Peripheral Memory for LLMs: Integration of Sequential Memory Banks with Adaptive Querying

Songlin Zhai, Yuan Meng, Yongrui Chen, Yiwei Wang, Guilin Qi

Southeast University, Nanjing China; University of California, Merced USA

Motivation

Large Language Models (LLMs) have revolutionized various natural language processing tasks with their remarkable capabilities. However, a challenge persists in effectively processing new information, particularly in the area of long-term knowledge updates without compromising model performance. This paper introduces a novel memory augmentation framework that conceptualizes memory as a peripheral component (akin to physical RAM), with the LLM serving as the information processor (analogous to a CPU). Memory is designed as a sequence of memory banks, each modeled using Kolmogorov-Arnold Network (KAN) to ensure smooth state transitions.

Memory Types	Working Memory	Implicit Memory	Explicit Memory	**Peripheral Memory**
Scalability	▭	▭	▭	▬
Reusability	▭	▭	▭	▬
Configurability	▭	▭	▭	▬

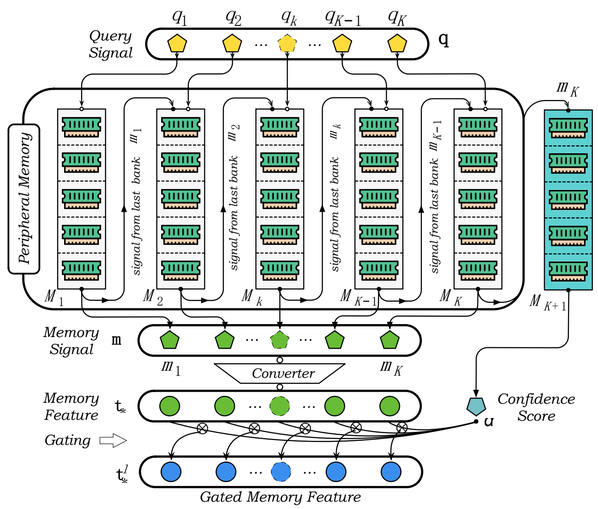
Memory Read & Write

Before reading from a peripheral memory, we first ***plug it into*** an LLM. Then, a query feature from the LLM is used to retrieve the memory:

$$\boldsymbol{t}_*^l = \sigma(\alpha) \cdot \boldsymbol{t}_* = \sigma(\alpha) \cdot \{\mathcal{M}((\boldsymbol{t}_*^l)^\top \boldsymbol{W}_0)^\top \boldsymbol{W}_1\}$$

$\boldsymbol{W}_0$converts the token feature into the query signal, acting as the ***outlet cable***. Similarly, $\boldsymbol{W}_1$maps the memory data into the memory feature adapted to LLM hidden feature space, which can be understood as the l***eading in cable***. This allows the memory and the LLM to be used effectively as a unified system. As such, the memory writing could be simply achieved by performing a fine-tuning process with setting requiring gradients of the memory.

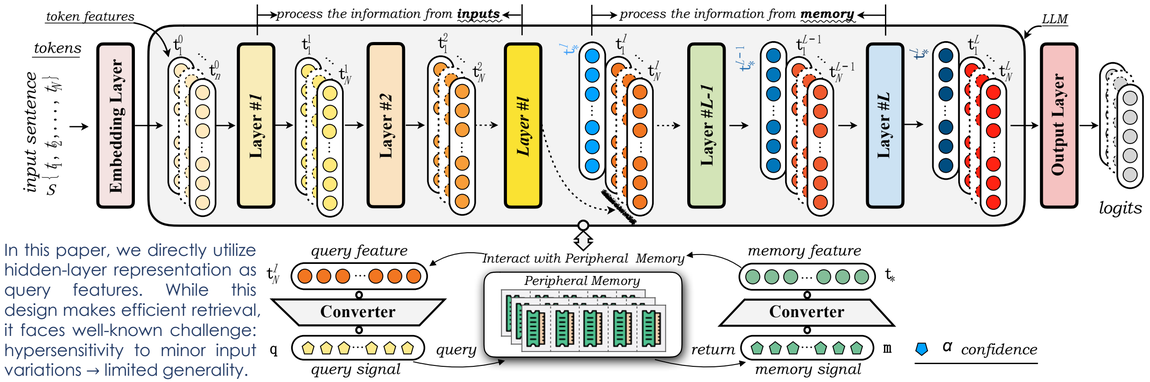
Peripheral Memory



The peripheral memory is composed of two components: ***memory banks to store data*** and ***the confidence bank to evaluate the relevance*** between the query feature and the retrieved information. The number of memory banks determine the memory bandwidth. Moreover, these banks are interconnected with each other, *i.e.*, the output of the one bank will be regarded as the input of the next one.

Overall Framework (LLM + Peripheral Memory)

Our peripheral memory decouples the memory module from the LLM, treating it as an independent storage component. In this design, the content stored in memory is treated as a dynamic variable, allowing the LLM to dynamically retrieve or write the necessary information as needed. The following figure illustrates the overall framework. The memory is connected to the LLM through a specialized converter, which acts like signal cables to facilitate feature conversion between the two components. Memory operations are driven by the hidden state of the last token from the LLM (serves as the query signal), and the query result is then integrated into the LLM as a prefix, enhancing the generation process.



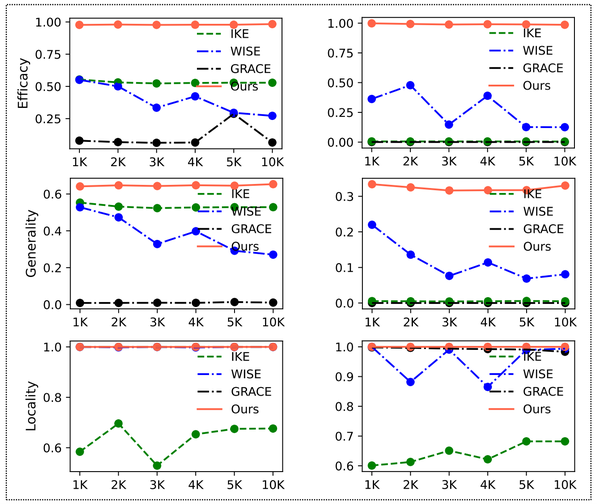
In this paper, we directly utilize hidden-layer representation as query features. While this design makes efficient retrieval, it faces well-known challenge: hypersensitivity to minor input variations → limited generality.

Experimental Results on Knowledge-based Model Editing Task

Editing performance of all compared methods under ***3K consecutive editing***. For our model, query features are derived from the last token hidden states in model's 24-th layer. Three fundamental metrics are used to evaluate model performance, including ***Efficacy***, ***Generality***, and ***Locality***.

Editor	**ZsRE** (Levy et al., 2017)				**CounterFact** (Meng et al., 2022)			
	Efficacy	Generality	Locality	Score	Efficacy	Generality	Locality	Score
Llama3 (8B)	0.2627	0.2598	/	0.2613	0.0087	0.0075	/	0.0081
FT-L 2020	0.0769	0.0666	0.0069	0.0501	0.0575	0.0047	0.0013	0.0212
LoRA 2022	0.1145	0.1116	0.0535	0.0932	0.0077	0.0117	0.0017	0.0070
ROME 2022	0.0339	0.0280	0.0015	0.0211	0.2507	0.1323	0.0097	0.1309
R-ROME 2024	0.0271	0.0243	0.0035	0.0183	0.4892	0.3662	0.0147	0.2900
MEMIT 2023	0.0000	0.0000	0.0396	0.0132	0.0000	0.0000	0.0722	0.0241
AlphaEdit 2025	0.0001	0.0000	0.0003	0.0001	0.0033	0.0017	0.0007	0.0019
PMET 2024	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EMMET 2024	0.0517	0.0486	0.0043	0.0349	0.5450	0.3882	0.0128	0.3153
GRACE 2023	0.0624	0.0095	1.0000	0.3573	0.0003	0.0000	0.9938	0.3314
IKE 2023	0.5233	0.5231	0.5289	0.5251	0.0055	0.0043	0.6509	0.2202
WISE 2024	0.3348	0.3283	0.9997	0.5543	0.1473	0.0763	0.9907	0.4048
Ours -1K archive	0.9597	0.5619	**1.0000**	0.8405	0.9038	0.2168	**1.0000**	0.7069
Ours +1K archive	**0.9805**	**0.6123**	**1.0000**	**0.8643**	**0.9915**	0.3108	**1.0000**	**0.7674**

To further investigate the effects of long-term editing, we conduct experiments by increasing the number of reads and writes in memory from 1K to 10K, and analyzing. The following figures present the results of our methods compared to several strong memory-based methods (including working & implicit memory) across varying numbers of edits.



Memory Sharing

The peripheral memory endows our method with an advantage, i.e., reusability. This follows the principle: ***store once, use everywhere***. The following tables present the results of 1K memory sharing, where knowledge is first stored using Llama3 (8B) and then transferred to two different LLMs: including Gemma2-it (2B) and Phi3 (3.8B).

Models on ZsRE	Efficacy	Generality	Locality	Score
Llama3 (8B)	0.9907	0.6090	1.0000	0.8666
Gemma2-it (2B)	0.9918	0.6989	1.0000	0.8969
	0.9713	0.6635	1.0000	0.8783
Phi3 (3.8B)	1.0000	0.6547	1.0000	0.8849
	0.0000	0.0000	1.0000	0.3333

Models on CounterFact	Efficacy	Generality	Locality	Score
Llama3 (8B)	0.9990	0.3150	1.0000	0.7713
Gemma2-it (2B)	0.9890	0.1948	1.0000	0.7279
	0.9790	0.2045	1.0000	0.7278
Phi3 (3.8B)	0.9997	0.2875	1.0000	0.7624
	0.9997	0.3075	1.0000	0.7691

Acknowledgment

This work is partially supported by National Nature Science Foundation of China under No. 62476058. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.