Supplementary Material for "Hierarchical Knowledge Masking for Edge-Deployable Mixed Language Models"

Hui Li, Shanying Zhu, and Yan Song

Table R1. Results on Military Entity Recognition Dataset

Model	Precision	Recall	F1	Token Acc.	Macro-F1	Micro-F1
BERT	54.82	58.45	56.16	91.2	55.7	56.2
MERNIE	55.59	59.27	56.60	91.8	56.7	56.6
SpanBERT	55.91	59.88	57.02	92.0	56.3	56.9
Multi-BERT	57.40	60.50	58.80	92.7	56.1	58.7
T5-base	56.12	56.91	56.53	91.7	55.9	56.5
LLaMA-7B*	59.04	59.61	59.32	92.9	58.9	59.3
Ours	55.89	59.73	57.46	92.5	57.2	57.5

Table R2. Results on Entity Typing Dataset

Model	Macro-F1	Micro-F1	Accuracy	Hamming loss	Macro-AUC	Precision@3
BERT	70.05	77.61	78.20	0.094	0.62	70.1
MERNIE	72.00	78.72	79.10	0.088	0.64	70.7
SpanBERT	73.11	78.95	79.50	0.087	0.63	70.4
Multi-BERT	74.90	80.40	81.20	0.079	0.62	70.2
T5-base	75.20	80.33	81.69	0.082	0.61	69.8
LLaMA-7B*	76.69	80.73	82.19	0.076	0.66	71.2
Ours	76.51	79.50	81.63	0.072	0.65	70.9

Table R3. Results on Military Entity Relation Extraction Dataset

Model	Precision	Recall	F1	Macro-F1	AUC-PR	MCC
BERT	47.48	48.54	44.52	44.11	0.51	0.47
MERNIE	44.46	41.87	40.51	40.13	0.53	0.48
SpanBERT	48.10	49.70	46.80	45.30	0.52	0.48
Multi-BERT	50.31	51.60	49.60	49.36	0.51	0.47
T5-base	54.51	50.33	50.19	48.17	0.50	0.46
LLaMA-7B*	52.13	53.72	50.49	50.35	0.55	0.50
Ours	55.12	51.52	50.26	50.71	0.54	0.49

Table R4. Deployment Runtime & Energy Efficiency

Model	Runtime (min)	Idle=105 min	Efficiency vs. Idle	Energy (J/inf)	Latency (ms)
BERT	48.0	–	45.7%	0.82	82
MERNIE	51.0	–	48.6%	0.80	84
SpanBERT	50.0	–	47.6%	0.82	81
Multi-BERT	49.0	–	46.7%	0.81	82
T5-base	46.0	–	43.8%	0.95	96
LLaMA-7B*	– (OOM)	–	infeasible	–	–
Ours	57.0	–	54.3%	0.78	80