# AlayaDB: Co-Optimizing Attention and KV Cache for Long-Context LLM Inference

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of different dynamic memory allocation strategies (e.g., FlowKV, PagedAttention) on the inference efficiency of BSFA-enhanced Llama-3-70b when scaling to contexts beyond 128K. 10 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: AlayaDB: The Data Foundation for Efficient and Effective Long-context LLM Inference. Research question: What is the impact of different dynamic memory allocation strategies (e.g., FlowKV, PagedAttention) on the inference efficiency of BSFA-enhanced Llama-3-70b when scaling to contexts beyond 128K tokens in legal benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

## 3 Results

11 papers retrieved. 10 claims extracted; 3 independently verified. Quality review score: 5.8/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| vLLM, SGLang, and HuggingFace transformers are widely-used LLM inference systems in coupled architecture. | × | 0.06 |
| Mooncake and LMCache are representative LLM inference systems in KV cache disaggregation. | × | 0.15 |
| InfLLM and RetrievalAttention are examples of retrieval-based sparse attention solutions. | × | 0.12 |
| AlayaDB is designed to overcome the limitations of existing LLM inference systems/solutions and enable efficient and eff | ✓ | 0.23 |
| AlayaDB solves a bi-objective optimization problem: meeting the SLOs of different workloads by consuming less GPU memory | × | 0.09 |
| AlayaDB decouples both KV cache and attention computation and encapsulates them into a monolithic vector database. | ✓ | 0.25 |
| AlayaDB lightens the burden of the LLM inference engine by separating cache management and attention computation. | × | 0.14 |
| AlayaDB simplifies the interface between the LLM inference engine and KV cache service by only returning the attention r | × | 0.13 |
| AlayaDB sheds light on co-optimizing attention computation and KV cache management in a monolithic vector database. | ✓ | 0.18 |
| AlayaDB achieves small GPU memory consumption, low inference latency, good generation quality, and good solution usabili | × | 0.05 |

## References


- `http://arxiv.org/abs/2504.10326v1`

- http://arxiv.org/abs/2508.06447v2
- http://arxiv.org/abs/2310.05276v1