Reading List

- **Preliminary**
  - Understand how Transformer works, including training and inference. What is self attention? What is K Q V? What are autoregressive models? Why do we need to cache K V during inference?

    https://www.zhihu.com/question/485876732

    https://zhuanlan.zhihu.com/p/27876127626

    https://zhuanlan.zhihu.com/p/718086680

Base:

- **Excellent Open Source Programs**
  - **(vLLM) Efficient Memory Management for Large Language Model Serving with PagedAttention, SOSP 23**
  - **SGLang: Efficient Execution of Structured Language Model Programs**
- **Prefill-Decode**
  - DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving, OSDI 24
  - SarathiServe:Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve, OSDI 24
- **Re-Scheduling**
  - Llumnix: Dynamic Scheduling for Large Language Model Serving

First Direction: utilize sparsity to re-design systems

（实验验证cross-attention这一现象）

- **Sparsity**
  - StreamingLLM: Efficient Streaming Language Models With Attention Sinks
  - H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models
  - InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management
- **KVCache Reuse**
  - CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion

Second Direction: Accelerating Attention Computation

（硬核CUDA代码）

- **Attention**
  - **FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness**
  - **FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning**

- **FlashInfer: Efficient And Customizable Attention Engine For LLM Inference Serving**

Thrid Direction: Deploy LLM on single device

（可以先在vLLM、SGLang尝试，后续再到FlexGen）

- FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU

Fourth Direction：Store new KVCache

（探索性工作）

- CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion

Fifth Direction: Survey and Benchmarking LLM Inference Simulators

（评估调研工作）

- VIDUR: https://github.com/microsoft/vidur
- Splitwise-Sim: https://github.com/mutinifni/splitwise-sim
- DistServe Simulator: https://github.com/LLMServe/DistServe/tree/main/simdistserve