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

Ziyong Lin[*1,2], Haoyi Wu[*3], Shu Wang[4], Kewei Tu[3✉], Zilong Zheng[1✉], Zixia Jia[1✉]
[1] BIGAI [2] THU [3] ShanghaiTech [3] UCLA
https://github.com/bigai-nlco/NoSink.git

## 1. Method

**In Low-Resource, Domain-Specific scenarios, general representation models often fail due to limited domain data and domain gaps**
**⇒A novel Training-Free conversion approach that significantly improves the performance of pretrained decoders in various domains!**

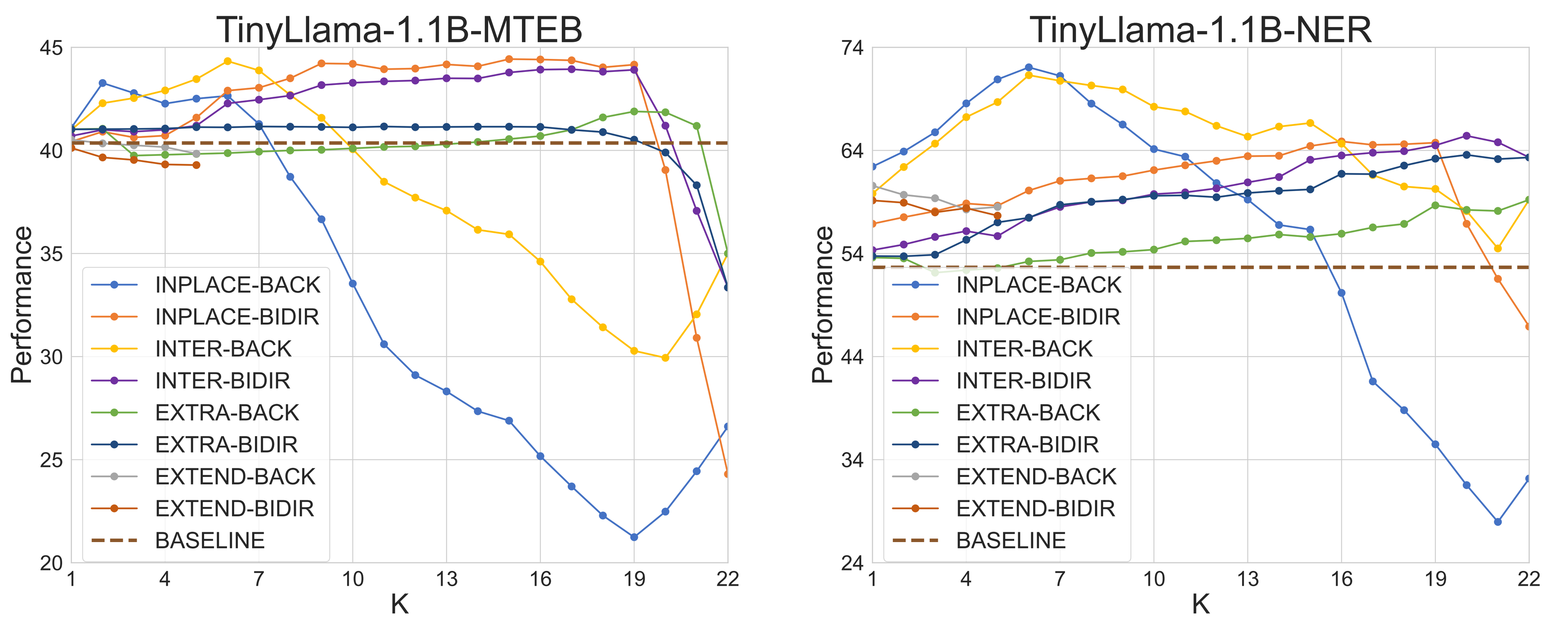
**Algorithm 1** Model Forward Pass

```
def forward(self, x = (x1, x2, ..., xT)):
    h = self.embedding(x) {Embed input sequence}
for i, layer in enumerate(self.decoder_layers) do
    mask = self.get_mask(i) {Select Mask based on layer ID}
    h_attn = layer.attn(h, mask) {Self Attention layer}
    h_ffn = layer.ffn(h_attn) {Feed-forward layer}
    h = h + h_ffn {Residual connection}
end for
    return self.head(h) {Generate output sequence} =0
```

**Attention Mask Types:**
- **FWD**: standard causal mask.
- **BACK**: transpose of causal mask (each token attends to itself and subsequent).
- **BIDIR**: zero matrix (no masking).
- **NoSink-X**: X with the first column elements set to negative infinity.

**Architecture Types:**(assume the model has $L$ layers)
- **INPLACE-BIDIR**: BIDIR if $i > L - k$; otherwise FWD.
- **INPLACE-BACK**: BACK if $i > L - k$; otherwise FWD.
- **MASK0-BIDIR**: NoSink-BIDIR if $i > L - k$; otherwise FWD.
- **MASK0-ALL**: NoSink-BIDIR if $i < L - k$; otherwise NoSink-FWD.
- **MASK0&BIDIR**: first FWD, then BIDIR, finally NoSink-BIDIR

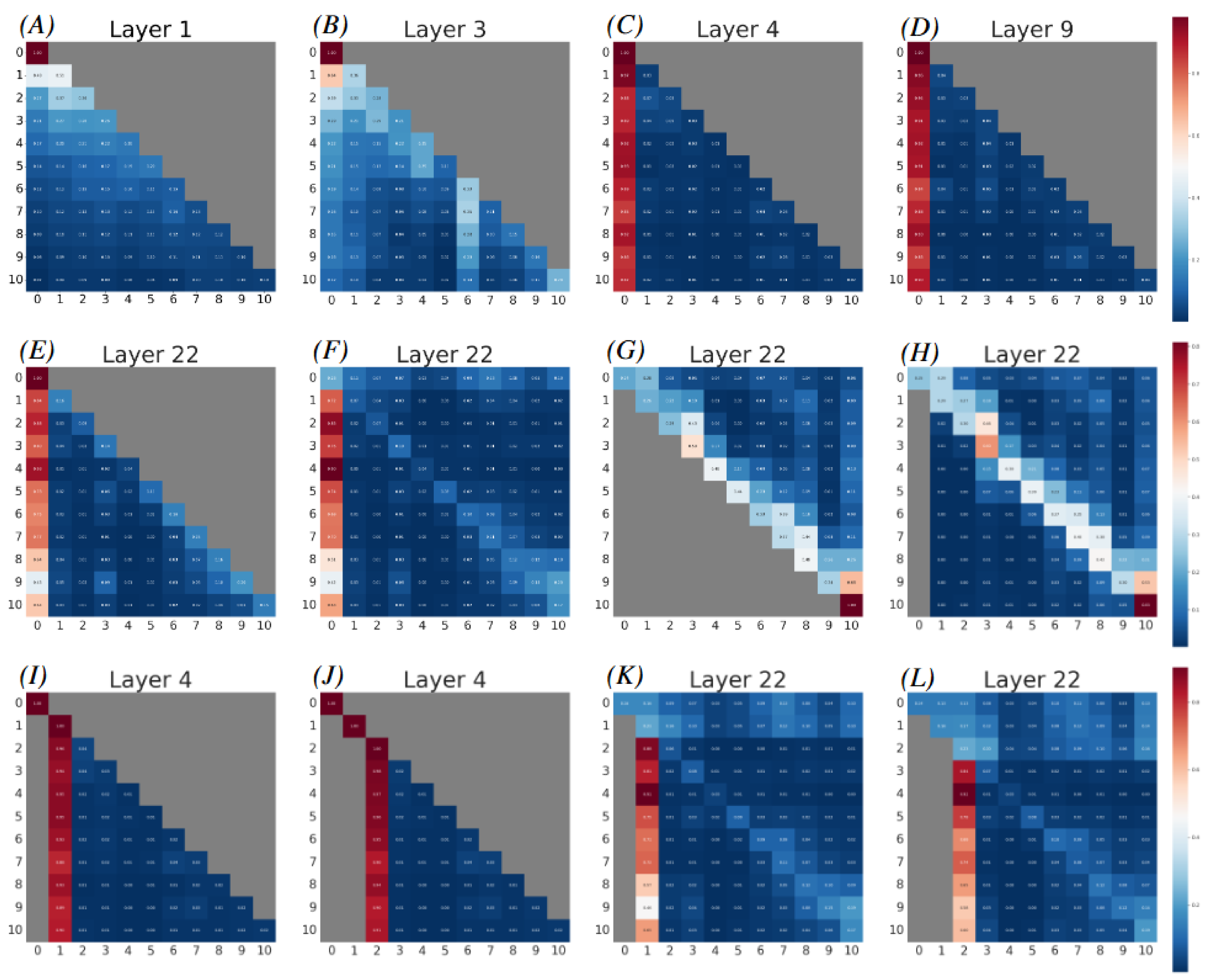
## 2. Experimental Results



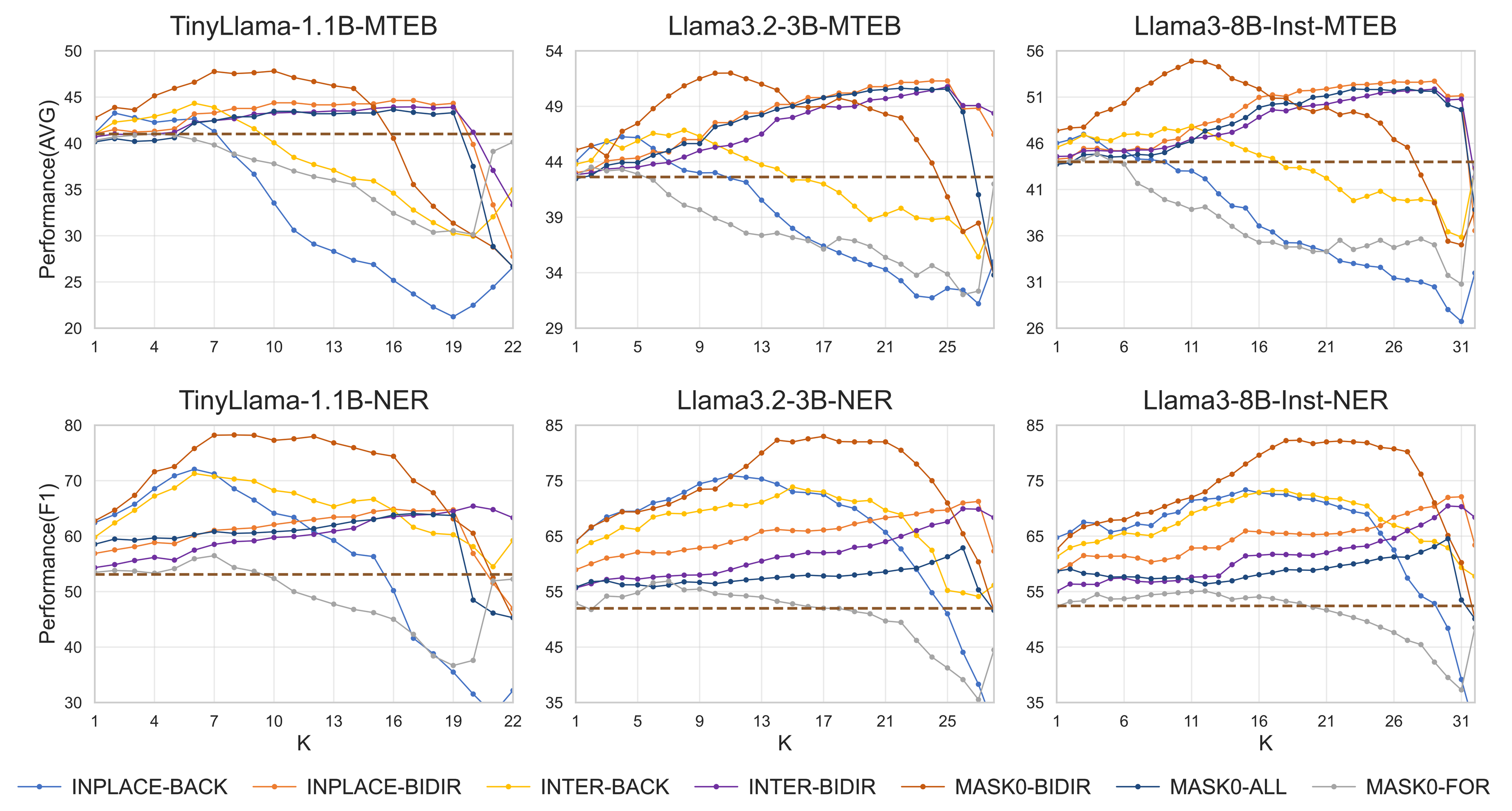

- **Maintaining some layers as FWD can be beneficial.**
- **INPLAC-BACK surpasses INPLACE-BIDIR when $k$ is small.**

## 3. Attention Map Visualization

**Figure:** (A)-(E) show attention score distributions in specific layers of a standard Transformer decoder. (F)-(H) display attention scores in the final layer with various masking methods. (I)-(K) demonstrate the attention sink shifting phenomenon, where masking initial tokens shifts the sink token to subsequent ones.

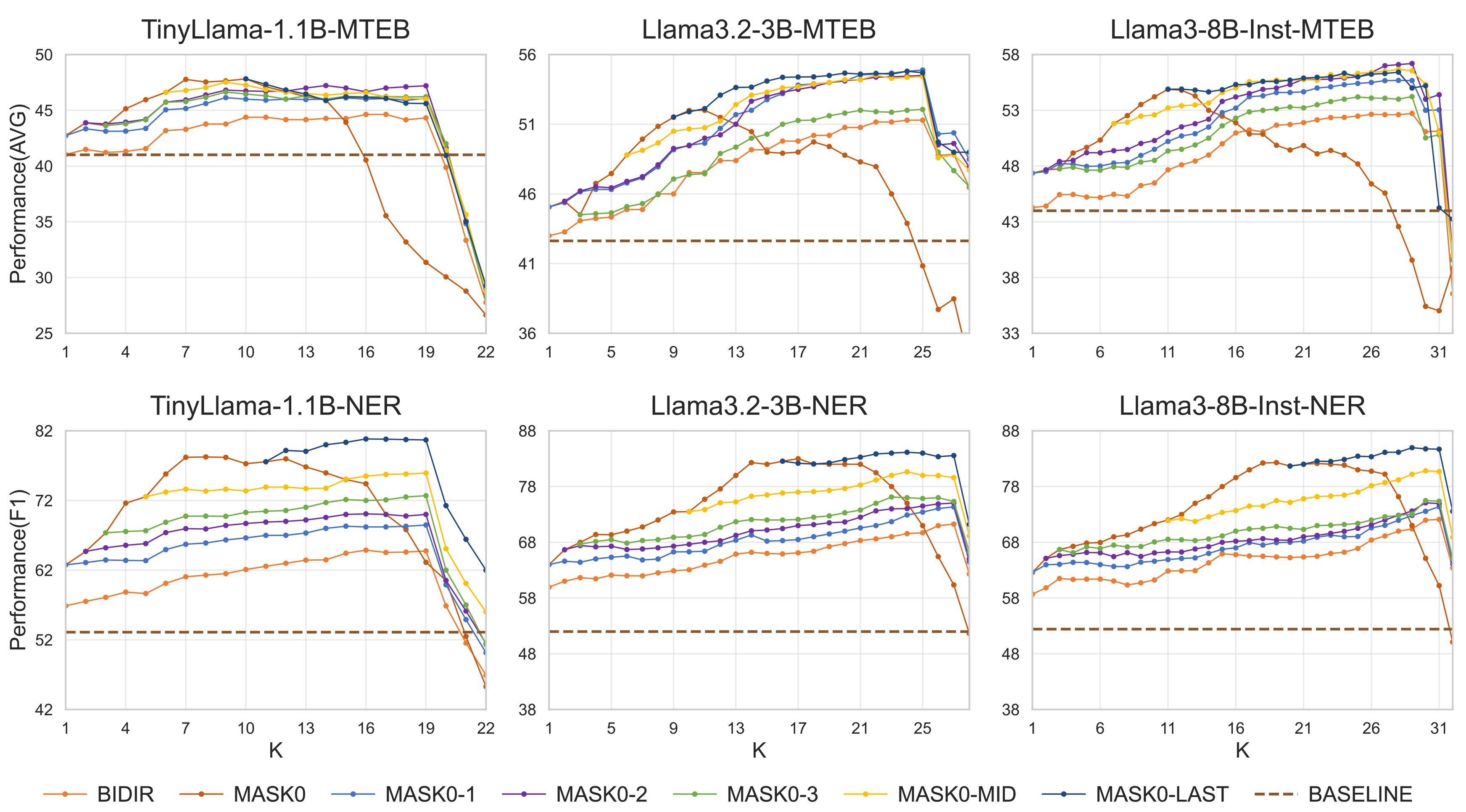



## 4. Experimental Results




- **MASK0-BIDIR achieves optimal performance by eliminating sinks.**
- **Note: directly masking the first token in all layers(MASK0-ALL) will not eliminate the sink!(see the last row of the score visualization on the left)**

## 5. Experimental Results




- **By combining the effectiveness of MASK0-BIDIR and the stability of INPLACE-BIDIR, MASK0&BIDIR can further enhance model performance.**

## 6. Experimental Results

| 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** |

| 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** |

| 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** |

| 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** |