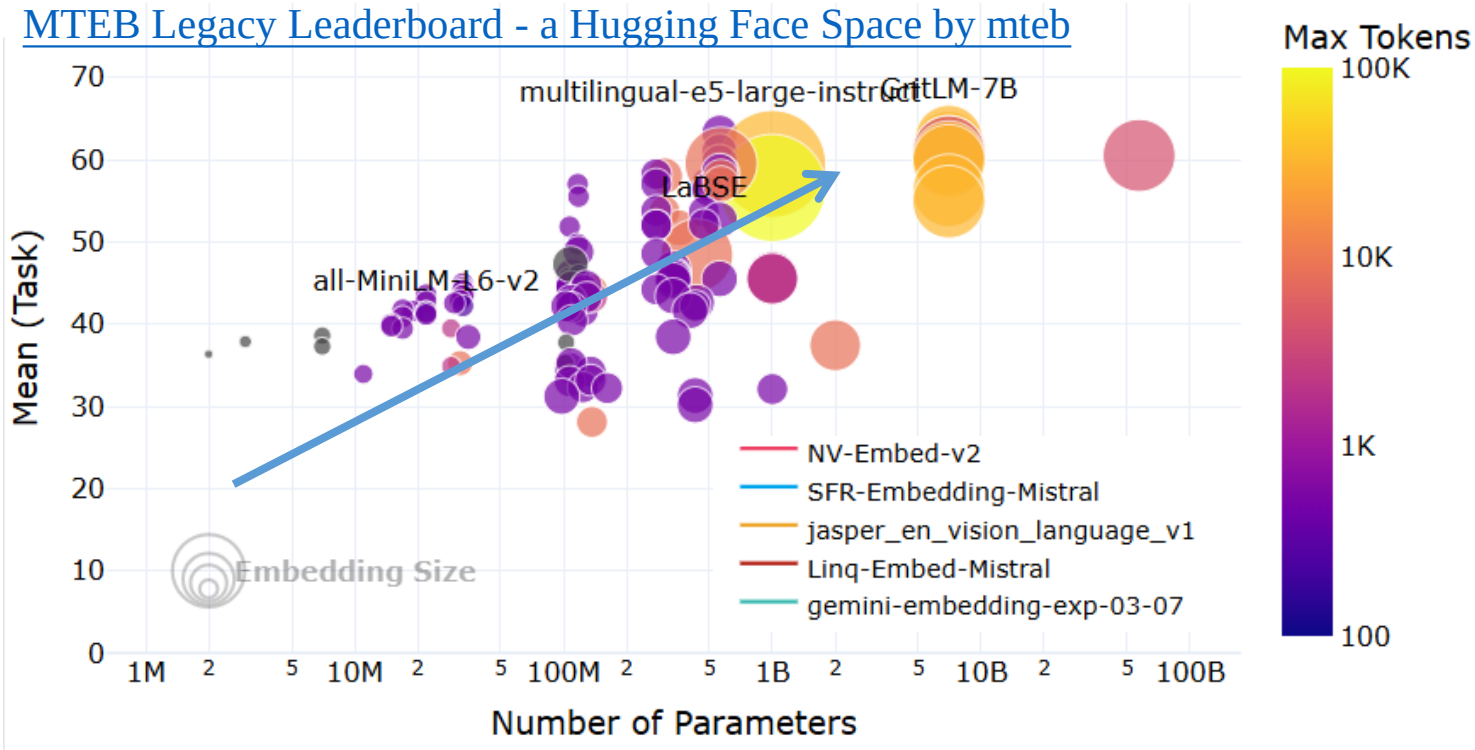
Look Both Ways and No Sink: Converting LLMs into Text Encoders without Training

Ziyong Lin[*1,4], **Haoyi Wu**[*2,3], **Shu Wang**[5], **Kewei Tu**[†2,3], **Zilong Zheng**[†1], **Zixia Jia**[†1]
[1]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
[2]School of Information Science and Technology, ShanghaiTech University
[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[4]Tsinghua University [5]University of California, Los Angeles

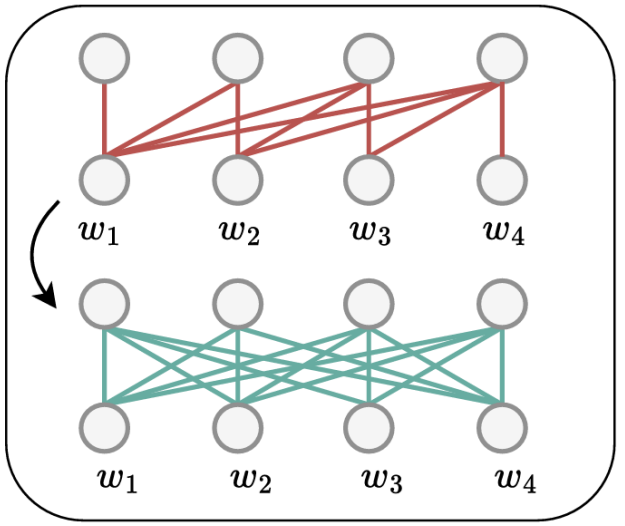
Background

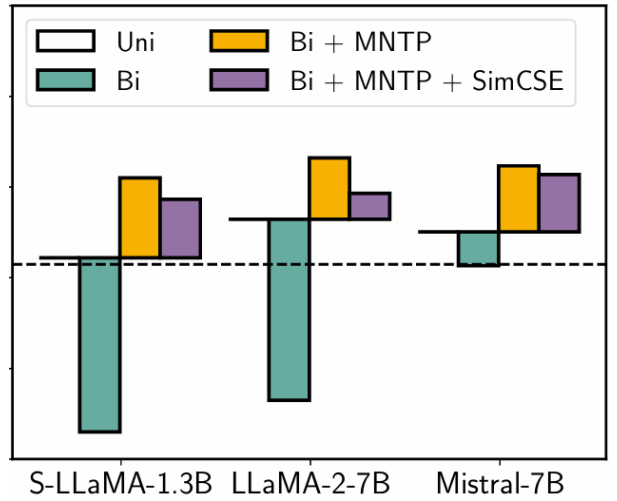
- **Larger Models, Better Performance (MTEB)**
- **Converting Unidirectional Models to Bidirectional Models**

MTEB Legacy Leaderboard - a Hugging Face Space by mteb



Enabling Bidirectional Attention





LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders, Parishad et al.

Motivation

4.1 Public Retrieval Datasets

We adopt the retrieval datasets as follows: MSMARCO (Bajaj et al., 2016), HotpotQA (Yang et al., 2018), Natural Question (Kwiatkowski et al., 2019), PAQ (Lewis et al., 2021), Stack Exchange (Stack-Exchange-Community, 2023), Natural Language Inference (Group et al., 2022), SQuAD (Rajpurkar et al., 2016), ArguAna (Wachsmuth et al., 2018), BioASQ (Tsatsaronis et al., 2015), FiQA (Maia et al., 2018), FEVER (Thorne et al., 2018), HoVer (Jiang et al., 2020), SciFact (Wadden et al., 2022), NFCorpus, MIRACL (Zhang et al., 2023) and Mr.TyDi (Zhang et al., 2021).

For semantic textual similarity datasets, we use the training splits of three semantic similarity datasets STS12 (Agirre et al., 2012), STS22 (Chen et al., 2022), STS-Benchmark (Cer et al., 2017) from MTEB Huggingface datasets. For any pair of texts with associated relevance scores $(t_a, t_b, score)$, we create two examples $(q^+ = t_a, d^+ = t_b)$ and $(q^+ = t_b, d^+ = t_a)$ if $score \geq 4$. We mine the hard negatives d_k^- from the pool of other texts using the same technique as section 4.1.1. Task instructions are appended to d^+, d^- since they are symmmetric with the query.

For classification, we utilize the English training splits of various datasets from MTEB Huggingface datasets (Muennighoff et al., 2022; Lhoest et al., 2021). The classification datasets that we use are as follows: AmazonReviews (McAuley & Leskovec, 2013a), AmazonCounterfactual (O'Neill et al., 2021), Banking77 (Casanueva et al., 2020), Emotion (Saravia et al., 2018), IMDB (Maas et al., 2011), MTOPDomain/MTOPIntent (Li et al., 2021), ToxicConversations (Adams et al., 2019), TweetSentimentExtraction (Maggie, 2020), AmazonPolarity (McAuley & Leskovec, 2013b), MassiveScenario/MassiveIntent (FitzGerald et al., 2022). For the Emotion and AmazonCounterfactual classification datasets we use BM25 (Robertson et al., 2009) similarity thresholds to filter out training data that is similar to the MTEB evaluation set.

For clustering datasets, we utilize the raw_arxiv, raw_biorxiv and raw_medrxiv datasets from MTEB Huggingface datasets, TwentyNewsgroups (Lang, 1995), Reddit (Geigle et al., 2021), StackExchange (Geigle et al., 2021), RedditP2P (Reimers, 2021b) and StackExchangeP2P (Reimers, 2021a) We filter out any training data that match the MTEB evaluation set.

- **Previous works require a significant amount of computational resources and time.**

GTE-Qwen2 -7B and NV-embed V2, involved first converting the model into a fully bidirectional model and then conducting additional training on a large dataset.

NV-EMBED: IMPROVED TECHNIQUES FOR TRAINING LLMS AS GENERALIST EMBEDDING MODELS, Chankyu Lee et al

Motivation

- In low-resource, domain-specific scenarios, general representation models often fail to perform optimally due to the scarcity of domain data and the presence of domain gaps.
  - Investigate the impact of different strategies to convert a pretrained transformer decoder into an encoder model.
  - Novel training-free conversion approach that significantly improves the performance of pretrained decoders in various (domain-specific) domains.

Architecture

For each attention layer in the Transformer, the attention score computation is given by:

$$\text{AttnLLM}_i(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + M\right)V$$

where AttnLLM_i is the i-th head of multi-head self-attention (Vaswani et al., 2017) in the LLM.

The query Q, key K, and value V are computed as:

$$Q = W_q x + b, \quad K = W_k x + b, \quad V = W_v x + b$$

Here, M represents the mask. For the forward layer with causal attention mask:

$$M_{FWD} = \begin{bmatrix} 0 & -\infty & -\infty & \dots & -\infty \\ 0 & 0 & -\infty & \dots & -\infty \\ 0 & 0 & 0 & \dots & -\infty \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

For the backward layer, where each token can only see itself and subsequent tokens:

$$M_{BWD} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ -\infty & 0 & 0 & \dots & 0 \\ -\infty & -\infty & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\infty & -\infty & -\infty & \dots & 0 \end{bmatrix}$$

For the bidirectional layer, M is a zero matrix.

Architecture

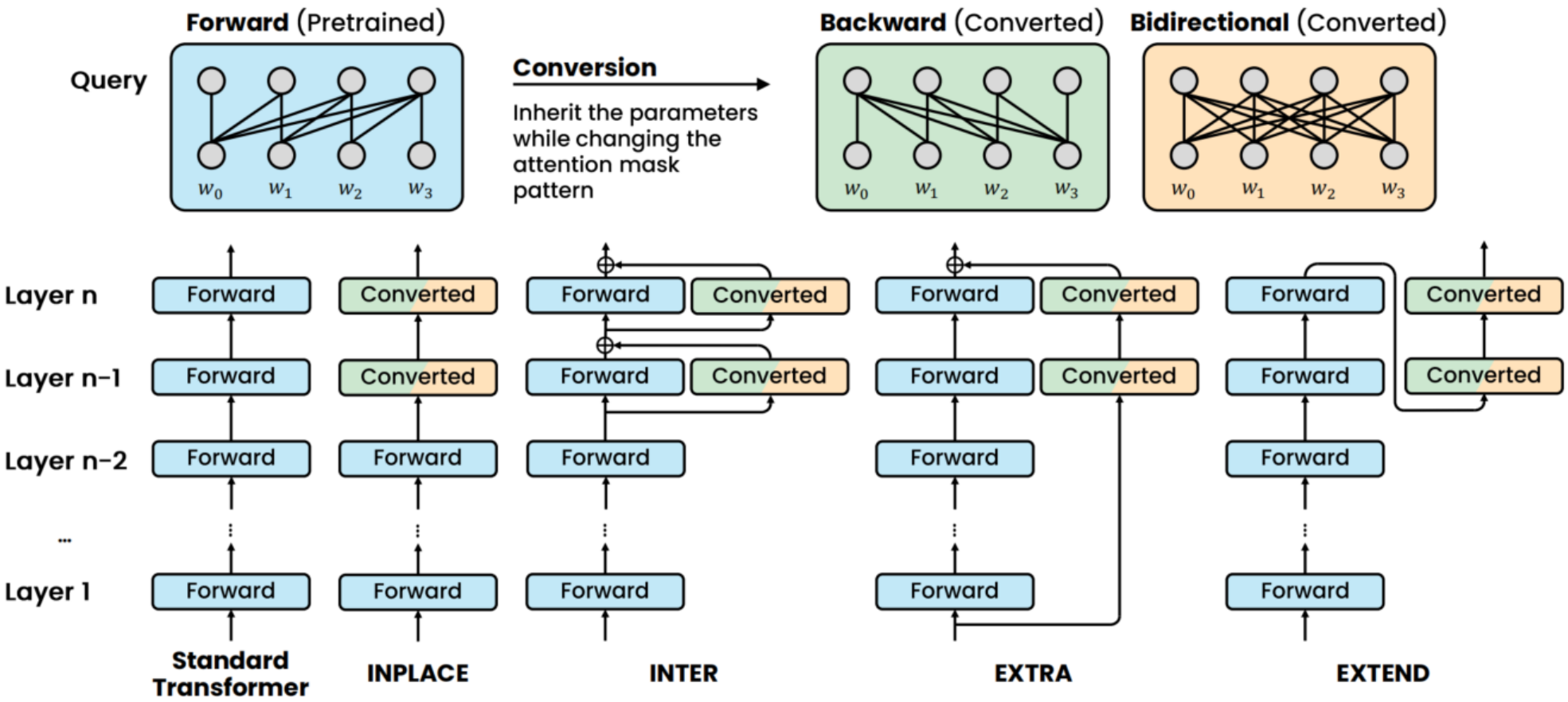
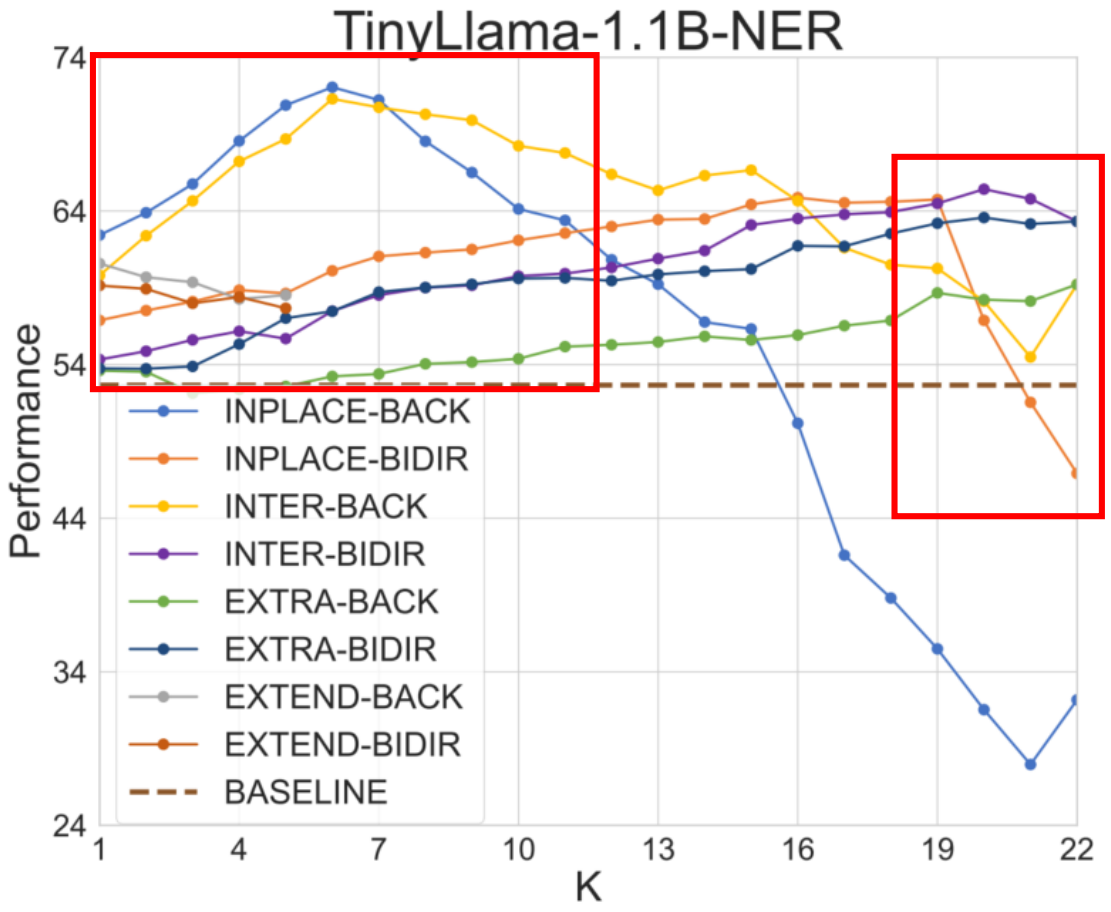
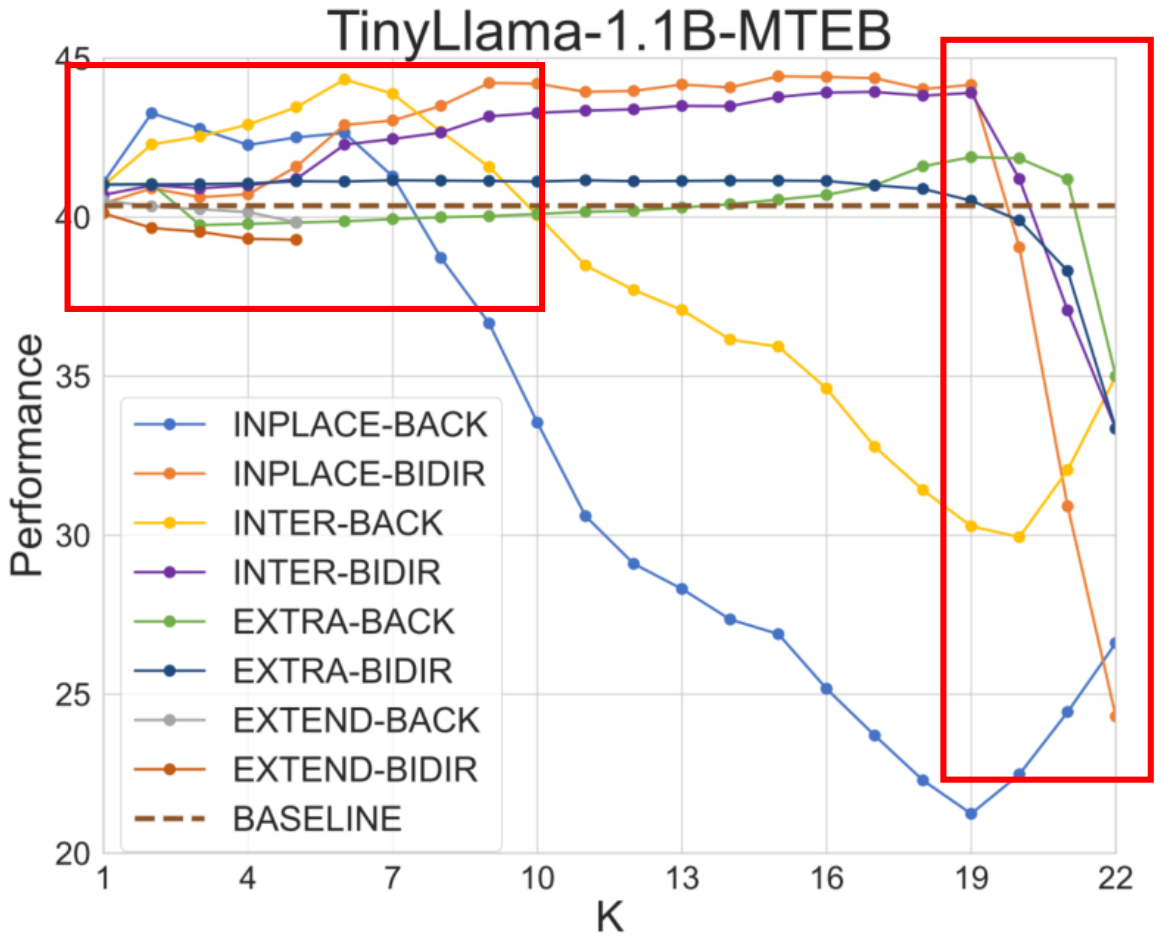


Figure 1: The schematic diagrams of four model architectures (number of converted layers $k = 2$). Either converted backward attention or bidirectional attention is incorporated into all the architectures.

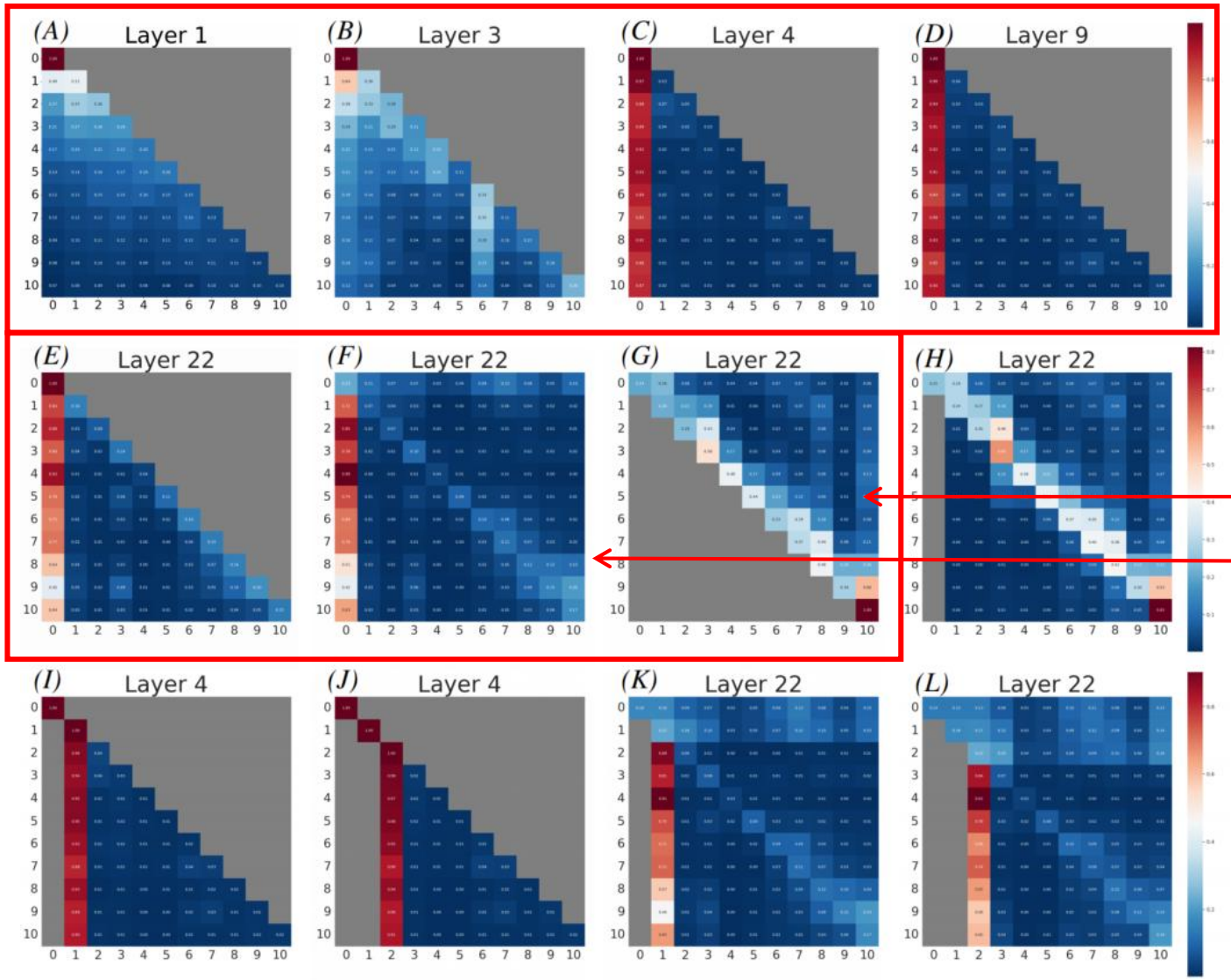
Empirical Study

- INPLACE and INTER achieve better performance;
- Maintaining some layers as FWD layers can be beneficial;
- BACK surpasses BIDIR when k is small.

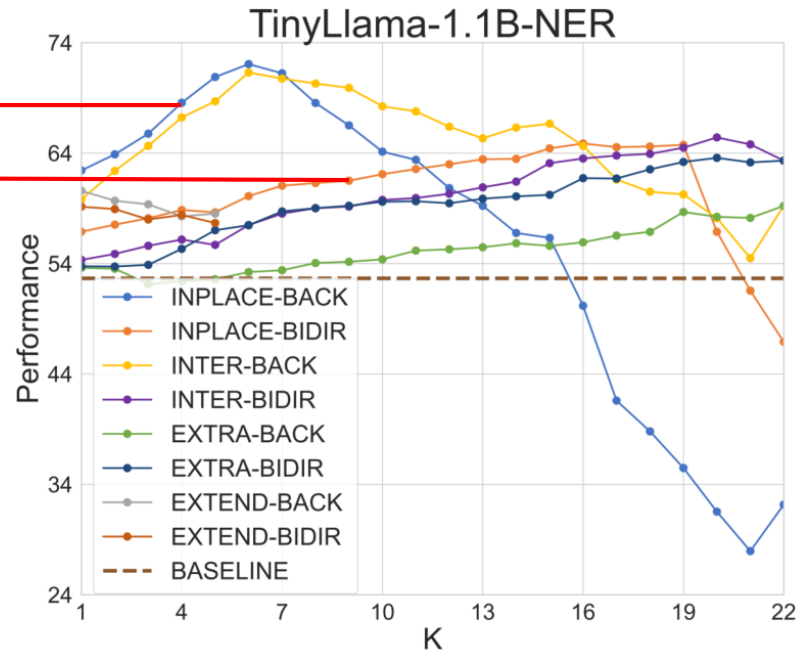




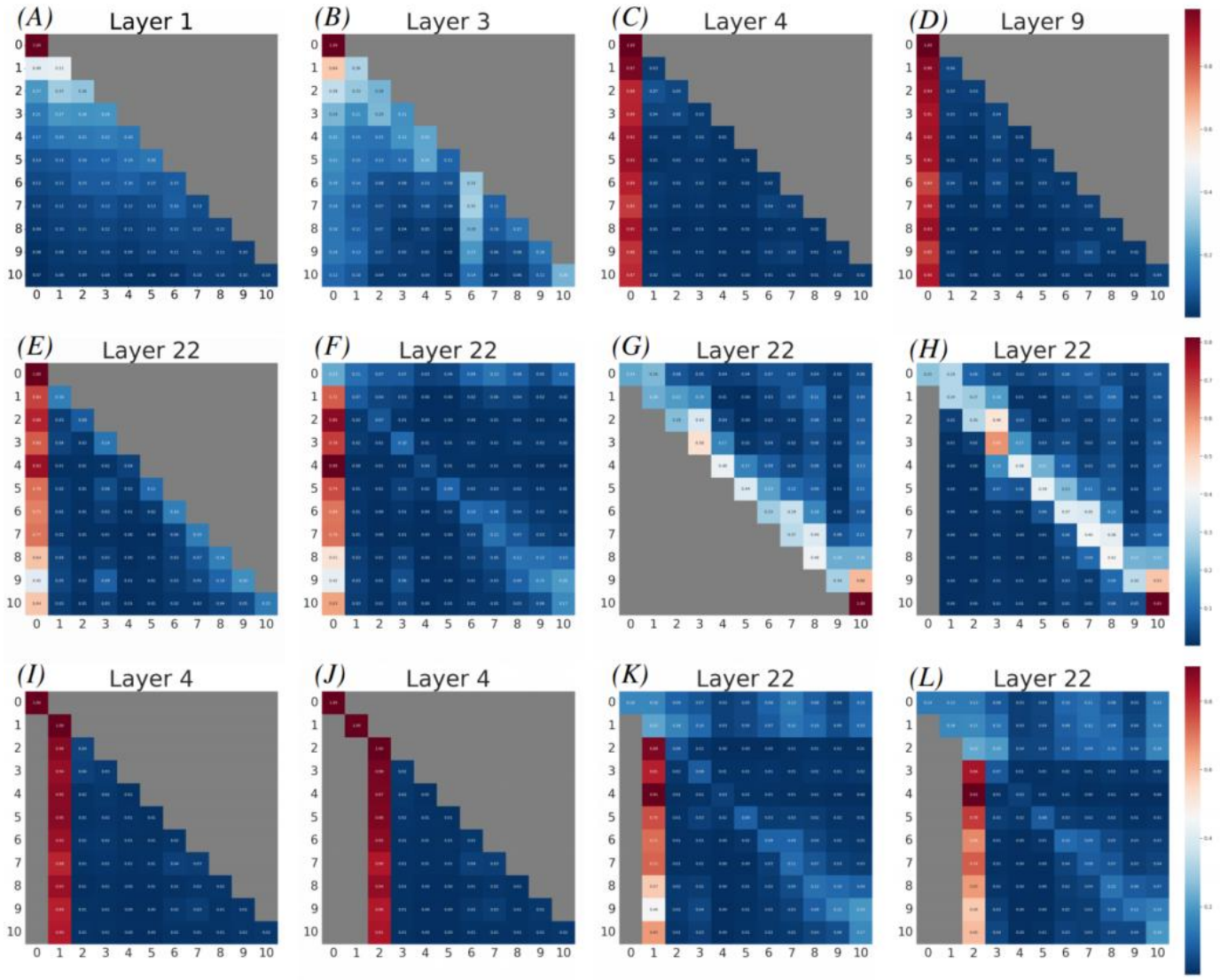
Empirical Study

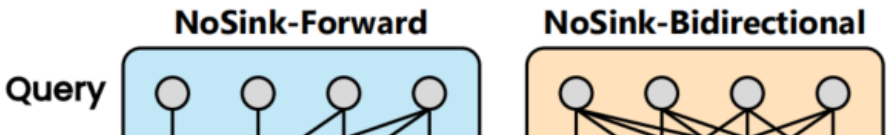


- "Attention Sink" limits the model's access to context. (Fig. A-E)
- Converting "Sink Layer" to BIDIR doesn't eliminate the sink. (Fig. F)
- Converting to BACK naturally removes the sink, as each token can't see the first token. (Fig. G)



Empirical Study





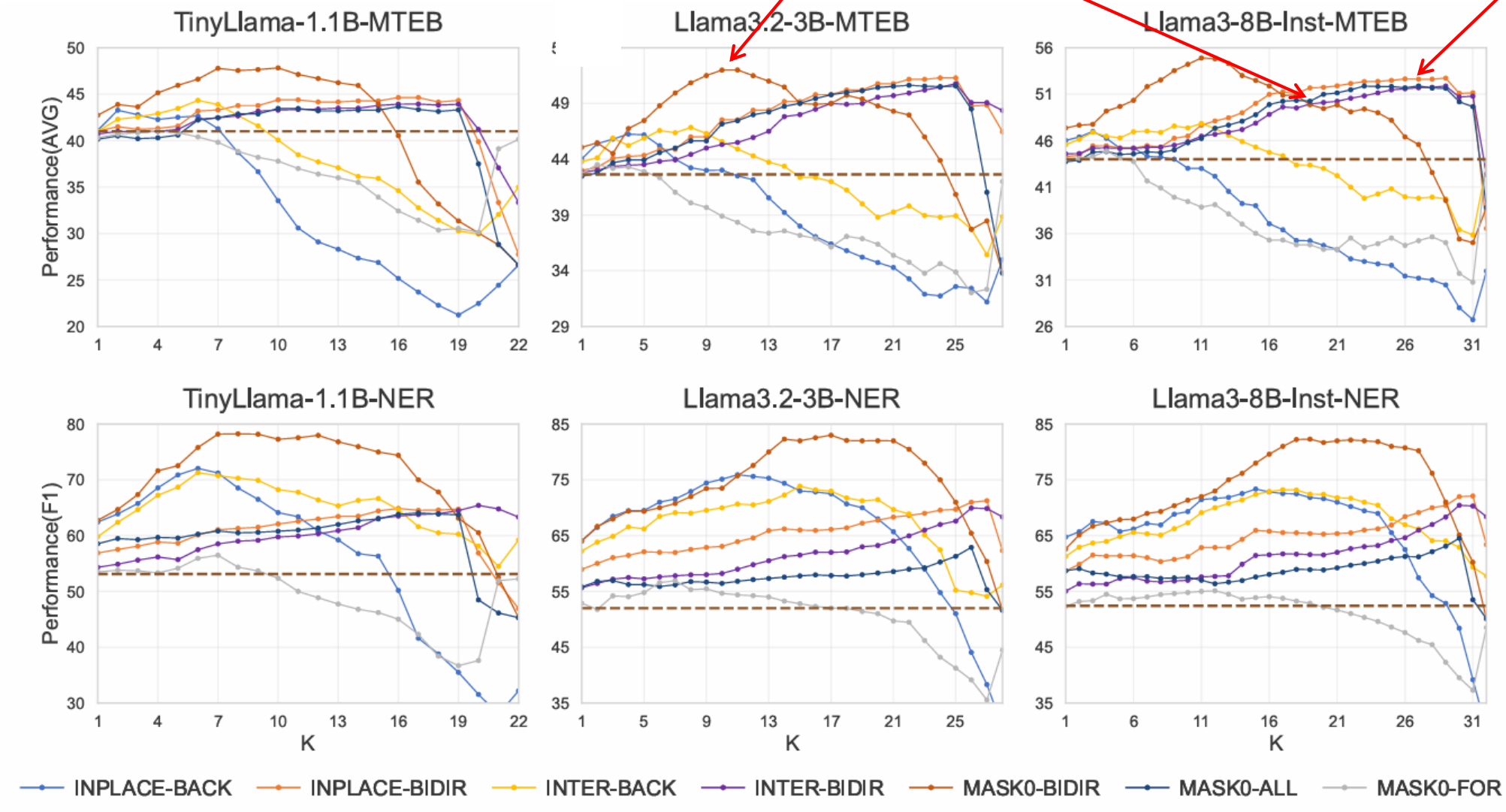
$$M_{NoSink0-FWD} = \begin{bmatrix} 0 & -\infty & -\infty & \dots & -\infty \\ -\infty & 0 & -\infty & \dots & -\infty \\ -\infty & 0 & 0 & \dots & -\infty \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\infty & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$M_{NoSink0-BIDIR} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ -\infty & 0 & 0 & \dots & 0 \\ -\infty & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\infty & 0 & 0 & \dots & 0 \end{bmatrix}$$

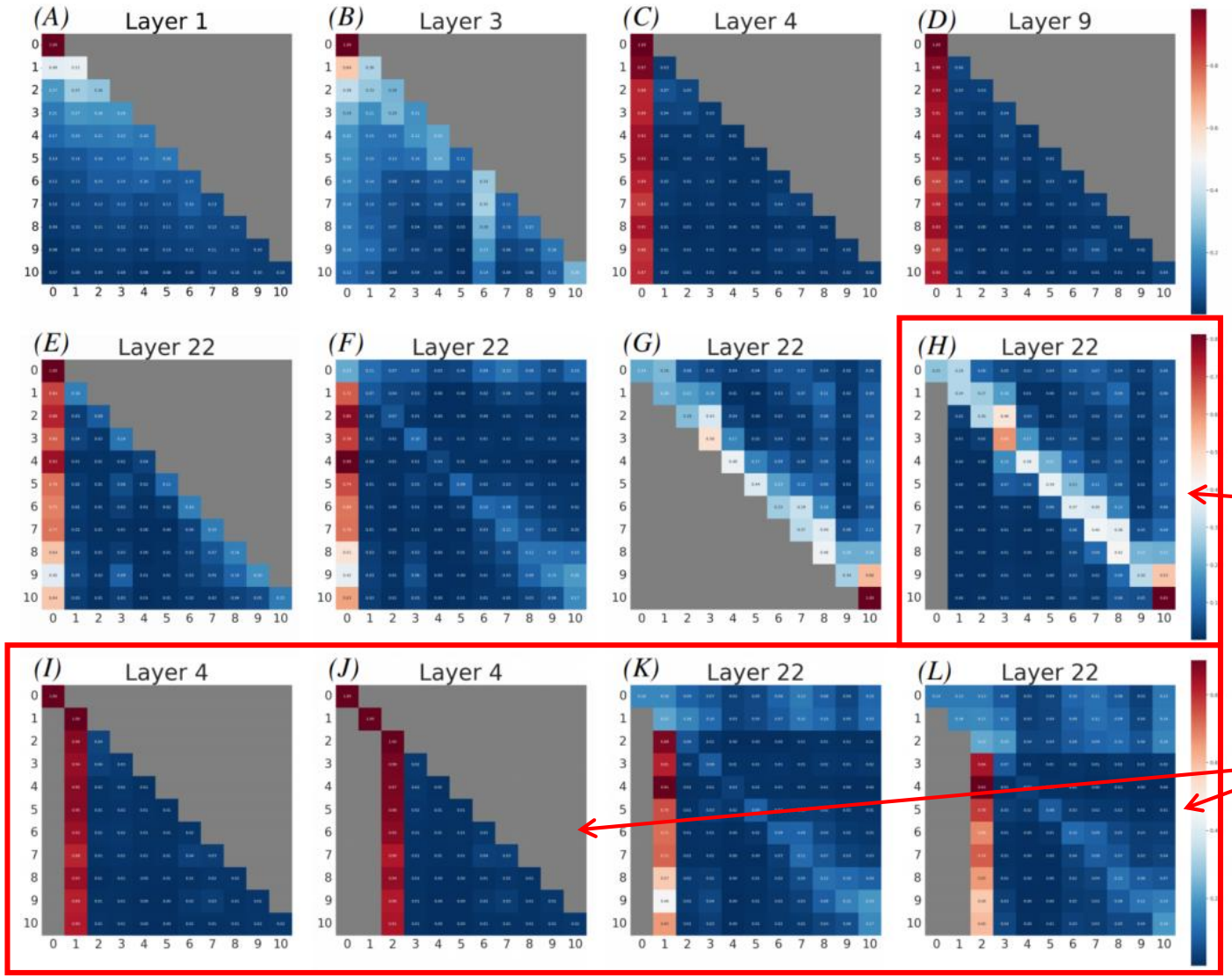
Figure 5: The schematic diagrams of the three model structures. NoSink-For and NoSink-Bi are abbreviations of NoSink-Forward and NoSink-Bidirectional, respectively.

Empirical Study

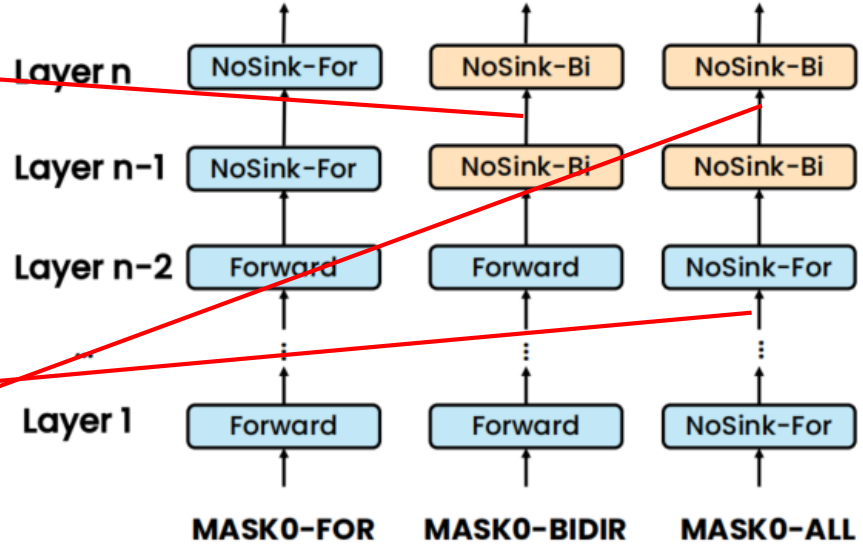
- MASK0-BIDIR(*Red line*) achieves the best performance
- MASK0-ALL(*Dark Blue*) and INPLACE-BIDIR (*Orange line*) perform similarly.



Empirical Study

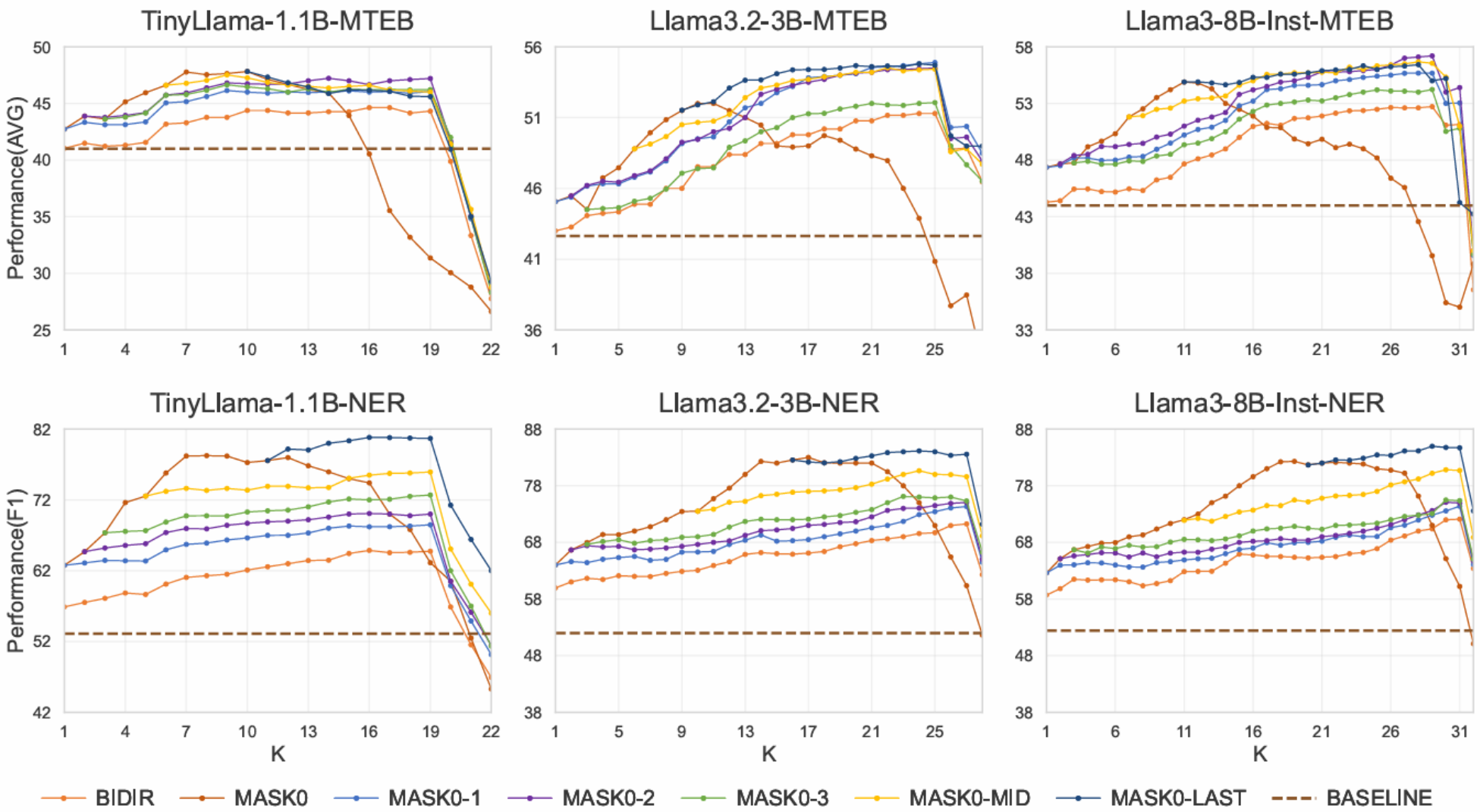


- MASK0-BIDIR eliminates the sink, enabling each token to focus more on other words. (H)
- Masking the first token of all layers(MASK0-ALL) doesn't eliminate the attention sink! (I-L)



Empirical Study

- By combining MASK0-BIDIR and INPLACE-BIDIR, MASK0&BIDIR can further enhance model performance.



Experiments

General	MTEB	CoNLL	AVG
BGE-M3	66.8	70.1	68.4
NV-Embed-V2	74.7	70.7	72.7
+*MASK0&BIDIR*	**75.8**	**85.9**	**80.8**
GTE-Qwen2-7B	70.9	71.5	71.2
+*MASK0&BIDIR*	74.1	80.2	77.2
E5-mistral	68.4	57.9	63.1
+*INPLACE-BIDIR*	70.9	74.8	72.9
+*INTER-BACK*	69.5	79.8	74.7
+*MASK0-BIDIR*	72.8	83.6	78.2
+*MASK0&BIDIR*	**74.0**	**85.2**	**79.6**

Medicine	PQAL	Chem	MQP	RCT	AVG
In-Context Learning	**66.0**	47.6	73.0	56.7	60.8
BGE-M3	63.0	68.9	77.7	69.8	69.9
MedicalBert	54.0	53.1	64.8	53.4	56.3
NV-Embed-V2	57.0	77.1	80.8	74.3	72.3
+*MASK0&BIDIR*	62.5	**77.9**	**81.4**	**84.6**	**76.6**
GTE-Qwen2-7B	56.2	78.0	77.7	77.1	72.3
+*MASK0&BIDIR*	63.0	77.9	78.8	82.9	75.7
BioMed-Llama	56.0	77.9	66.8	81.6	70.6
+*INPLACE-BIDIR*	59.0	78.4	73.4	82.5	73.3
+*INTER-BACK*	57.5	78.2	72.8	80.5	72.3
+*MASK0-BIDIR*	63.0	79.0	74.8	**88.0**	76.2
+*MASK0&BIDIR*	**65.5**	**79.4**	**75.2**	87.5	**76.9**

Finance	FIQA	FPB	NER	AVG
In-Context Learning	75.1	62.5	68.2	68.6
BGE-M3	81.0	94.2	68.2	81.3
FinBert	78.0	98.7	59.1	78.6
NV-Embed-V2	86.9	96.2	77.6	86.9
+*MASK0&BIDIR*	**87.9**	**97.8**	**83.1**	**89.6**
GTE-Qwen2-7B	83.3	95.8	67.6	82.2
+*MASK0&BIDIR*	85.0	97.6	81.3	88.0
Finma-7B	91.3	99.1	67.8	86.1
+*INPLACE-BIDIR*	93.2	99.1	69.6	87.3
+*INTER-BACK*	91.9	98.3	81.6	90.3
+*MASK0-BIDIR*	93.9	99.8	87.1	93.6
+*MASK0&BIDIR*	**94.4**	**99.8**	**89.5**	**94.5**

Law	SCOTUS		ToS	AVG
	mic-F1	mac-F1		
In-Context Learning	30.0	20.1	84.9	45.0
BGE-M3	68.4	46.8	84.7	66.6
InlegalBert	66.0	40.1	82.3	62.8
NV-Embed-V2	76.3	68.2	88.8	77.8
+*MASK0&BIDIR*	**78.7**	**71.5**	**91.0**	**80.4**
GTE-Qwen2-7B	73.9	62.6	89.1	75.2
+*MASK0&BIDIR*	77.3	69.9	91.5	79.6
LLama-Lawyer	73.7	61.1	89.7	74.8
+*INPLACE-BIDIR*	75.9	65.6	90.5	77.3
+*INTER-BACK*	74.1	64.6	90.3	76.3
+*MASK0-BIDIR*	75.9	65.6	90.7	77.4
+*MASK0&BIDIR*	**76.8**	**66.0**	**91.5**	**78.1**

Conclusion

- Investigated various strategies converting transformer decoder to encoder model.
- Discovered that attention sink phenomenon affects converted encoder performance.
- propose a novel training-free decoder-to-encoder conversion approach that significantly improves the performance of models in various domains.

Thanks!