UC SANTA BARBARA

Microsoft Research

Augmenting Language Models with Long-Term Memory

To appear on Proceedings of NeurIPS 2023.

Weizhi Wang[1], Li Dong[2], Hao Cheng[2],
Xiaodong Liu[2], Xifeng Yan[1], Jianfeng Gao[2], Furu Wei[2]
University of California Santa Barbara[1] & Microsoft Research[2]

Augmenting Language Models with Long-Term Memory

Motivation:

- The input length prevents LLMs from processing long-form information beyond a fix-sized session.
- Previous memory-augmented model faces the memory staleness issue. The eldest memory are stale to current inputs because they are generated by stale parameters. Additionally, it requires training from scratch.

Contributions:

- We proposed **LongMem** framework to augment language models with long-term memory to read and comprehend up to 65k tokens.
- **LongMem** can easily adapt current SOTA LLMs to utilize long-term memory via efficient continual training.

LongMem Memory Flow

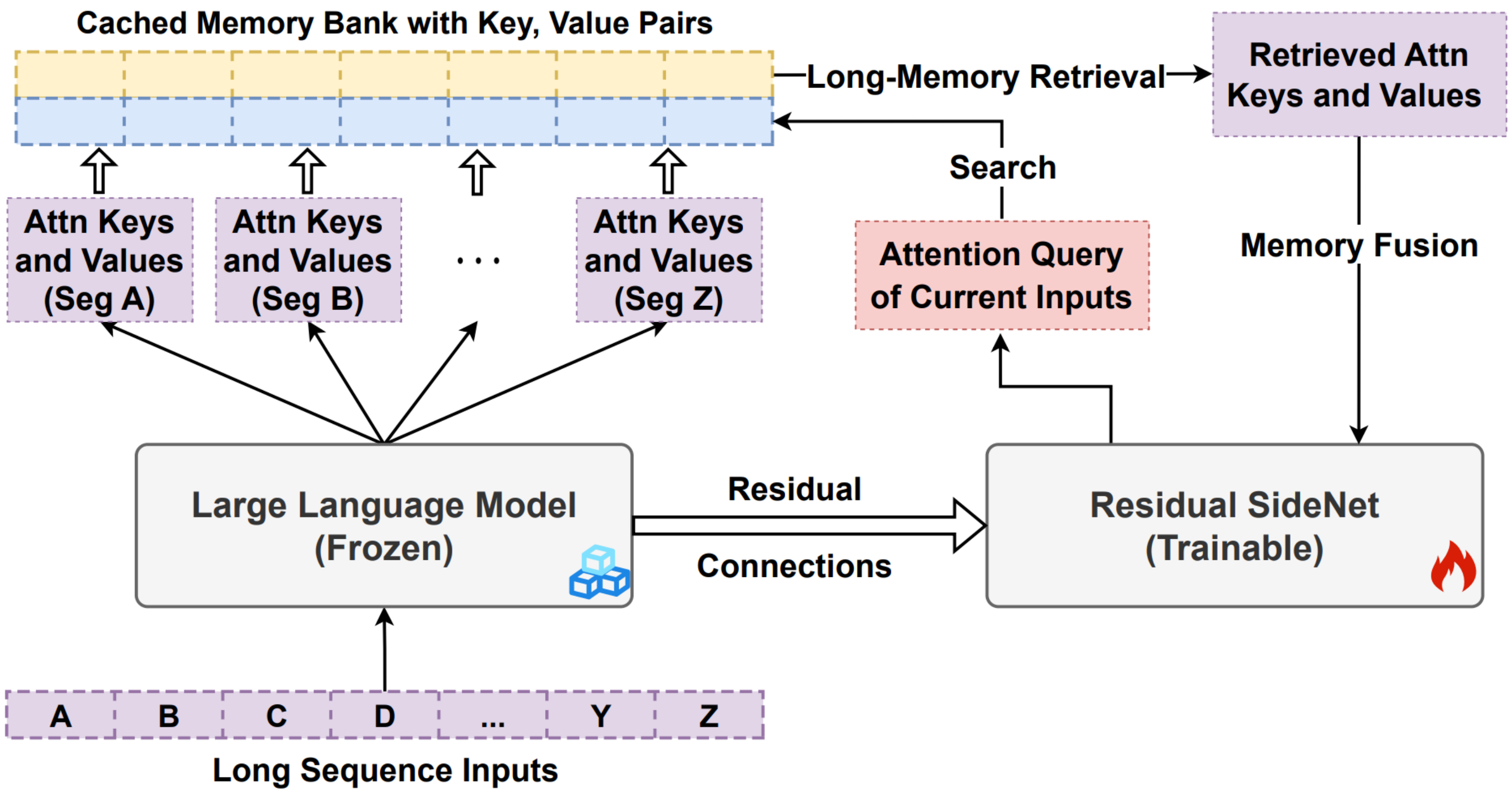


Figure 1: Overview of the memory caching and retrieval flow of LongMem. The long text sequence is split into fix-length segments, then each segment is forwarded through large language models and the attention key and value vectors of m-th layer are cached into the long-term memory bank. For future inputs, via attention query-key based retrieval, the top-k attention key-value pairs of long-term memory are retrieved and fused into language modeling.

LongMem Architecture

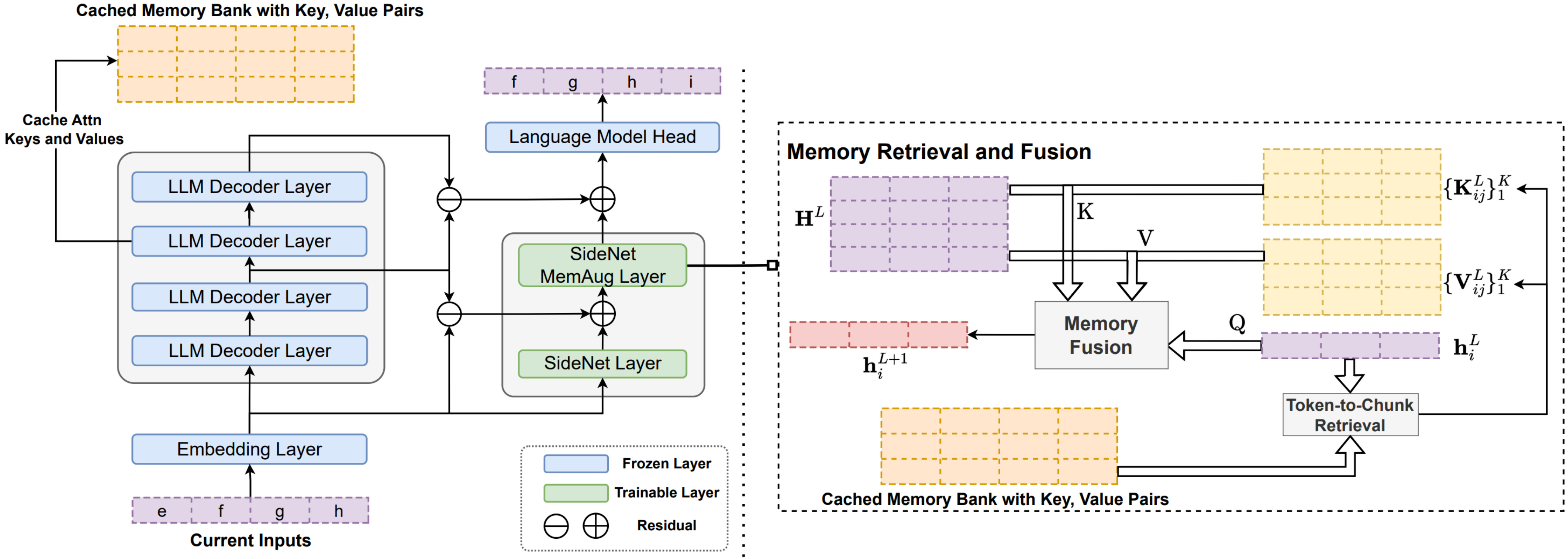


Figure 2: Overview of LongMem architecture. "MemAug" represents Memory-Augmented Layer.

SideNet Architecture

- LongMem = Backbone LLM (L' Layers) + SideNet (L Layers), L'=2L=24

- SideNet is composed of (L-1) transformer layer and 1 Memory-Augmented Layer (**9-th**):

$$\mathbf{H}_{\text{Side}}^{m_s} = f_{\theta_{\text{Mem}}}(\mathbf{H}_{\text{Side}}^{m_s-1}, \{\{\widetilde{\mathbf{k}}_{ij}, \widetilde{\mathbf{v}}_{ij}\}_{j=1}^{K}\}_{i=1}^{|x|})$$

- SideNet Initialization: $\Theta_{\text{Side}}^{\frac{l'}{2}} = \Theta_{\text{LLM}}^{l'}$

- Output attention keys-values pairs of 18-th layer of frozen backbone LLM are cached

- Cross-Network Residual Connections:

$$\mathbf{H}_{\text{Side}}^{l} = f_{\Theta_{\text{Side}}^{l}}(\mathbf{H}_{\text{Side}}^{l-1}) + (\mathbf{H}_{\text{LLM}}^{2l} - \mathbf{H}_{\text{LLM}}^{2l-2}), \forall l \in [1, L],$$

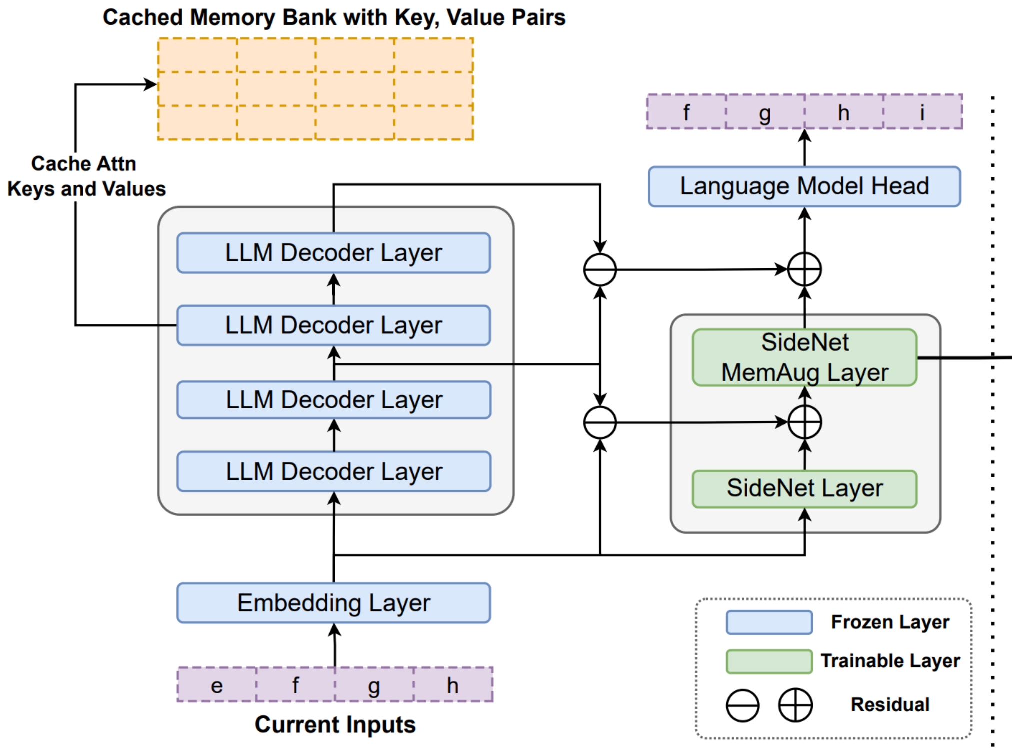


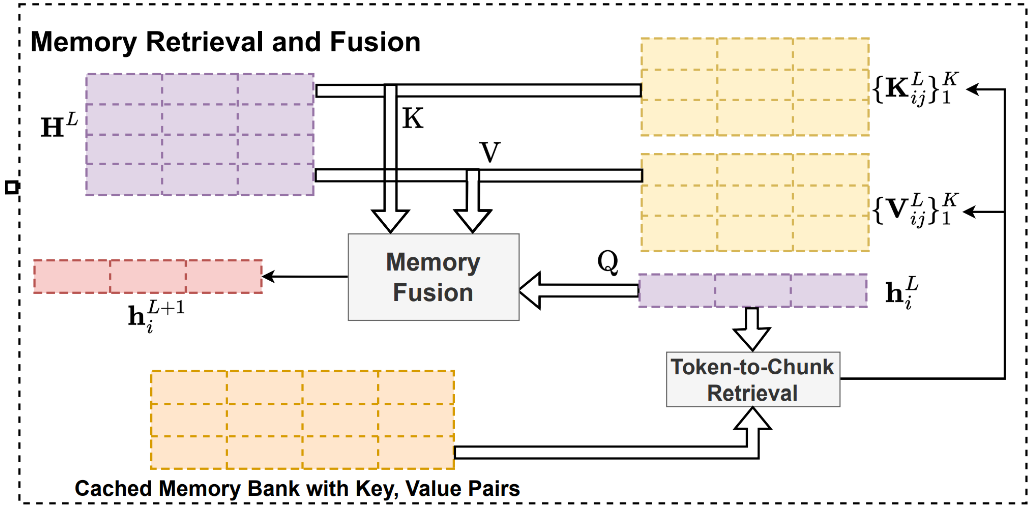
Figure 2: Overview of LONGMEM architecture.

Memory Retrieval and Fusion

- Memory Bank
  - Head-wise vector queue which maintains attention key-value pairs of latest **M** previous tokens, updates during the iteration
- Token-to-Chunk Retrieval:
  - A chunk contains chunk-size (csz) tokens, **csz=4**
  - The retrieval key for the chunk is computed via mean pooling on 4 attention keys of the tokens
  - Adopting token-to-chunk retrieval reduces the size of the retrieval index and accelerates the process.
- Memory Fusion:

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \quad \mathbf{M} = \mathrm{Concat}\left\{\mathrm{softmax}\left(\frac{\mathbf{Q}_i\widetilde{\mathbf{K}}_i^T}{\sqrt{d}}\right)\widetilde{\mathbf{V}}_i\right\}_{i=1}^{|x|},$$

$$\mathbf{H}^l = \mathrm{sigmoid}(g)\cdot\mathbf{A} + (1-\mathrm{sigmoid}(g))\cdot\mathbf{M},$$



Training and Evaluation Details

- Training Text corpus: Pile corpus, 26B tokens iterated
- Model Architecture: GPT2 Medium, 407M
- Training Objective: Maximum Likelihood on tokens
- 256 batch-size and 1024 sequence length
- Memory bank: 65k tokens, K=64, which is 16 chunks
- Batchfying: following bptt, truncate a document into segments and distribute them in consecutive batches
- Evaluation Tasks:
  - Language Modeling Tasks
  - Memory-Augmented In-Context Learning Tasks

Results on Memory-Augmented In-Context Learning

Model	In-Context #Demons.	In-Memory #Demons.	SST-2 ACC↑	MR ACC↑	Subj ACC↑	SST-5 ACC↑	MPQA ACC↑	Avg.
Majority	N/A	N/A	50.9	50.0	50.0	20.0	50.0	44.2
GPT-2*	4	N/A	$68.3_{11.6}$	$64.7_{12.5}$	$51.9_{4.2}$	$31.4_{4.4}$	$61.5_{11.8}$	55.6
MemTRM	4	2000	$67.5_{12.4}$	$64.6_{11.3}$	$53.2_{6.0}$	$29.6_{4.4}$	$63.0_{12.1}$	55.6
TRIME	4	2000	$69.5_{14.5}$	$63.8_{9.8}$	$51.5_{1.5}$	$31.8_{6.7}$	$63.6_{12.9}$	56.0
LongMem	4	2000	$\mathbf{71.8}_{14.0}$	$\mathbf{65.1}_{11.0}$	$\mathbf{53.8}_{3.7}$	$\mathbf{36.0}_{6.8}$	$\mathbf{65.4}_{12.8}$	**58.4**
GPT-2*	20	N/A	$68.2_{11.5}$	$63.4_{5.2}$	$57.6_{10.2}$	$33.6_{6.0}$	$70.8_{7.6}$	58.7
MemTRM	20	2000	$65.1_{9.6}$	$65.1_{9.3}$	$58.2_{10.6}$	$31.9_{6.3}$	$72.7_{7.4}$	58.6
TRIME	20	2000	$74.3_{13.9}$	$71.5_{2.5}$	$57.5_{11.4}$	$33.0_{4.6}$	$69.8_{7.8}$	61.1
LongMem	20	2000	$\mathbf{78.0}_{14.1}$	$\mathbf{78.6}_{3.3}$	$\mathbf{65.6}_{8.5}$	$\mathbf{36.5}_{7.5}$	$\mathbf{74.6}_{7.3}$	**66.7**

Table 5: Accuracy [%] of 4-shot and 20-shot ICL on 5 NLU tasks (SST-2, mr, subj, SST-5, mpqa). We sample 2000 extra demonstration examples and load them into cached memory. The subscript is the standard deviation across 6 runs. Avg. refers to the average accuracy on 5 datasets.

Due to the input length limit, LLMs can only adopt at most 20 demonstration examples in in-context learning. With LongMem, it can load up to 2000 demonstrations examples into the long-term memory and perform the in-context learning much better.

Results on Language Modeling Datasets

Dataset Splits	PG-22 S1	S2	S3	S4	S5	ArXiv
Len. Range	5K-10K	10K-100K	100K-500K	500K-1M	>1M	<60K
#Documents	500	100	30	8	1	100
Avg. #tokens	7.6K	47.6K	140K	640K	1.2M	15.4K

Table 1: Dataset Statistics of five splits of PG-22 based on length range and ArXiv.

Model	In-Context Len.	In-Memory Len.	PG-22 5K-10K	10K-100K	100K-500K	500K-1M	>1M	ArXiv
GPT-2*	1k	N/A	22.78	24.39	24.12	24.97	18.07	11.05
MemTRM	1k	65K	21.77	23.56	23.23	24.16	17.39	10.81
TRIME	1k	65K	22.21	23.50	23.74	24.32	17.80	10.95
LongMem	1k	65K	**21.29**	**23.01**	**22.55**	**23.35**	**16.71**	**10.05**

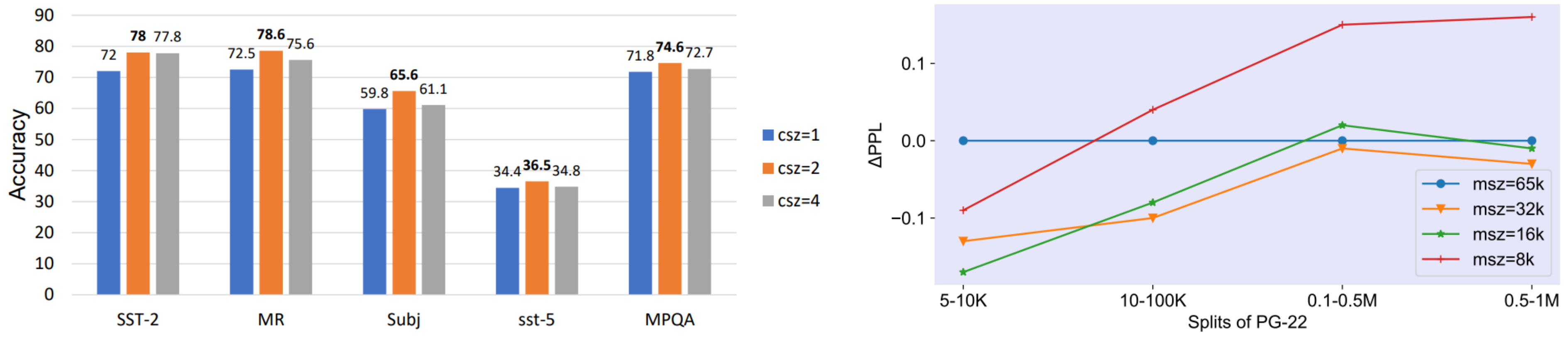
Table 2: Evaluation results on long-context language modeling datasets. We report token-level perplexity (PPL) (lower the better) on all datasets.

ChapterBreak Long-Context Modeling Benchmark

Model	#Params	In-Context Len.	In-Memory Len.	ChapterBreak$_{ao3}$ ctx-4k	ctx-6k	ctx-8k
GPT-2-XL† [RWC+19]	1.5B	1K	N/A	24%	24%	24%
GPT-3† [BMR+20]	175B	2K	N/A	28%	28%	28%
LocalTRM† [RSVG21]	516M	8K	N/A	24%	24%	24%
RoutTRM† [RSVG21]	490M	8K	N/A	25%	24%	24%
Bigbird† [ZGD+20]	128M	4K	N/A	26%	26%	26%
GPT-2*	407M	1K	N/A	18.4%	18.4%	18.4%
MemTRM	407M	1K	∞	28.3%	28.7%	28.7%
LongMem	558M	1K	∞	**37.7%**	**39.4%**	**40.5%**

Table 3: Zero-shot Suffix Identification Accuracy on AO3 subset of ChapterBreak. Baselines marked with † are directly cited from [STI22]. The MemTRM and LongMem loads the given 4k/6k/8k prefix contexts into cached memory, while the input length to local context is still 1k tokens.

Ablation Study on the Chunk-Size and Memory Size



(a)

(b)

Figure 4: (a) Accuracy on 5 NLU datasets given different chunk size during inference; (b) ΔPerplexity on 4 splits of PG-22 given different memory size during inference, in which the perplexity when msz=65k is used as baseline.

- Chunk-size of 2 performs best on in-context learning tasks, and we adopt csz=2 in inference.
- 32k memory size performs best on language modeling.

Ablation Study on the Effects of Memory Augmentation

Model	In-Context #Demons.	In-Memory #Demons.	SST-2 ACC↑	MR ACC↑	Subj ACC↑	SST-5 ACC↑	MPQA ACC↑	Avg.
Majority	N/A	N/A	50.9	50.0	50.0	20.0	50.0	44.2
GPT-2*	4	N/A	$68.3_{11.6}$	$64.7_{12.5}$	$51.9_{4.2}$	$31.4_{4.4}$	$61.5_{11.8}$	55.6
MemTRM	4	2000	$67.5_{12.4}$	$64.6_{11.3}$	$53.2_{6.0}$	$29.6_{4.4}$	$63.0_{12.1}$	55.6
TRIME	4	2000	$69.5_{14.5}$	$63.8_{9.8}$	$51.5_{1.5}$	$31.8_{6.7}$	$63.6_{12.9}$	56.0
LongMem	4	2000	$\mathbf{71.8}_{14.0}$	$\mathbf{65.1}_{11.0}$	$\mathbf{53.8}_{3.7}$	$\mathbf{36.0}_{6.8}$	$\mathbf{65.4}_{12.8}$	**58.4**
w/o Memory	4	0	$69.4_{12.4}$	$64.3_{12.1}$	$53.4_{7.7}$	$29.0_{5.2}$	$62.5_{12.3}$	55.7
GPT-2*	20	N/A	$68.2_{11.5}$	$63.4_{5.2}$	$57.6_{10.2}$	$33.6_{6.0}$	$70.8_{7.6}$	58.7
MemTRM	20	2000	$65.1_{9.6}$	$65.1_{9.3}$	$58.2_{10.6}$	$31.9_{6.3}$	$72.7_{7.4}$	58.6
TRIME	20	2000	$74.3_{13.9}$	$71.5_{2.5}$	$57.5_{11.4}$	$33.0_{4.6}$	$69.8_{7.8}$	61.1
LongMem	20	2000	$\mathbf{78.0}_{14.1}$	$\mathbf{78.6}_{3.3}$	$\mathbf{65.6}_{8.5}$	$\mathbf{36.5}_{7.5}$	$\mathbf{74.6}_{7.3}$	**66.7**
w/o Memory	20	0	$70.0_{12.8}$	$70.8_{6.2}$	$52.9_{4.6}$	$30.9_{6.4}$	$72.5_{7.5}$	59.4

Table 6: Ablation study results on the effect of memory augmentation of 4-shot and 20-shot ICL on 5 NLU tasks (SST-2, mr, subj, SST-5, mpqa). We sample 2000 extra demonstration examples and load them into cached memory. The subscript is the standard deviation across 6 runs. Avg. refers to the average accuracy on 5 datasets. "w/o" is short for "without".