# 顶点课程期末答辩海报

22230608 范一泽 22230610 曹晨旭 22230616 黄勇鑫

## SSSKV-lite 推理框架优化:
## StreamingSummarySnapKV-lite
## 稀疏注意力+KVcache优化+KV量化

### 1. 三段式连续上下文（Prefix / Summary / Recent）

通过连续块重组**系统提示**、**长期摘要**与**短期对话**，在控制 KV 规模的同时保持语义连续性。

### 2. 长期记忆压缩

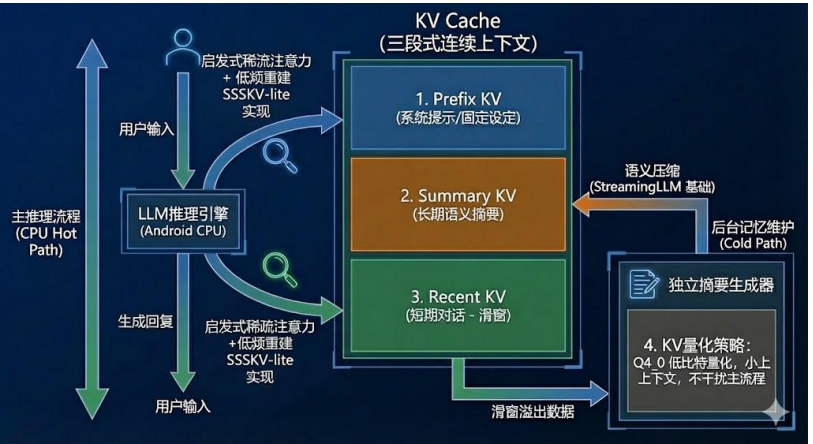
**StreamingLLM**的基础上，以语义**摘要**替代被滑窗裁剪的历史内容，实现有限 KV 条件下的长期对话记忆。

### 3. 面向 CPU 的 SnapKV-lite 实现

启发式**稀疏注意力**方法的 KVsparsing + 低频重建，替代 token 级稀疏 KV，在 Android CPU 上稳定实现 SnapKV 效果。

### 4. KV量化策略

摘要生成在独立小上下文、低比特(Q4_0)**KV量化**的 context 中完成，避免干扰主推理流程。




## PeakChat
## 本地高性能对话

### 基于MedChatZH数据集的指令微调

- Continued training with domain-specific corpus
- Fine-tuning for instructions
- Reward model filtering

MedChatZH

- Modeling of Chinese conversational language
- Understanding questions in TCM
- High-quality conversational outputs

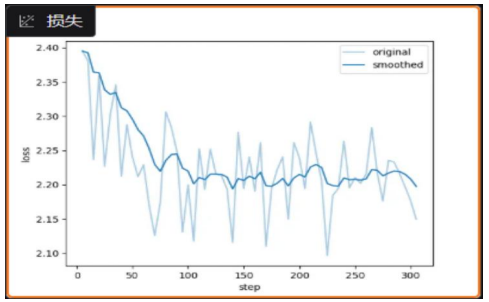



Figure 1 微调loss变化

```
{
    Epoch:3,
    N_tokens:102887824,
    Train_loss:2.154,
    Train_runtime:2:51:23.49,
    Template:"qwen3_no_think"
}
```

Figure 2 关键参数

### 不同量化方法的评估

| Model | Quantization/Format | Inference Time (s) | Perplexity |
|---|---|---|---|
| Llama-3.2-1B-Instruct-FP8-Dynamic | FP8 Dynamic | 4.3845 | 856 |
| Llama-3.2-1B-Instruct-W4A8-GPTQ | W4A8 GPTQ | 6.4398 | 664 |

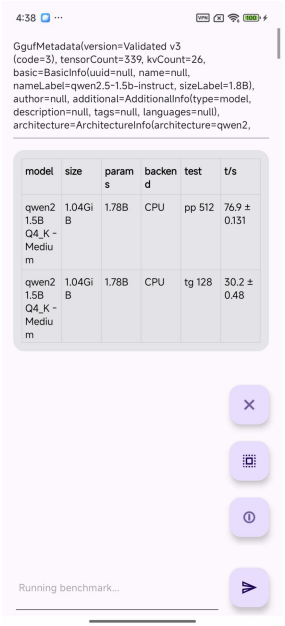


Figure 3 app界面-bench

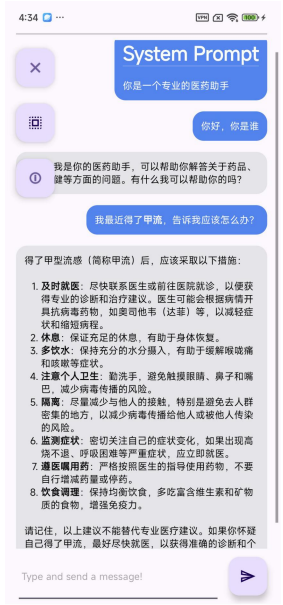



Figure 4 app界面-对话

### 对KV cache优化方法的调研与评估

| Method | Context 8192 | | Context 32768 | | Context=131072 | |
|---|---|---|---|---|---|---|
| | Lantency | MaxMem | Lantency | MaxMem | Lantency | MaxMem |
| fullkv | 0.83s | 2.44GB | 4.48s | 5.27GB | 30.25s | 16.60GB |
| streamingllm | 1.00s | 2.41GB | 4.44s | 5.15GB | 30.20s | 16.11GB |
| snapkv | 0.69s | 2.42GB | 3.62s | 5.19GB | 30.58s | 16.28G |
| h2o | 1.81s | 19.02GB | - | OOM | - | OOM |
| gemfilter | 0.53s | 2.41GB | 1.95s | 5.14GB | 14.37s | 16.06GB |
| pyramidinfer | 0.90s | 19.18GB | - | OOM | - | OOM |
| fastkv | 0.41s | 2.41GB | 2.06s | 5.15GB | 16.55s | 16.11GB |

Table 1 各方法在Llama模型上不同上下文长度的prefill延迟和最大显存占用