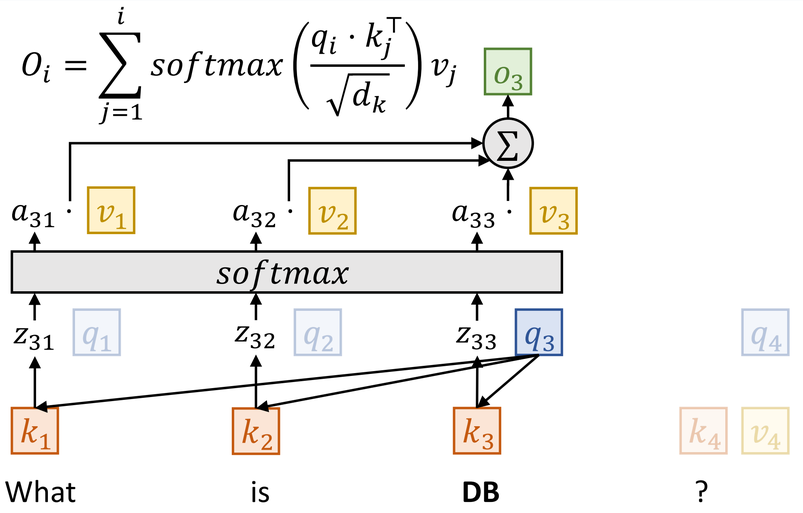
AlayaDB: The Data Foundation for Efficient and Effective Long-context LLM Inference

Yangshen Deng*, Zhengxin You*, Long Xiang*, Qilong Li, Peiqi Yuan, Zhaoyang Hong, Yitao Zheng, Wanting Li, Runzhong Li, Haotian Liu, Kyriakos Mouratidis, Man Lung Yiu, Huan Li, Qiaomu Shen, Rui Mao, **Bo Tang**
research@alayadb.ai

Attention in LLM

$$O_i = \sum_{j=1}^{i} softmax\left(\frac{q_i \cdot k_j^\top}{\sqrt{d_k}}\right) v_j$$



Sparse Attention

- Only consider tokens with high a_{ij}.
- Turns into a ***vector search problem***.

Given q_i, select token j with large $q_i \cdot k_j^\top$

Existing Algorithms

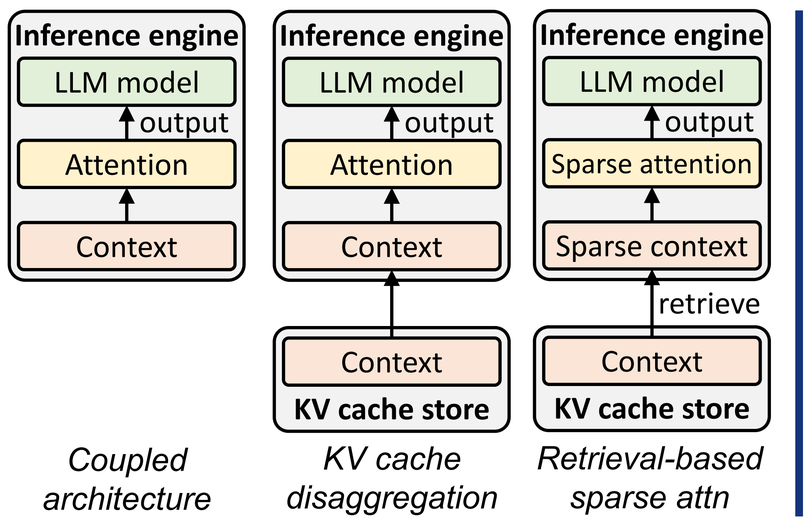
- Use **Top-k** to find critical tokens.
- k is usually statically determined.

However, it neglects the nature of *attention*.

Challenges of Long Context LLM Inference

- **Large KV Cache -> *High GPU memory consumption.*** *Solution: Offload & Reuse*
- **Heavy attention computation -> *High latency.*** *Solution: Sparse Attention*

New Disaggregation Level with Vector Database



Disaggregate both ***KV cache*** and ***attention*** from inference engine, and encapsulate them into a vector database.

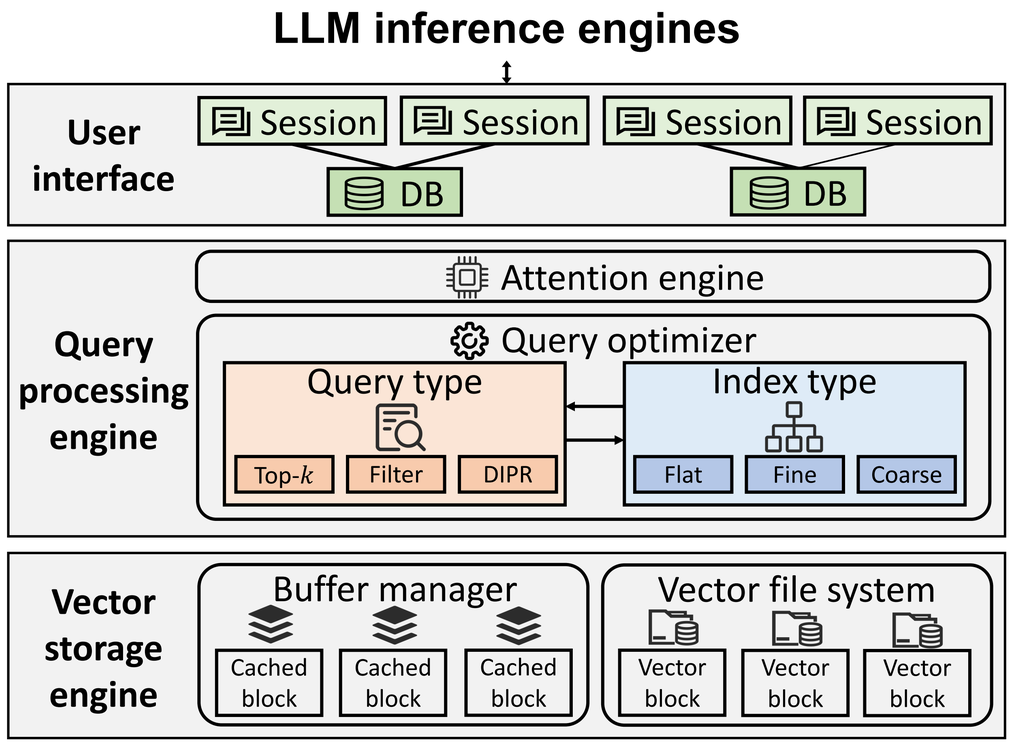
Store & Reuse Context *via vector storage engine.*
Sparse Attention *via vector search engine.*

Comparison of Different Architectures

	Latency	Quality	GPU memory	Usability
Coupled architecture	High	Good	Large	Good
KV cache disaggregation	Medium	High	Large	Medium
Retrieval-based sparse attn	—	Medium	Small	Bad
AlayaDB	**Low**	**Good**	**Small**	**Good**

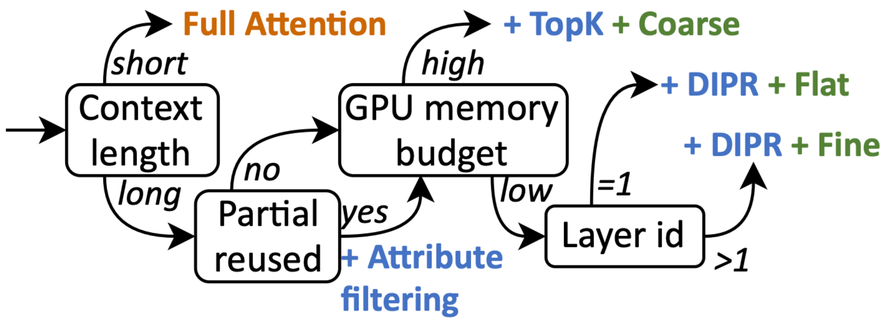
System Design and Optimizations

System Architecture



Query Optimizer

To choose the best index and query type.



Simple and Compatible Interface

Use AlayaDB out-of-the-box with modifying only two lines of code.

- Replace DynamicCache in **huggingface/transformers** with ***session***.
- Replace API of **flash-attention** with ***session.attention***.

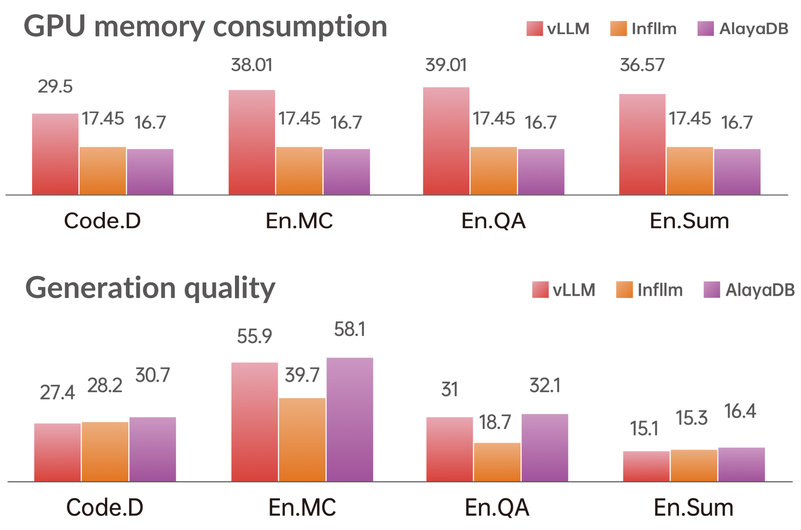
Dynamic Inner Product Range Query (*DIPR*)

A better vector search target than Top-k for sparse attention.

- Targets to select **all** tokens with **high attention scores.**
- Focuses on the **values** of attention scores, instead of the **rankings**.

$$a_{ij} > \alpha \times \max_{s \in [1,n]}(a_{is}) \longrightarrow q_i \cdot k_j^\top > \max_{s \in [1,n]}(q_i \cdot k_s^\top) - \beta$$

Experiment Results



Highest generation quality with minimal GPU memory.

Check our paper for more technical details and results.