LoKiFormer: Locality-aware Attention with Decoupled Knowledge Memory for Efficient LLM Pretraining

Qiuwu Chen*, Zimo Liu*, Yuchen Li, Ying Sun, Yifan Zhang, Zhijie Qiu, Zeng You, Ryan Dong, Simeng Ma, Yaofo Chen†, Mingkui Tan†

ICML International Conference On Machine Learning

Background

- **Problem**: Existing LLM architectures remain inefficient in pretraining due to redundant modeling of sequence-internal local patterns and inflexible access to sequence-external global knowledge.
- **Limitations of Existing Methods**: Self-attention lacks explicit locality modeling, while MoE implicitly couples knowledge storage with computational transformation.

We aim to design an **efficient LLM architecture** that explicitly captures local patterns and directly retrieves global knowledge, enabling faster and more effective pretraining.

Local Patterns vs. Global Knowledge

- **Local Patterns**: Sequence-internal information, such as token correlations, syntactic structures, and short-range semantics, forming the basis for fine-grained representation learning.
- **Global Knowledge**: Sequence-external information, such as factual, commonsense, and domain knowledge, supporting reasoning and generalization beyond the current input sequence.

Existing LLM architectures process these two levels of information in an inefficient manner :

- Self-attention spends substantial computation on modeling short-range token interactions that can be captured more efficiently by locality-aware operators.
- MoE stores large amounts of knowledge implicitly inside expert parameters, making knowledge access indirect and difficult to interpret.

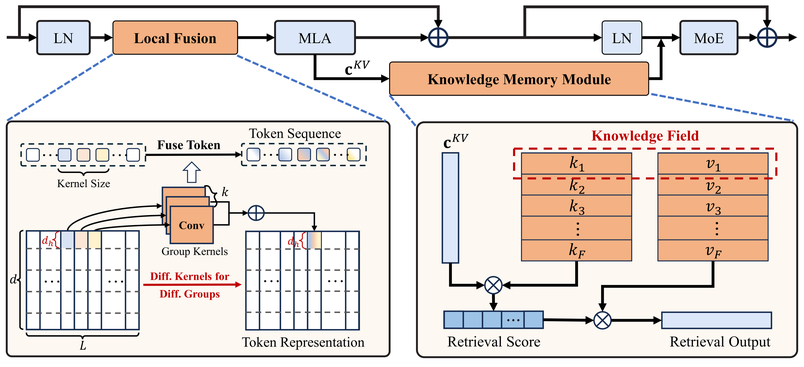
Our Goal: 1) explicitly capture local patterns for efficient attention; 2) enable direct access to global knowledge for flexible knowledge integration.

Contributions

- **Local Inductive Bias for Efficient Attention.** We identify that self-attention is inefficient for modeling short-range dependencies. To address this, we propose **Local Fusion Attention**, which introduces a convolutional fusion layer to explicitly capture local patterns before attention, reducing redundant local interactions and allowing attention to focus on more complex and global contextual relationships.
- **Decoupled Parametric Memory for Global Knowledge.** We argue that implicit knowledge representation in MoE couples knowledge storage with computation, limiting transparent access and flexible manipulation. To address this, we introduce a **Knowledge Memory Module**, which explicitly stores reusable knowledge abstractions in addressable slots, enabling direct knowledge retrieval and interpretable inference.
- **Improved Pretraining Efficiency.** Our architecture achieves **1.33× faster convergence** faster convergence during pretraining, reaching the same validation loss in **7.5K steps** steps compared to **10K steps** steps for the baseline. This shows superior optimization efficiency, leading to reduced computational cost and energy consumption.

Overview of LoKiFormer Architecture

LoKiFormer enhances the standard decoder block with two complementary modules for efficient pretraining: **1) Local Fusion Attention** introduces locality-aware representations before attention, enabling the model to capture short-range dependencies more efficiently and allowing attention to focus on broader contextual relationship; **2) Knowledge Memory Module** provides an explicit memory pathway for global knowledge retrieval, decoupling knowledge storage from computation and enabling direct, reusable, and interpretable access to global knowledge.



Local Fusion Attention

Insight: Natural language exhibits strong local syntactic and semantic dependencies.

Solution: We split hidden states into h groups and apply causal group convolution to fuse the previous k neighboring tokens before attention:

$$\hat{\mathbf{U}}_t^{(g)} = \sum_{s=0}^{k-1} \mathbf{U}_{t-s}^{(g)} \cdot \mathbf{\Theta}^{(g)}[s], \quad g = 1, \ldots, h. \tag{1}$$

Knowledge Memory Module

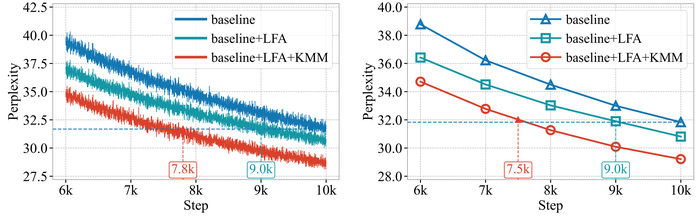
Insight: Global knowledge should be explicitly stored and directly retrieved.

Solution: KMM projects the latent KV representation $\mathbf{c}^{KV}$ into queries $\mathbf{H}$ and retrieves knowledge from learnable key–value fields $\mathcal{K} \in \mathbb{R}^{F \times d_u}$, $\mathcal{V} \in \mathbb{R}^{F \times d_v}$:

$$\mathrm{KMM}(\mathbf{H}; \mathcal{K}, \mathcal{V}) = [\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(c)}] \in \mathbb{R}^{L \times d},$$
$$\mathbf{Z}^{(i)} = \mathrm{softmax}\left(\frac{\mathbf{H}^{(i)} \mathcal{K}^{(i)\top}}{\sqrt{d_u}}\right) \mathcal{V}^{(i)}. \tag{2}$$

Comparisons with SoTA Methods

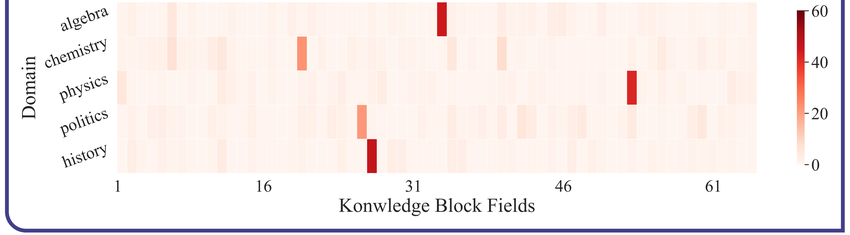
- Effect of LoKiFormer on pretraining performance. We reaches the same validation loss in 7.5K steps, compared to 10K steps for the baseline, achieving 1.33× faster pretraining convergence.



- Comparisons of LokiFormer-7B with **size-comparable LLMs** across four domains, including language, reasoning, code, math.

				Language			Reasoning		Code	Math
Model	Arch.	# Act.	# Total	MMLU	CMMLU	C-Eval	HellaSwag	ARC-C	HumanEval	GSM8K
Llama-3.1-8B	Dense	8B	8B	73.0	–	52.0	82.3	83.4	72.6	84.5
Qwen2.5-7B	Dense	7B	7B	76.6	79.1	76.2	81.5	63.4	84.8	**91.6**
Gemma-7B	Dense	7B	7B	64.3	–	–	81.2	53.2	32.3	46.4
InternLM2-7B	Dense	7B	7B	63.7	63.0	60.8	83.0	–	59.2	70.7
Phi-3 medium	Dense	3.8B	3.8B	78.0	–	–	82.4	91.6	62.2	91.0
Mixtral-8×7B	MoE	14B	46.7B	70.5	–	–	70.4	87.3	37.8	64.7
DeepseekMoE-16B	MoE	2.4B	16B	47.2	49.3	40.0	72.2	50.0	45.7	62.2
CCA-Attention-7B (Ours)	MoE	3B	7B	**91.5**	**93.4**	**92.8**	**89.5**	**96.2**	**87.0**	73.0

- Field–domain associations of KMM. Removing a specific field mainly degrades its corresponding MMLU domain, indicating explicit and interpretable knowledge retrieval.



Contact Information and Code

- Email: chenqiuwu@aigcode.net
- Email: mingkuitan@scut.edu.cn
- Code: github.com/zliu69/aigcode_lokiformer

PDF

Code