Feel free to contact me

Efficient Prompt Compression with Evaluator Heads for Long-Context Transformer Inference

Weizhi Fei Xueyan Niu* Guoqing Xie Yingqing Liu Bo Bai Wei Han

fwz22@mails.tsinghua.edu.cn

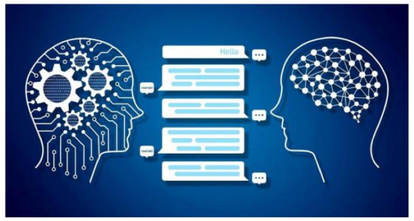
NEURAL INFORMATION PROCESSING SYSTEMS

1.Background: Long context inference

With the rapid application of LLMs, long context situations have emerged in many fields:

Long Doc

Agentic trajectory



2.Methods: Evaluator Heads

1. Find the *evaluator heads* with pilot experiments, using the attention scores of *evaluator heads* to evaluate the importance of tokens

2. Validating the properties of *evaluator heads*: Existence, Generalizability, Robustness.

3. Great efficiency in compressing long *prompts, building on the pre-filling stage.*

3.Experiments: Efficient and Effective

Methods	LongBench									ZeroSCROLLS		
	SingleDoc	MultiDoc	Summ.	FewShot	Synth.	Code	Avg.	# Tokens	κ_2	Avg.	# Tokens	κ_2
Original Prompt	39.7	38.7	26.5	67.0	37.8	54.2	44.0	10,295	-	32.5	9,788	-
Zero-shot	15.6	31.3	15.6	40.7	1.6	36.2	23.5	214	48×	10.8	32	306×
2,000 tokens constraint												
Retrieval-based Methods												
BM25	30.1	29.4	21.2	19.5	12.4	29.1	23.6	1,985	5×	20.1	1,799	5×
SBERT	33.8	35.9	25.9	23.5	18.0	17.8	25.8	1,947	5×	20.5	1,773	6×
OpenAI	34.3	36.3	24.7	32.4	26.3	24.8	29.8	1,991	5×	20.6	1,784	5×
Compression-based Methods												
Selective-Context	16.2	34.8	24.4	15.7	8.4	49.2	24.8	1,925	5×	19.4	1,865	5×
LLMLingua	22.4	32.1	24.5	61.2	10.4	56.8	34.6	1,950	5×	27.2	1,862	5×
LLMLingua-2	29.8	33.1	25.3	66.4	21.3	58.9	39.1	1,954	5×	33.4	1,898	5×
LongLLMLingua	39.0	42.2	**27.4**	**69.3**	**53.8**	56.6	48.0	1,809	6×	32.5	1,753	6×
EHPC (EMI)	**44.5**	**50.7**	24.8	68.9	51.5	**61.9**	**49.6**	2,004	5×	**34.6**	2,041	5×
1,000 tokens constraint												
EHPC (EMI)	**45.0**	**49.5**	23.9	67.4.9	59.0	**49.5**	**48.4**	1,024	10×			10×

Table 1: Overall comparison of the proposed method in terms of average performance and latency on the LongBench dataset, under the constraint of a compressed prompt length of 2048 tokens. For comprehensive results, please see Table 4 and Table 5.

Method	Performance	Latency	Training-free
LongLLMLingua	48.0	67.44	✓
LLMLingua	34.6	7.51	✓
LLMLingua-2	39.1	1.27	✗
EHPC (*ours*)	**49.6**	**0.88**	✓

Compressed prompt:
1. general for LLMs
2. preserving keys
3. high comp ratios

Advantages :
1. Train free
2. Efficiency
3. Effective