpolation

Transformer

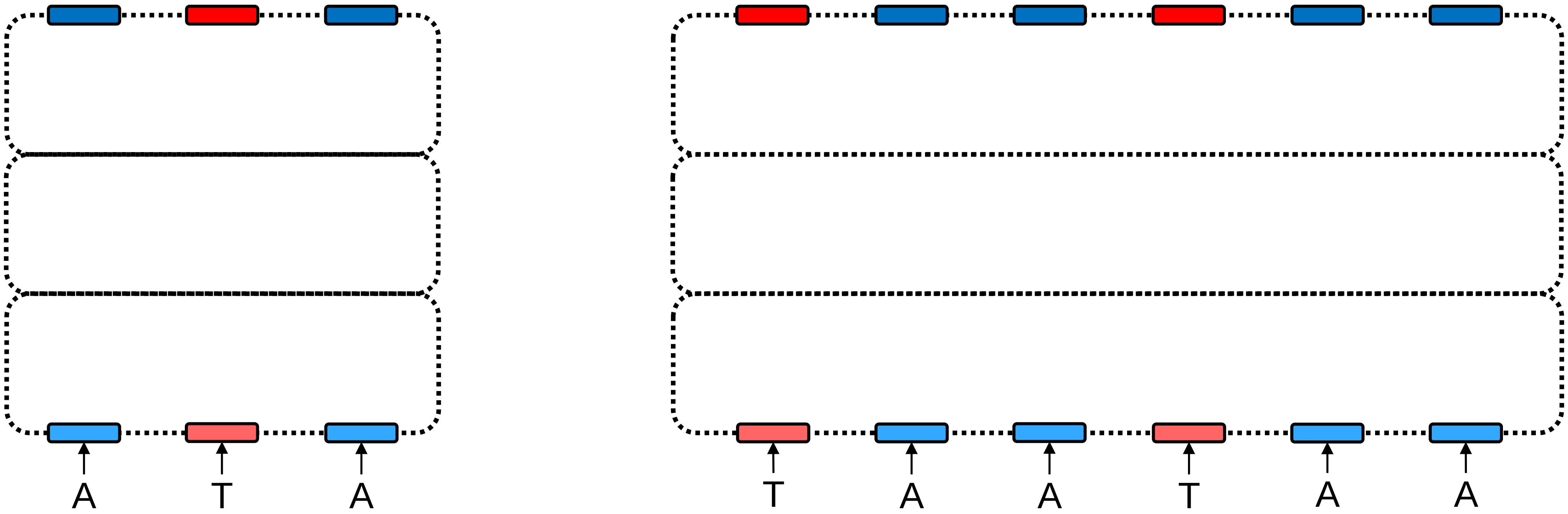


Transformer: Permutation Equivariant

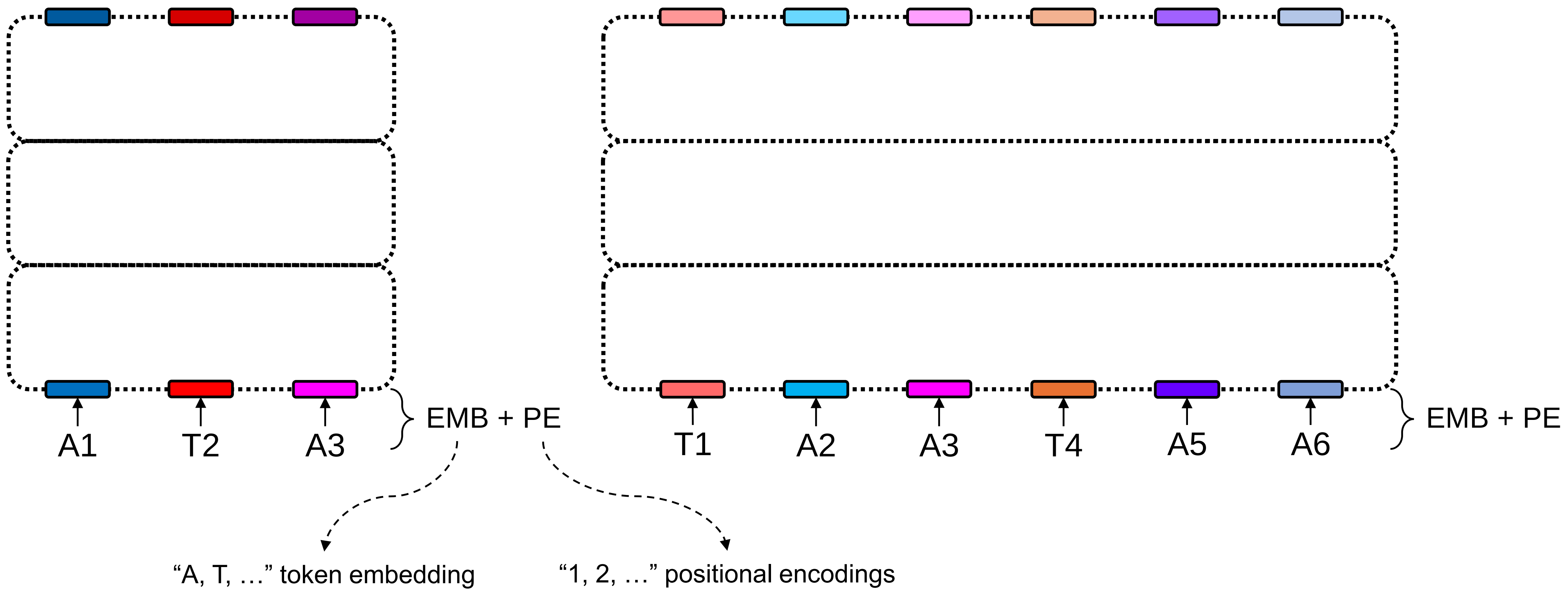




Transformer: Proportion Equivariant



Transformer: Sequence Modeling



Positional Encodings

APE

→ Encodes ***A***bsolute positions

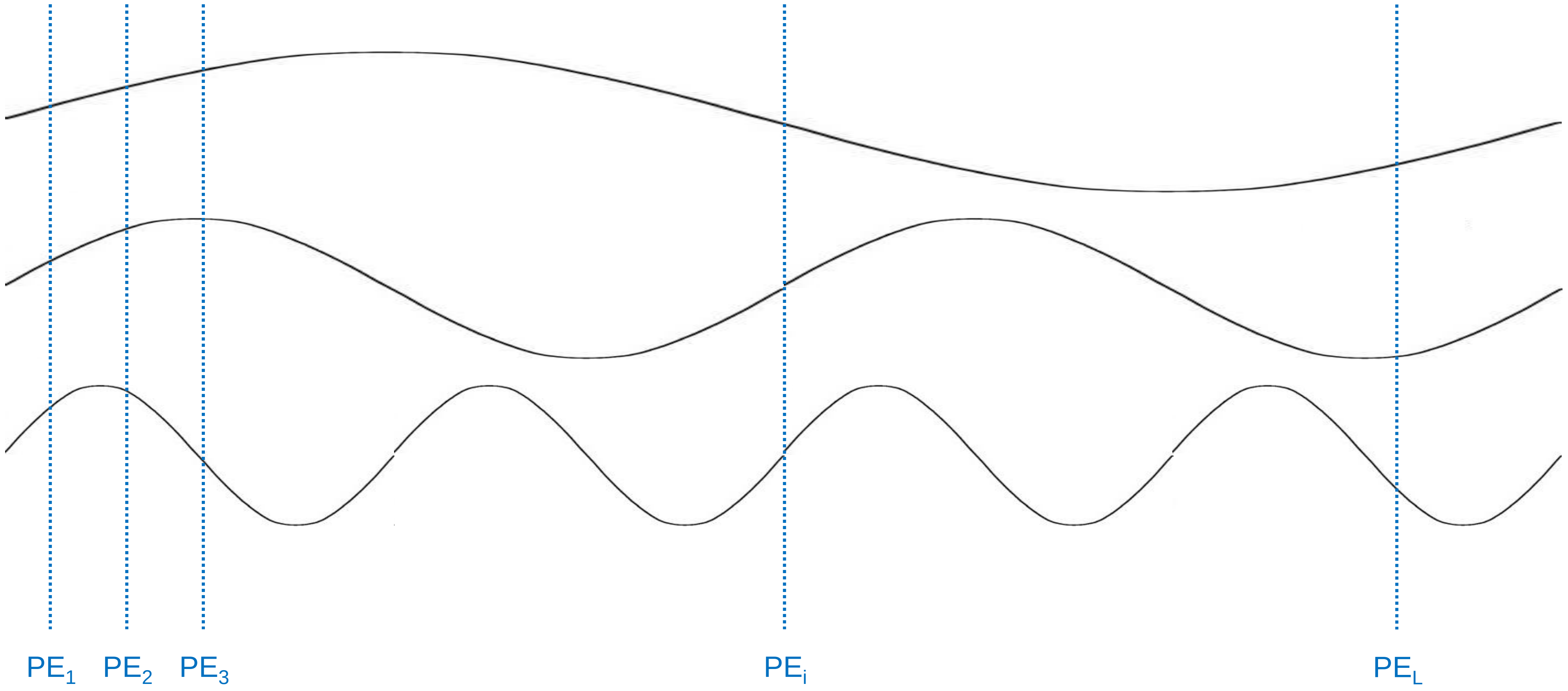
→ Adds to input embedding

RPE

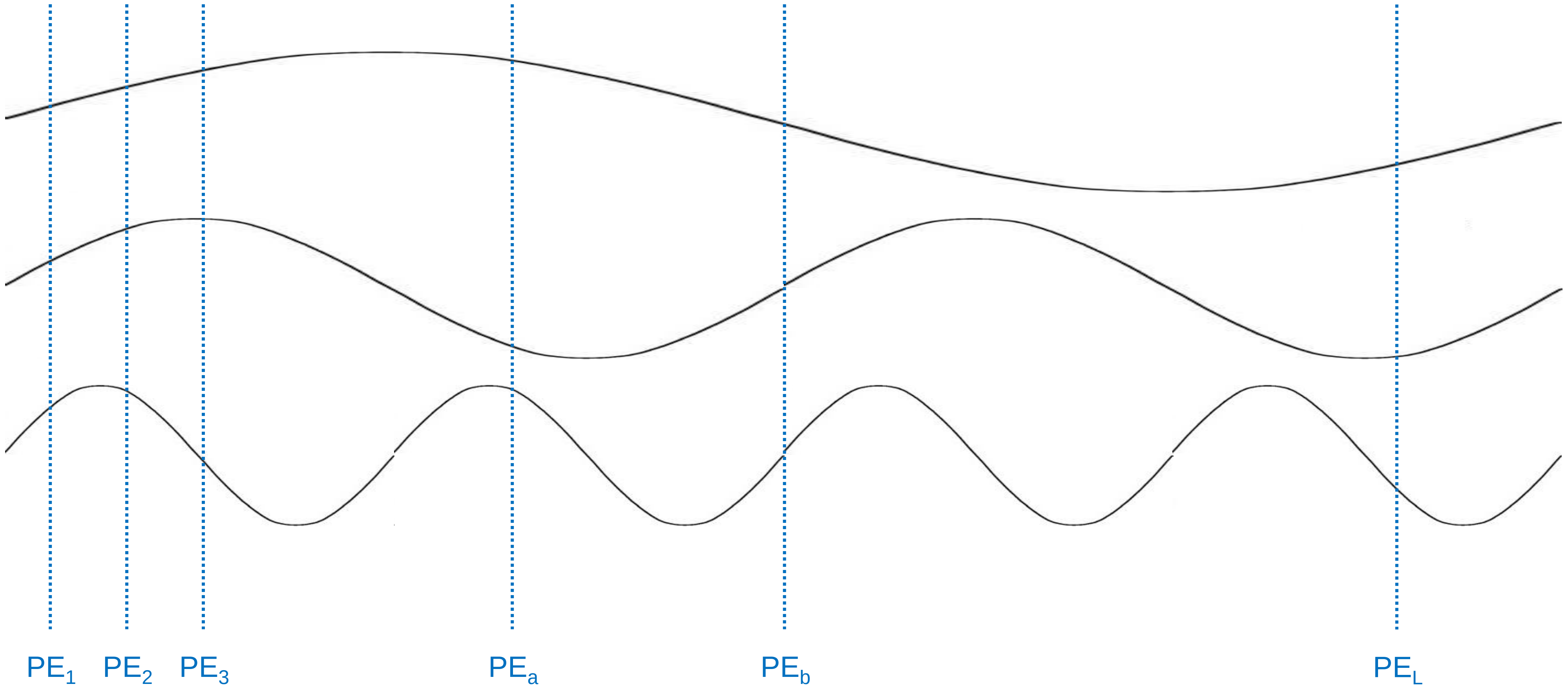
→ Encodes token-token ***R***elative distances

→ Modifies attention weights

Sine Wave APE



Sinusoidal APE: Sine & Cosine Waves



$$\mathrm{PE}_a \cdot \mathrm{PE}_b = \cos a \cos b + \sin a \sin b = \cos(a - b)$$

Additive RPE

Normal attention score between $\boldsymbol{q}, \boldsymbol{k}$

$$\mathrm{score}(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{q} \cdot \boldsymbol{k}$$

T5/Alibi attention score between $\boldsymbol{q}@a$ and $\boldsymbol{k}@b$

$$\mathrm{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \boldsymbol{q} \cdot \boldsymbol{k} + f(|a - b|)$$

f: a decreasing function

Rotary RPE (RoPE)

RoPE attention score between $\boldsymbol{q}@a$ and $\boldsymbol{k}@b$

$$\text{rotate}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$\text{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \text{rotate}(a)\boldsymbol{q} \cdot \text{rotate}(b)\boldsymbol{k} = \boldsymbol{q}^T \text{rotate}(a-b)\boldsymbol{k}$$

Positional Encodings

Sinusoidal APE[1]

→ $\text{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \left(\boldsymbol{q} + \begin{pmatrix}\cos a \\ \sin a\end{pmatrix}\right)^T \left(\boldsymbol{k} + \begin{pmatrix}\cos b \\ \sin b\end{pmatrix}\right) = \boldsymbol{q}^T\boldsymbol{k} + \cos(a-b) + \boldsymbol{q}^T \begin{pmatrix}\cos b \\ \sin b\end{pmatrix} + \begin{pmatrix}\cos a \\ \sin a\end{pmatrix}^T \boldsymbol{k}$

T5[2] / Alibi[3] additive RPE

→ $\text{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \boldsymbol{q}^T\boldsymbol{k} + f(|a-b|), \quad f: a \text{ decreasing function}$

RoPE[4]

→ $\text{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \boldsymbol{q}^T \begin{pmatrix}\cos(a-b) & -\sin(a-b) \\ \sin(a-b) & \cos(a-b)\end{pmatrix} \boldsymbol{k}$

[1] Attention Is All You Need.
https://doi.org/10.48550/arXiv.1706.03762

[2] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
https://doi.org/10.48550/arXiv.1910.10683

[3] Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation.
https://doi.org/10.48550/arXiv.2108.12409

[4] RoFormer: Enhanced Transformer with Rotary Position Embedding.
https://doi.org/10.48550/arXiv.2104.09864

Agenda

Positional Encodings

→ Transformer, absolute PE, relative PE

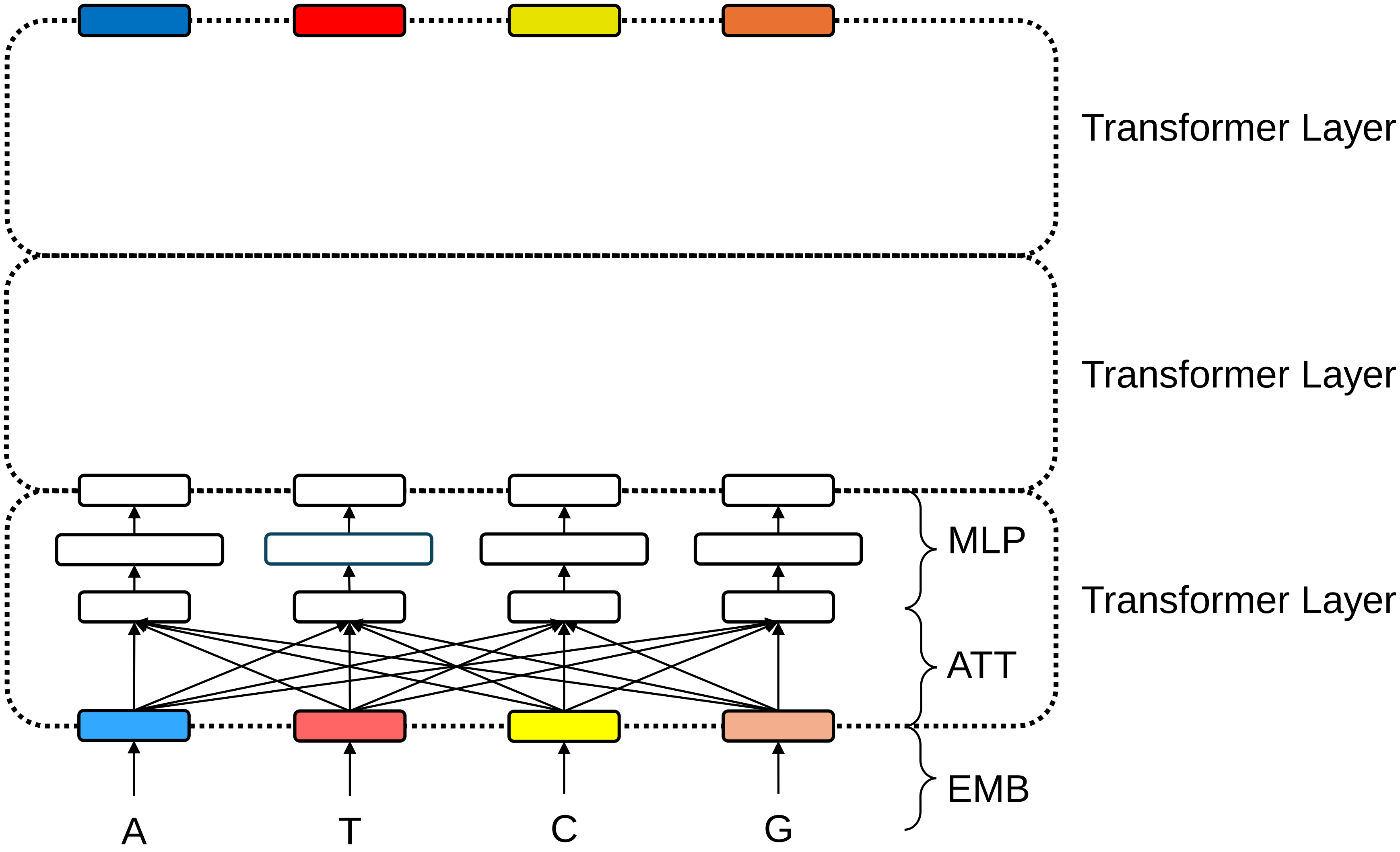
No Positional Encodings

→ Generative transformer, NoPE

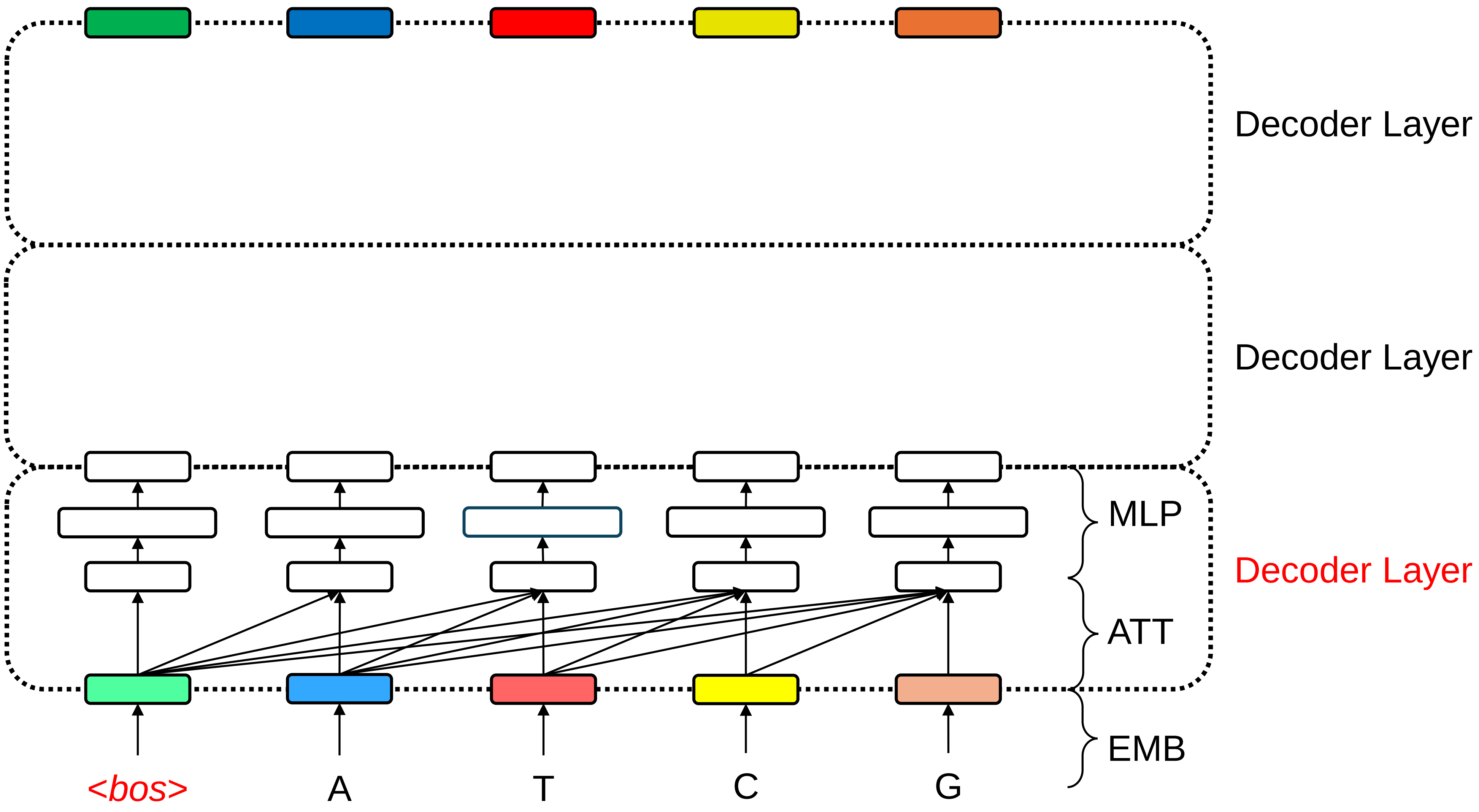
Length Generalization

→ Out-of-distribution, extrapolation, interpolation

Transformer



Generative Transformer



Generative Transformer

- Prepends a unique <*bos*> token to input sequence
- Only allows backward attention

Also called ***G***enerative ***P***re-***T***raining (GPT)

No Positional Encodings (NoPE/NoPos)

NoPE attention score between $\boldsymbol{q}@a$ and $\boldsymbol{k}@b$

$$\text{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \boldsymbol{q}^T \boldsymbol{k}$$

No Positional Encodings

NoPE attention score between $\boldsymbol{q}@a$ and $\boldsymbol{k}@b$

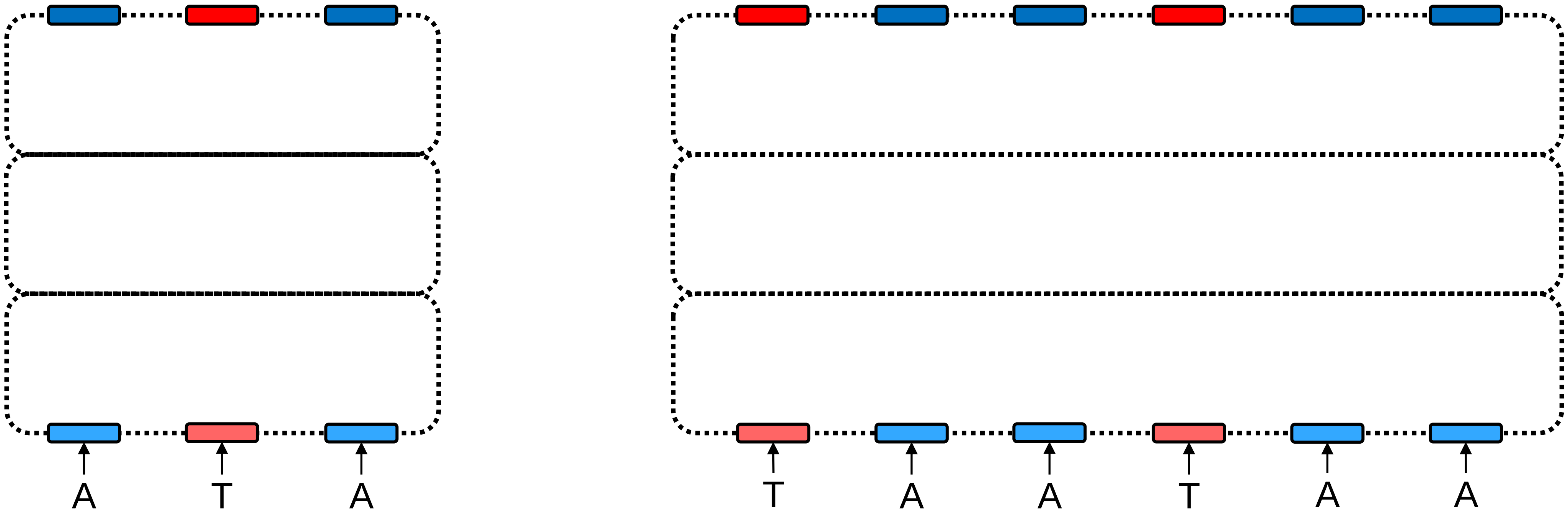
$$\text{score}(\boldsymbol{q}, \boldsymbol{k}, a, b) = \boldsymbol{q}^T \boldsymbol{k}$$

NoPE

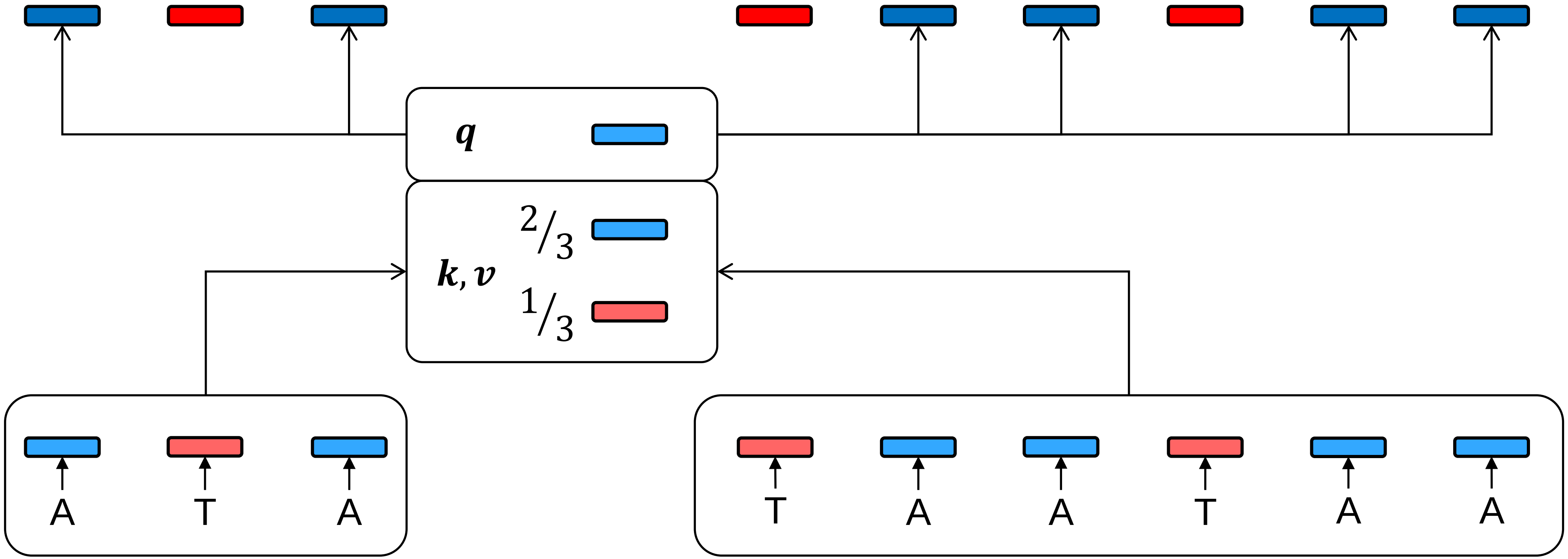
Theorem. Generative transformer with NoPE can encode both absolute and relative positions.

The Impact of Positional Encoding on Length Generalization in Transformers.
https://doi.org/10.48550/arXiv.2305.19466

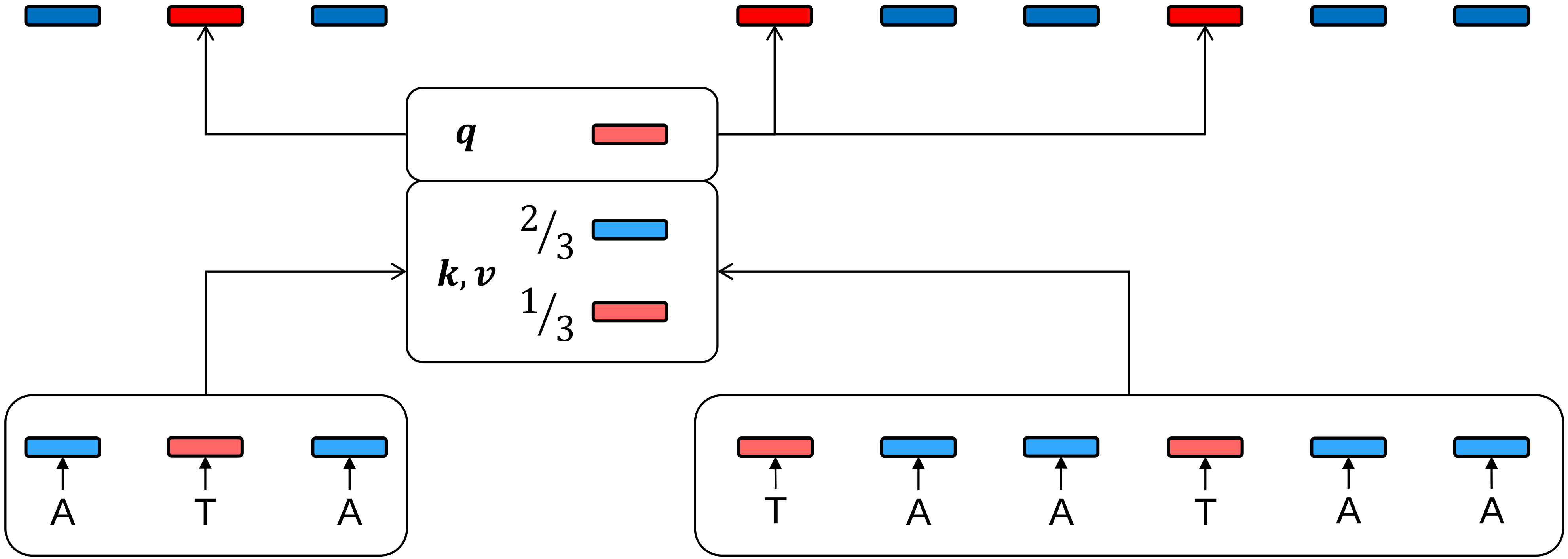
Transformer: Proportion Equivariant



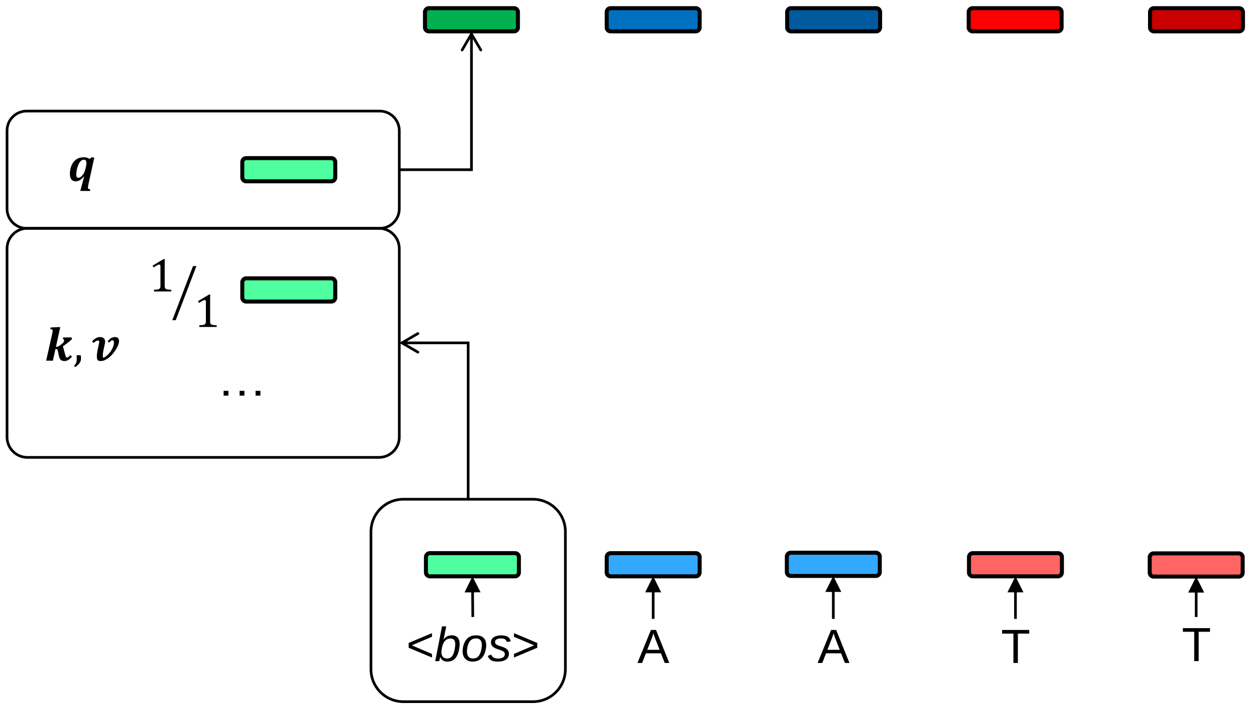
Transformer + NoPE



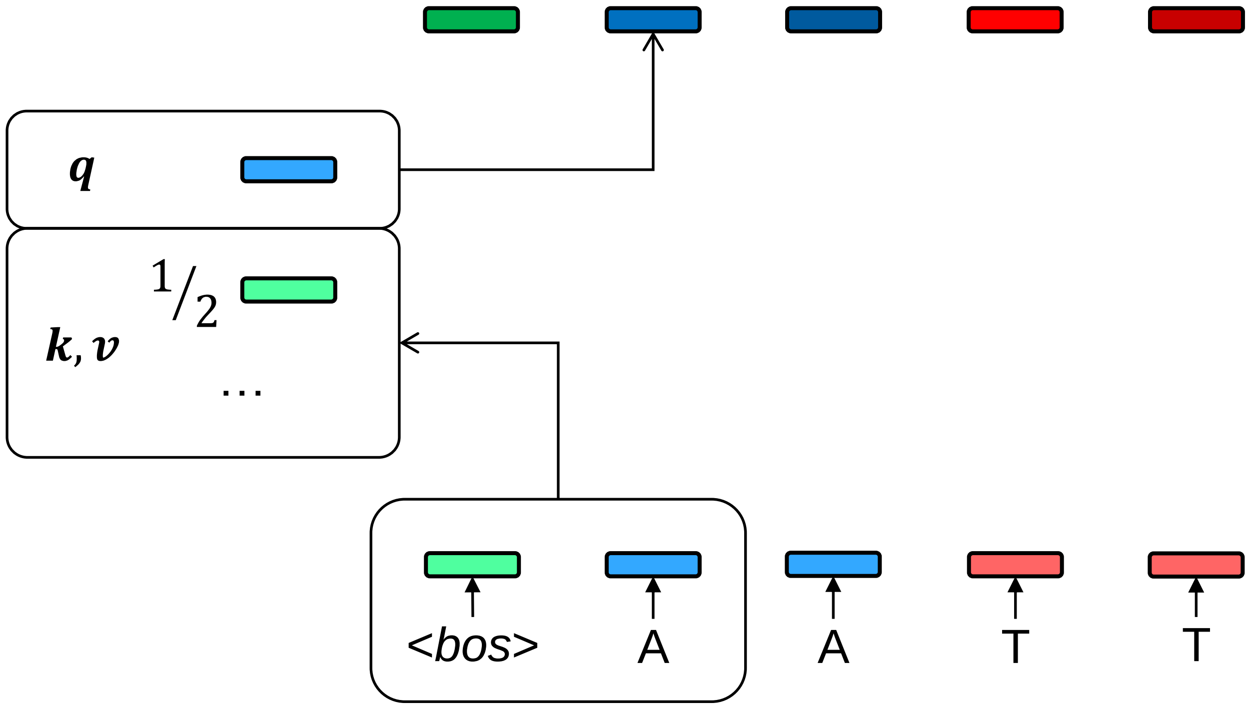
Transformer + NoPE



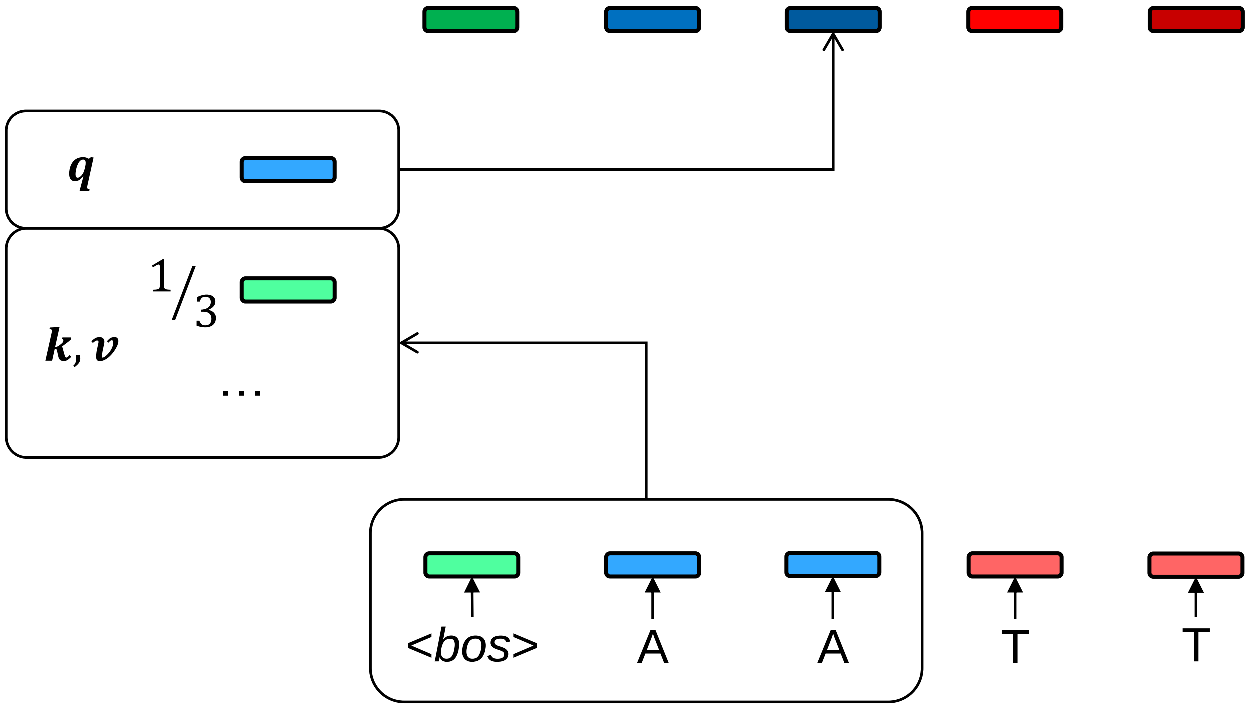
Generative Transformer + NoPE



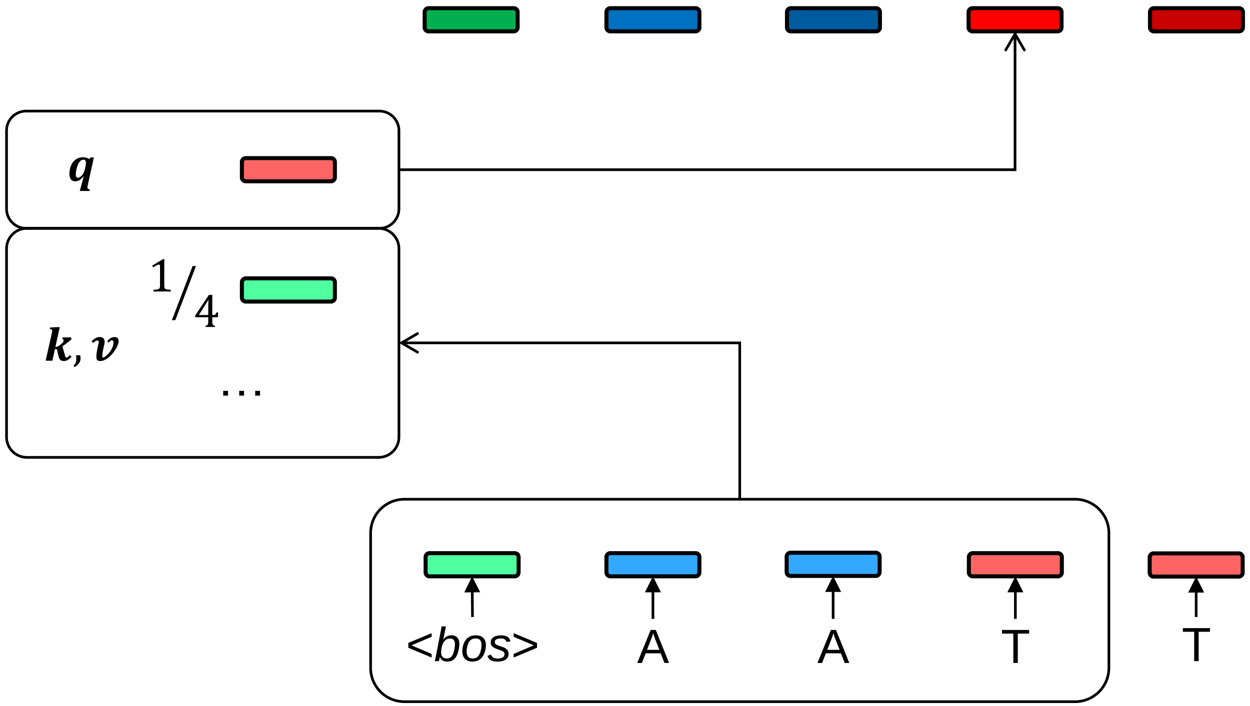
Generative Transformer + NoPE



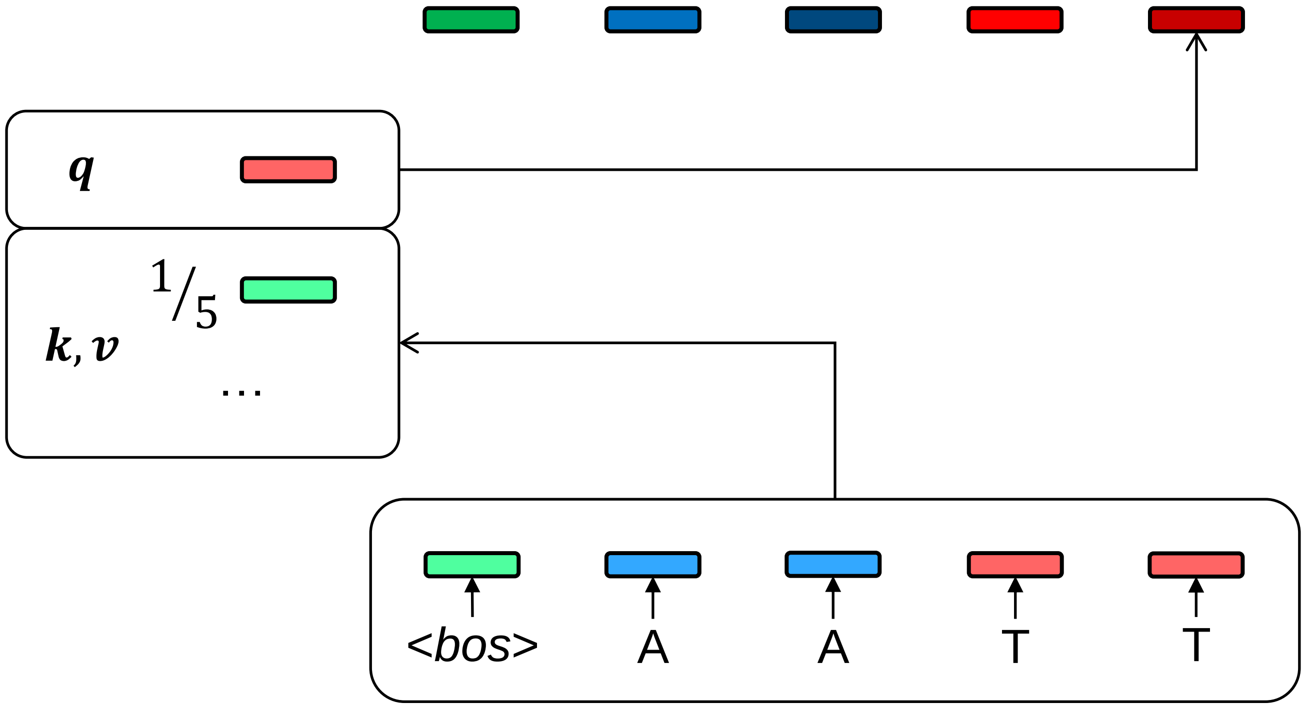
Generative Transformer + NoPE



Generative Transformer + NoPE



Generative Transformer + NoPE



In-Distribution Perplexity

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

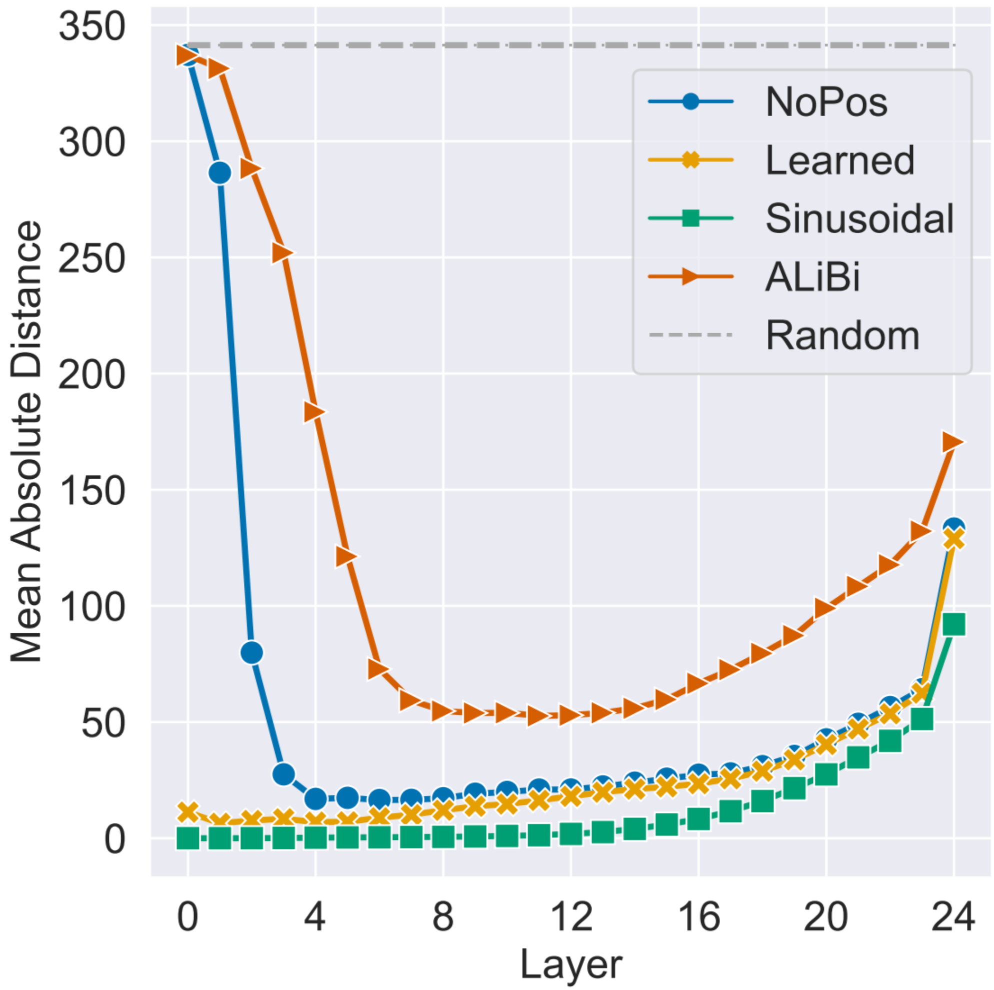
	WikiText-103	The Pile
NoPos	20.97	13.10
Learned	20.42	13.05
Sinusoidal	20.16	12.93
ALiBi	19.71	12.51

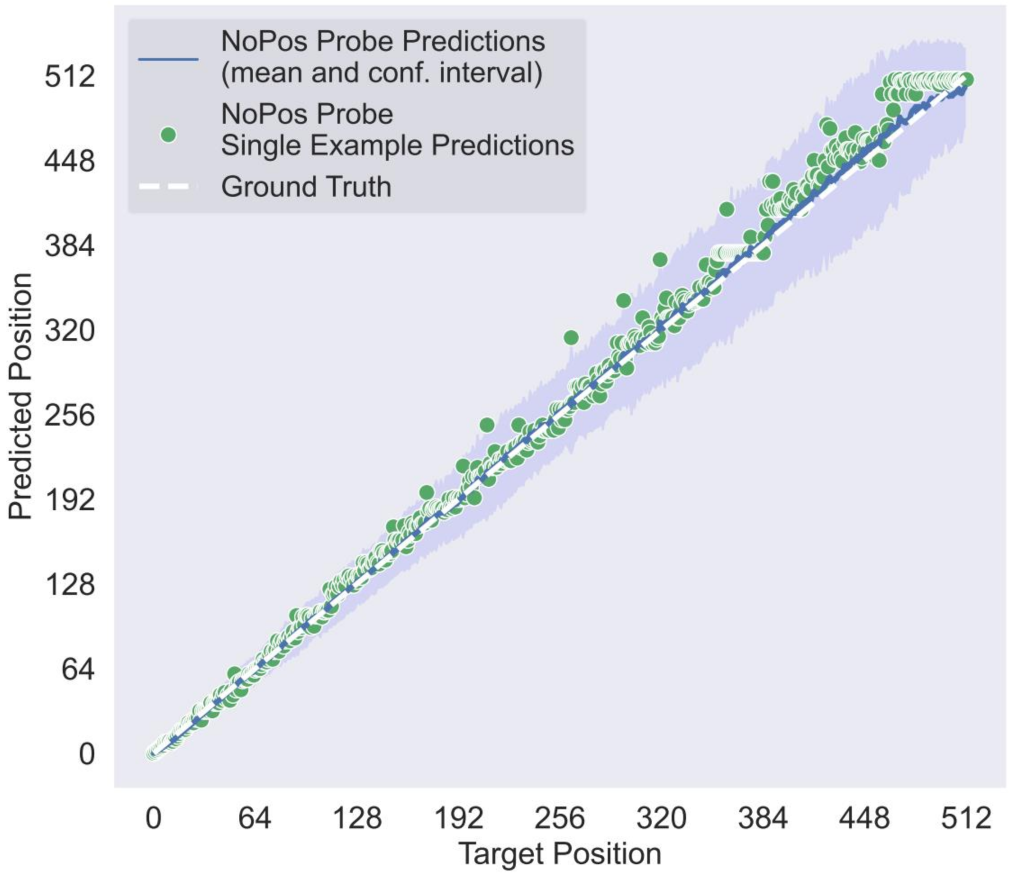
Model Size	125M	350M	760M	1.3B
NoPos	22.15	16.87	14.29	13.10
Learned	22.04	16.84	14.21	13.05
Sinusoidal	21.49	16.58	14.04	12.93
ALiBi	19.94	15.66	13.53	12.51

	MLM Perplexity
NoPos	147.18
Learned	4.06
Sinusoidal	4.07
ALiBi	4.00

Transformer Language Models without Positional Encodings Still Learn Positional Information.
https://doi.org/10.18653/v1/2022.findings-emnlp.99

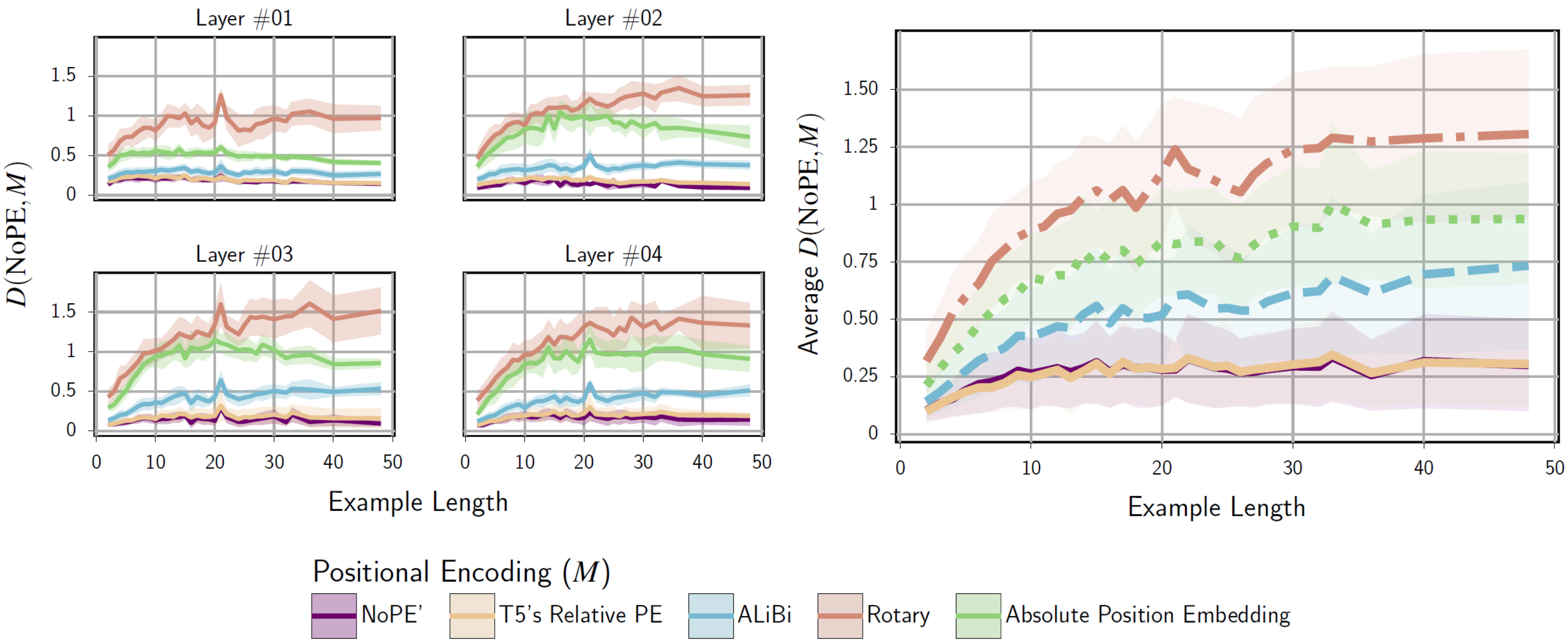
Absolute Position Inference





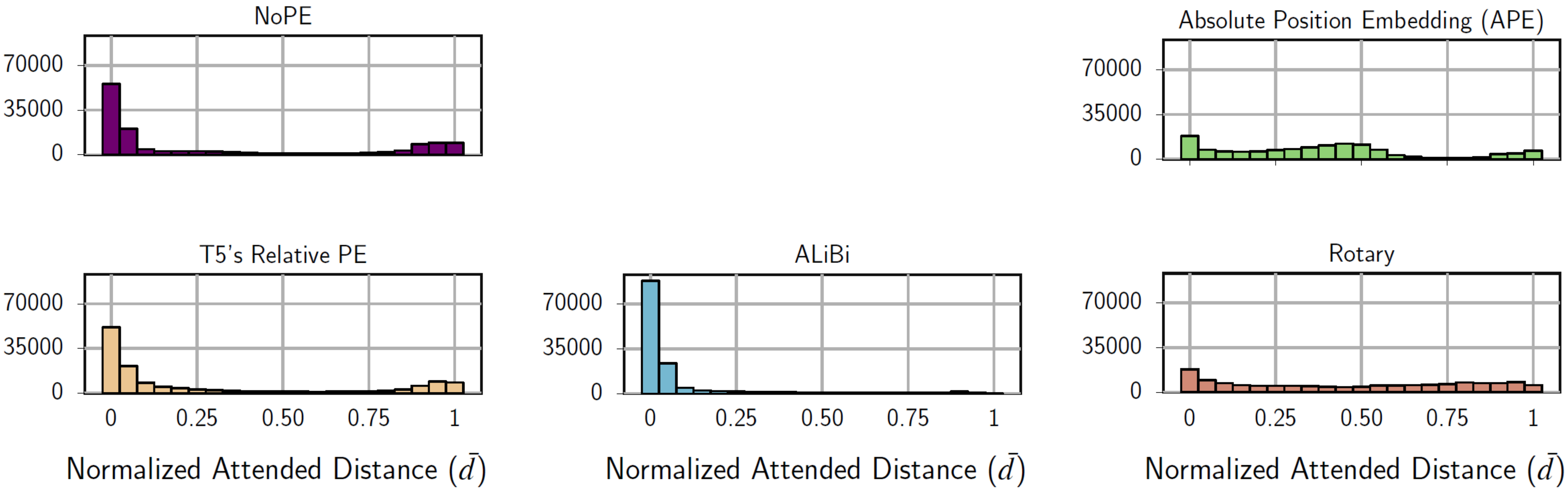
Transformer Language Models without Positional Encodings Still Learn Positional Information.
https://doi.org/10.18653/v1/2022.findings-emnlp.99

Attention Pattern Similarity



The Impact of Positional Encoding on Length Generalization in Transformers.
https://doi.org/10.48550/arXiv.2305.19466

Attention Distance Pattern



The Impact of Positional Encoding on Length Generalization in Transformers.
https://doi.org/10.48550/arXiv.2305.19466

Agenda

Positional Encodings

→ Transformer, absolute PE, relative PE

No Positional Encodings

→ Generative transformer, NoPE

Length Generalization

→ Out-of-distribution, extrapolation, interpolation

Sequence Lengths

L: max length that has sufficient training sequences

E.g., 3,072

L': max possible sequence length

E.g., 128,000

$L < l \leq L'$: out-of-distribution lengths → OOD positional encodings

Generalize to OOD lengths

- Negative reasons, L is limited in practice by

  Data sparsity

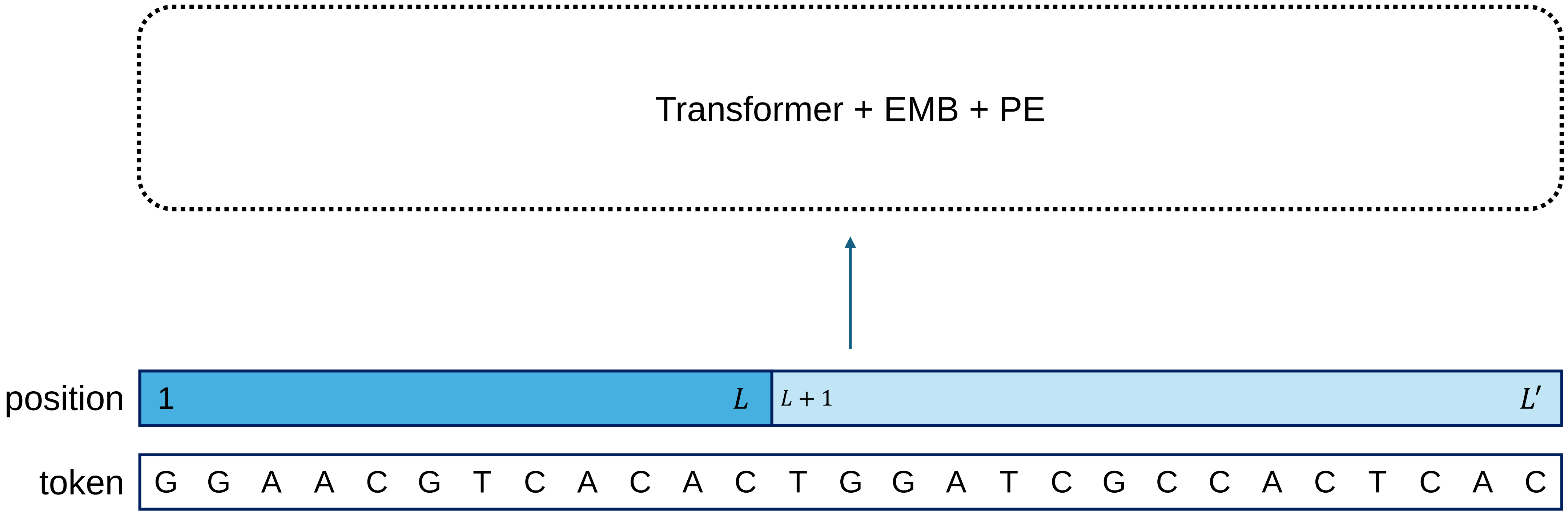
  Computation resources

- Positive reasons, large L' is often desirable for it enables

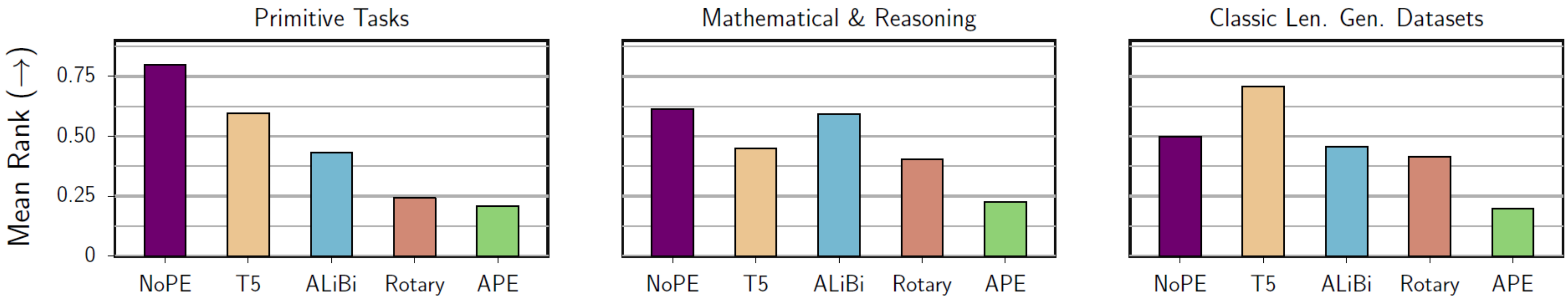
  Longer context, more complex instructions, more in-context examples

  Longer generation, more reasoning steps

Direct Extrapolation

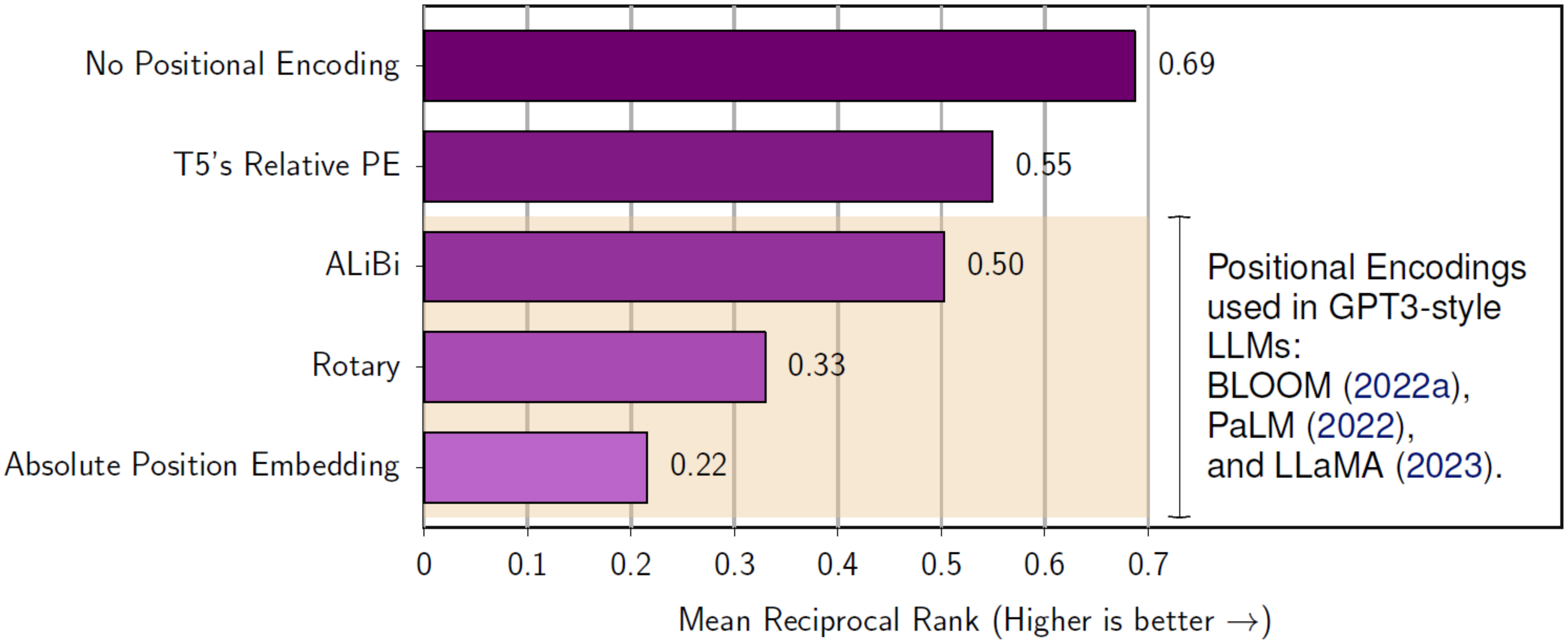


Direct Extrapolation Performance



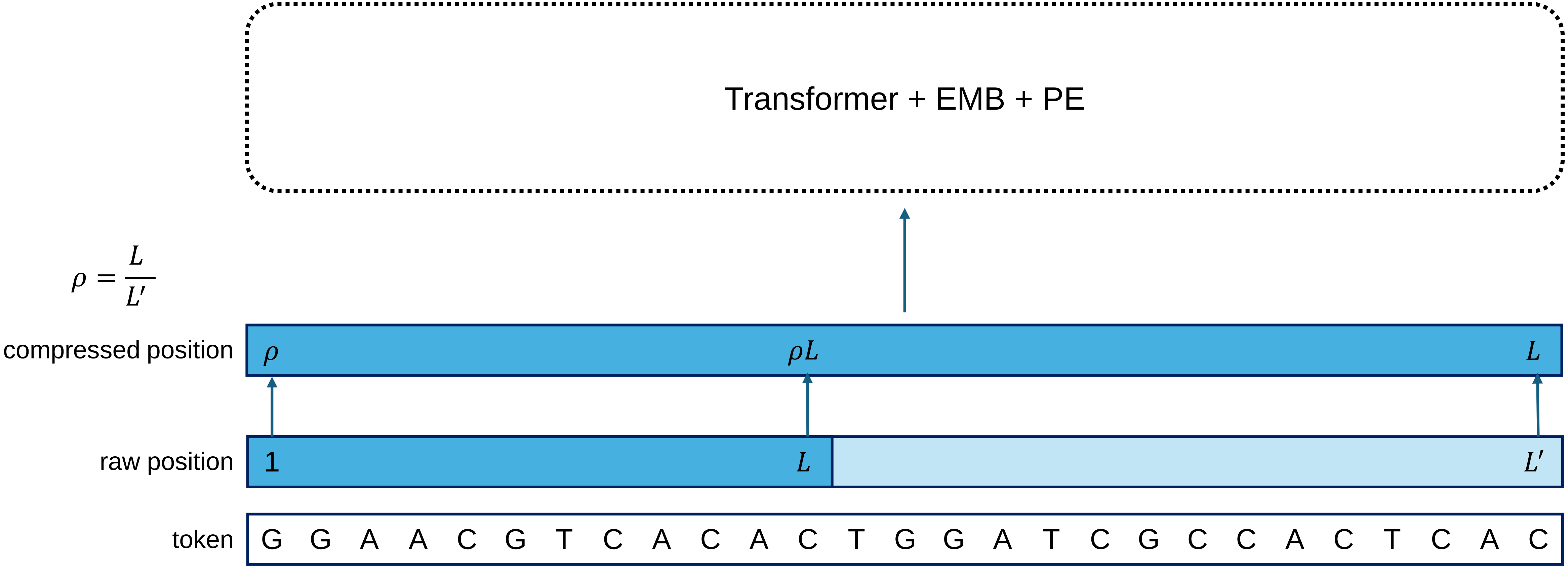
The Impact of Positional Encoding on Length Generalization in Transformers.
https://doi.org/10.48550/arXiv.2305.19466

Direct Extrapolation Performance

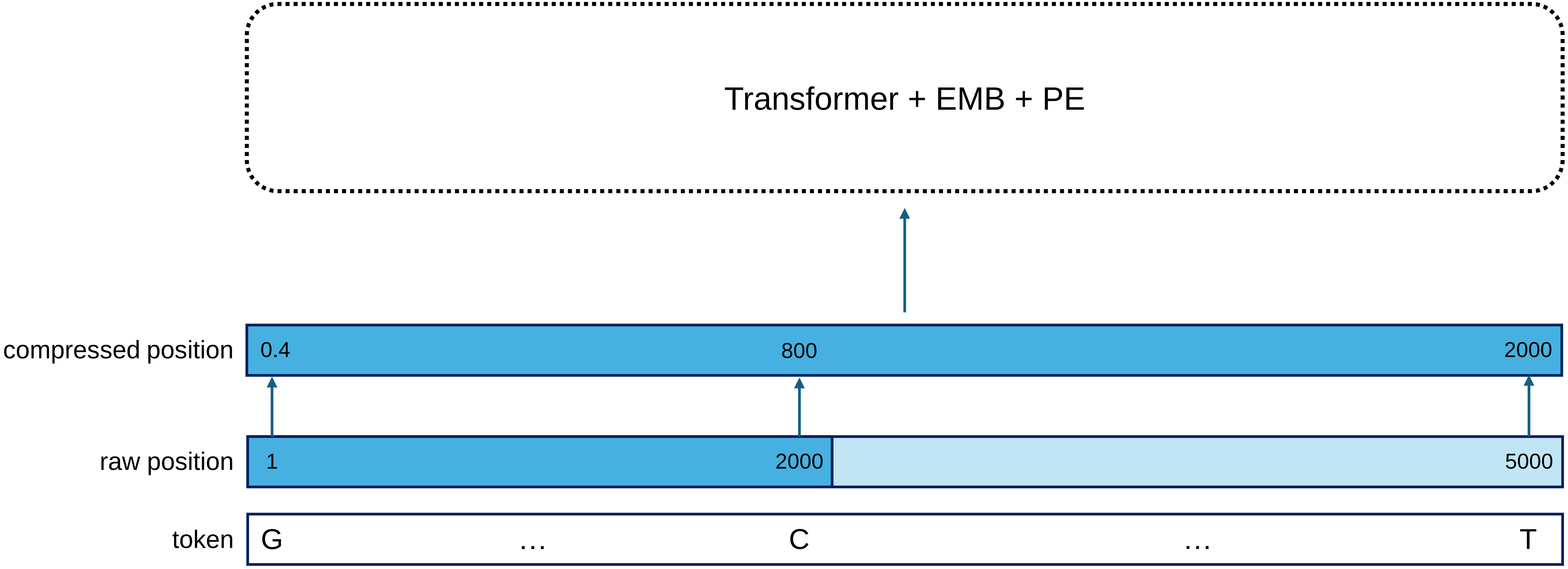


The Impact of Positional Encoding on Length Generalization in Transformers.
https://doi.org/10.48550/arXiv.2305.19466

Extrapolation by Interpolation



Extrapolation by Interpolation



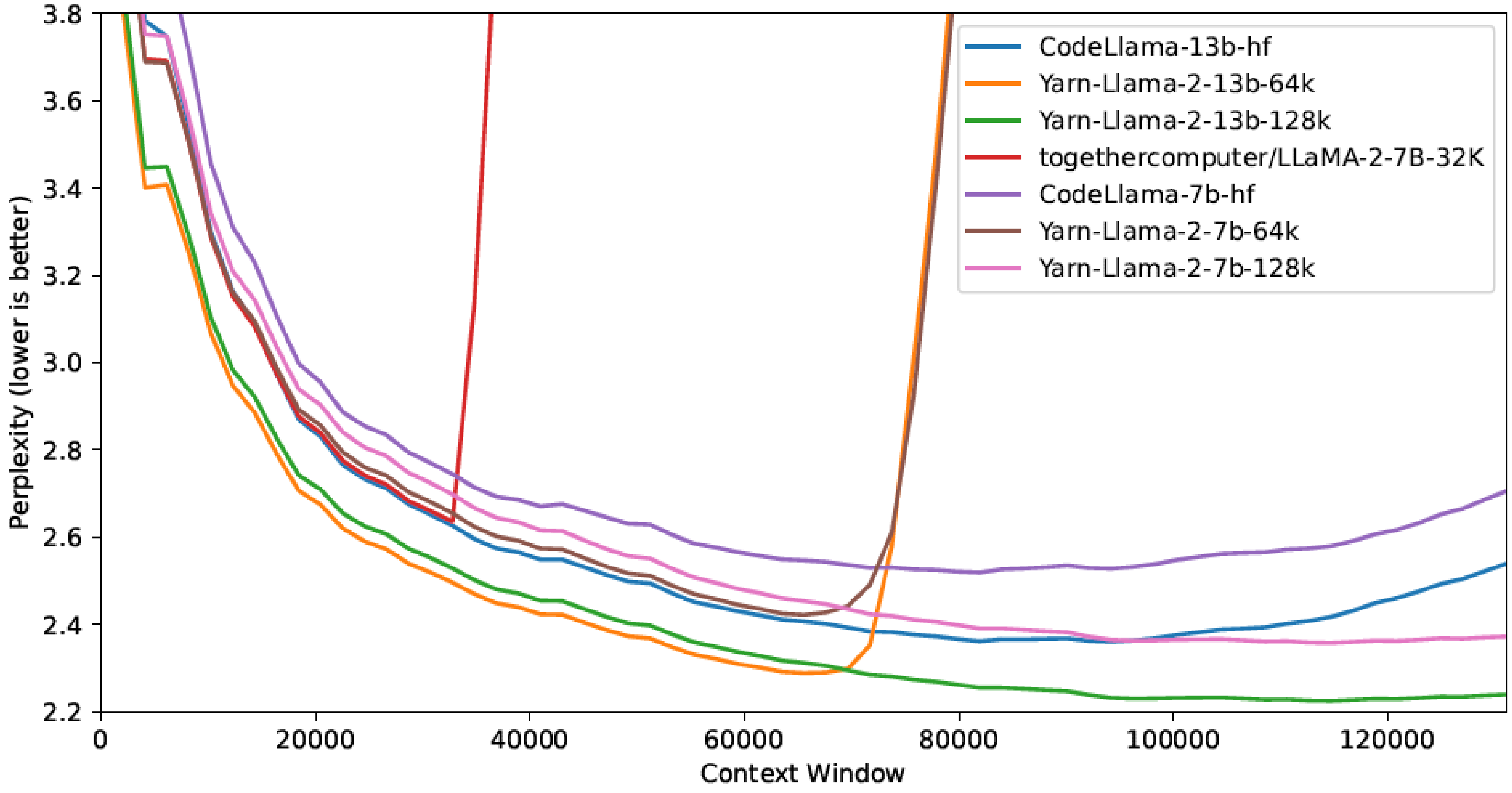
Interpolation for RoPE

YaRN*

- Scale rotation wavelengths
- Do not scale high frequency dimensions
- Change scale at each time step
- Finetuned on ~0.1% pretraining data size

*YaRN: Efficient Context Window Extension of Large Language Models.
https://doi.org/10.48550/arXiv.2309.00071

Interpolation for RoPE



*YaRN: Efficient Context Window Extension of Large Language Models.
https://doi.org/10.48550/arXiv.2309.00071

Agenda

Positional Encodings

→ Transformer, absolute PE, relative PE

No Positional Encodings

→ Generative transformer, NoPE

Length Generalization

→ Out-of-distribution, extrapolation, interpolation