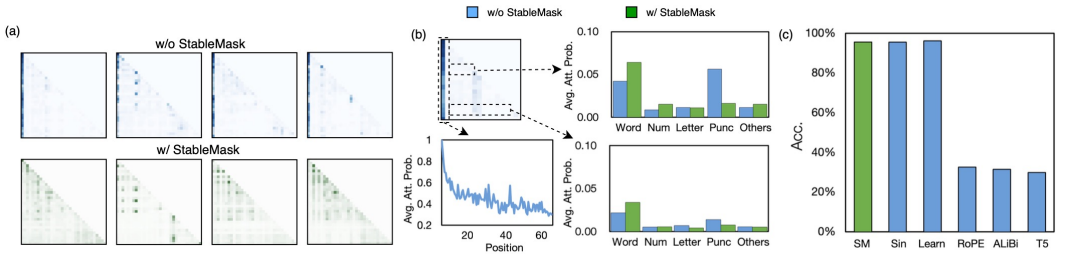
StableMask: Refining Causal Masking in Decoder-only Transformer

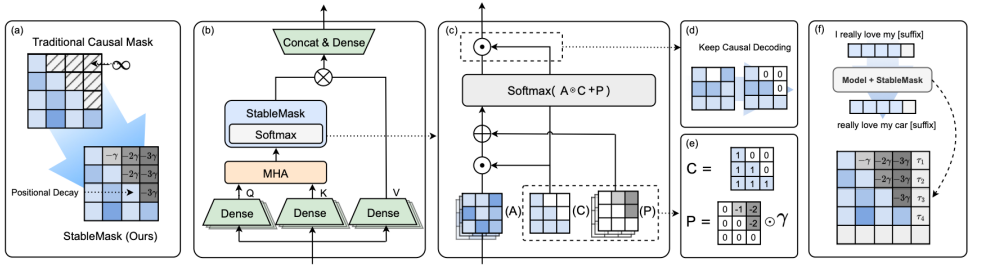
Qingyu Yin, Xuzheng He, Xiang Zhuang, Yu Zhao, Jianhua Yao, Xiaoyu Shen, Qiang Zhang
Zhejiang University, Peking University, Eastern Institute of Technology and Tencent AI



Two Issues in Decoder-only Transformers

Disproportional Attention: The softmax function used in self-attention requires all attention scores to be non-zero and sum up to 1. This often leads to an uneven distribution of attention across tokens, causing the model to allocate excessive attention to certain tokens like punctuation marks or initial tokens, which can degrade model performance and stability.

Inability to Encode Absolute Position: Relative Position Encoding (RPE), though beneficial for extrapolation and transformation invariance, struggles to encode absolute positional information. This limitation hampers the model's ability to perform tasks requiring precise position information.



One Stone, Two Bird: StableMask

$$p_{\text{base}} = 0, \quad p_{ij} = p_{\text{base}} - (j-1)\gamma$$

Pseudo-Attention Scores: The pseudo-attention scores are designed to decrease linearly along the sequence, adhering to the property of disproportionate attention in a decoder-only model.

$$\begin{aligned}\tilde{A} &= \text{Softmax}(A_{\text{SM}}) \odot C \\ &= \text{Softmax}(A \odot C + P) \odot C.\end{aligned}$$

Causal Mask Modification: The traditional causal mask is modified by incorporating pseudo-attention scores. This allows for the encoding of absolute positional information.

WikiText-103				MiniPile				
Model	*PE	#Params	PPL	Model	*PE	#Params	PPL 1 Epoch	PPL 2 Epoch
BLOOM	ALiBi	71M	$29.9_{\pm.1}$	BLOOM	ALiBi	160M	$25.8_{\pm.2}$	$23.3_{\pm.4}$
BLOOM-SM	ALiBi	71M	$\mathbf{29.0}_{\pm.1}$	BLOOM-SM	ALiBi	160M	$\mathbf{25.6}_{\pm.0}$	$\mathbf{22.9}_{\pm.2}$
OpenLLaMA	RoPE	71M	$27.4_{\pm.2}$	OpenLLaMA	RoPE	160M	$25.9_{\pm.1}$	$21.2_{\pm.1}$
OpenLLaMA-SM	RoPE	71M	$\mathbf{26.9}_{\pm.3}$	OpenLLaMA-SM	RoPE	160M	$\mathbf{25.0}_{\pm.0}$	$\mathbf{20.9}_{\pm.3}$
BLOOM	ALiBi	160M	$27.6_{\pm.9}$	BLOOM	ALiBi	430M	$20.6_{\pm.1}$	$15.6_{\pm.4}$
BLOOM-SM	ALiBi	160M	$\mathbf{26.1}_{\pm.2}$	BLOOM-SM	ALiBi	430M	$\mathbf{19.6}_{\pm.3}$	$\mathbf{15.5}_{\pm.2}$
OpenLLaMA	RoPE	160M	$22.5_{\pm.8}$	OpenLLaMA	RoPE	430M	$19.6_{\pm.2}$	$15.7_{\pm.5}$
OpenLLaMA-SM	RoPE	160M	$\mathbf{21.1}_{\pm.6}$	OpenLLaMA-SM	RoPE	430M	$\mathbf{19.5}_{\pm.4}$	$\mathbf{15.1}_{\pm.5}$

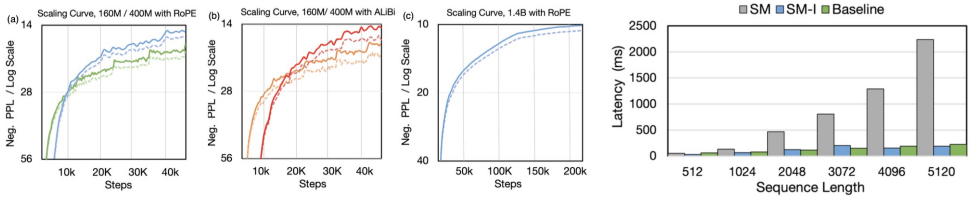
**: positional encoding type*

Methods	PPL
Baseline	22.5
Learnable AT	21.6
Fixed Value AT	22.4
StableMask	**21.1**

Pseudo Value	PPL
$-\infty$	22.5
0	21.5
1×10^{-2}	22.2
Positional Decay	**21.1**

Experiments

Addressing Two Issues: We calculated attention probability ratios for the first token and various token types. The results showed that StableMask significantly reduces abnormal attention distribution, particularly for initial tokens and punctuation marks. Models with StableMask displayed a significant reduction in attention values for these tokens, confirming that StableMask effectively mitigates the issue of disproportionate attention. Also, performance of StableMask was compared with various Position Encoding approaches, including RPE methods like ALiBi and RoPE. StableMask demonstrated superior capability in encoding absolute positional information, effectively addressing the limitations of RPE



Performance on Various Architectures and Tasks: Models with StableMask consistently achieved better perplexity (PPL) scores across different architectures and sizes, showing improved language modeling performance.On large-scale datasets like The Pile, a 1.4B parameter model with StableMask outperformed standard models, demonstrating better scaling with the number of tokens.

Pre-trained models with StableMask showed improved performance on downstream tasks such as LAMBADA, PIQA, ARC-Easy, ARC-Challenge, OpenbookQA, and Winogrande. This suggests that StableMask enhances both pretraining language understanding and downstream task effectiveness.