

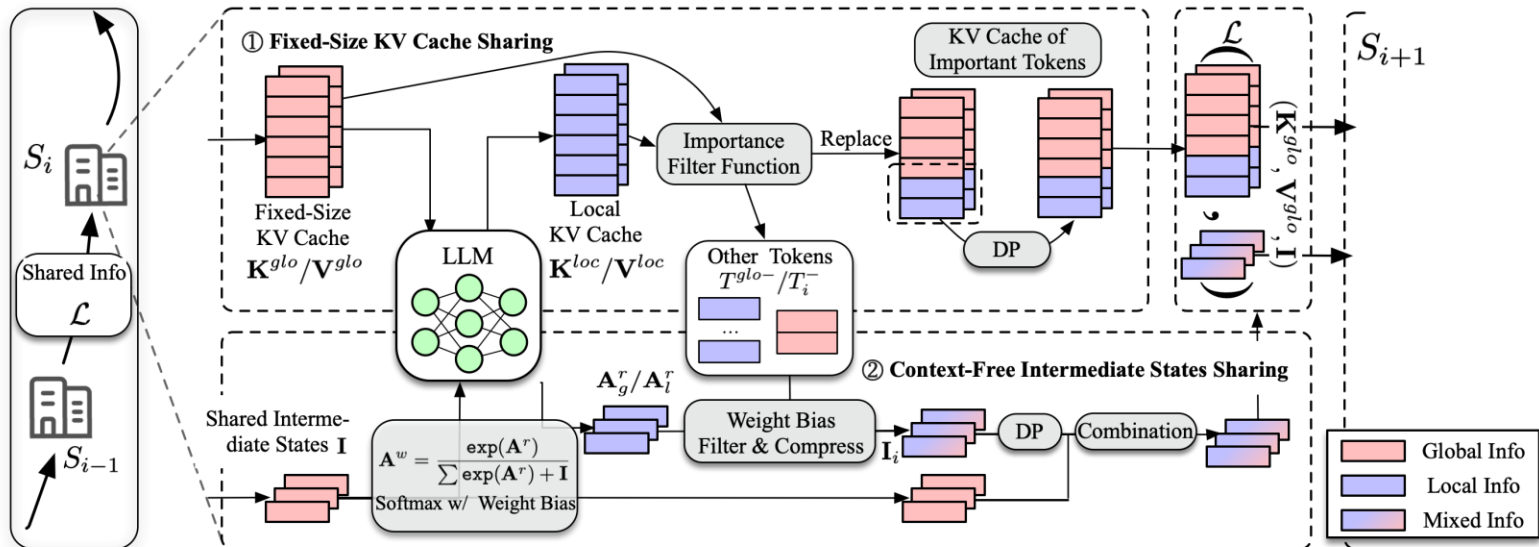
BEFI: Balanced and Efficient Federated Inference of Large Language Models

Lulu Zhang, Qian Tao, Zimu Zhou, Yuanyuan Zhang,
Yuxiang Wang, Yongxin Tong

Frontiers of Computer Science, DOI: [10.1007/s11704-026-60010-4](https://doi.org/10.1007/s11704-026-60010-4)

Problems & Ideas

- Problems: Federated LLM inference suffers from continuously growing and structurally imbalanced KV cache communication overhead across data silos, compounded by accumulating differential privacy noise.
- Ideas: BEFI resolves this by sharing only a fixed-size KV cache for important tokens and a constant-size intermediate state for the rest, achieving both communication efficiency and balance regardless of context length.



Main Contributions

- Contributions:
 - To the best of our knowledge, we are the first to define and study the problem of federated LLM inference with efficient and balanced communication overhead;
 - We propose BEFI, a novel framework designed to ensure balanced communication overhead through efficient sharing of limited-size KV cache information;
 - For other contexts, we introduce a context-free intermediate states sharing mechanism, which enables compressed intermediate states exchanged among silos, maintaining a constant communication size regardless of context length.

