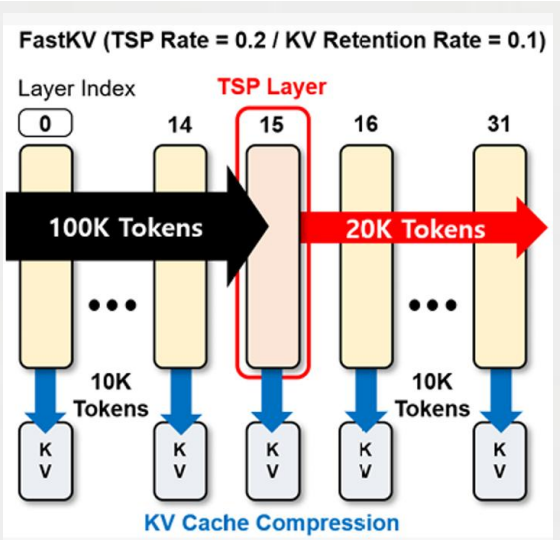
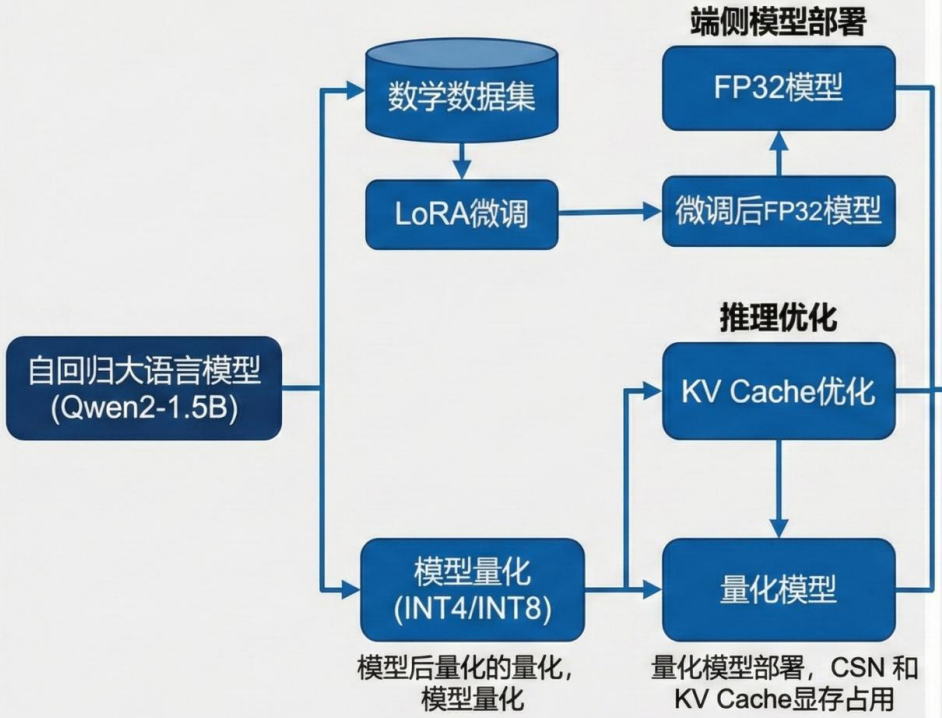
顶点课程期末展示——第六组

黄瑞翔　张翼飞　张雯幻　康硕



Method	Prefill	Decoding	Acc.
Full-context	Slow	Slow	High
StreamingLLM	Slow	Fast	Low
SnapKV	Slow	Fast	High
GemFilter	Fast	Fast	Low
FastKV	**Fast**	**Fast**	**High**

基于Qwen2-1.5B模型的部署与优化

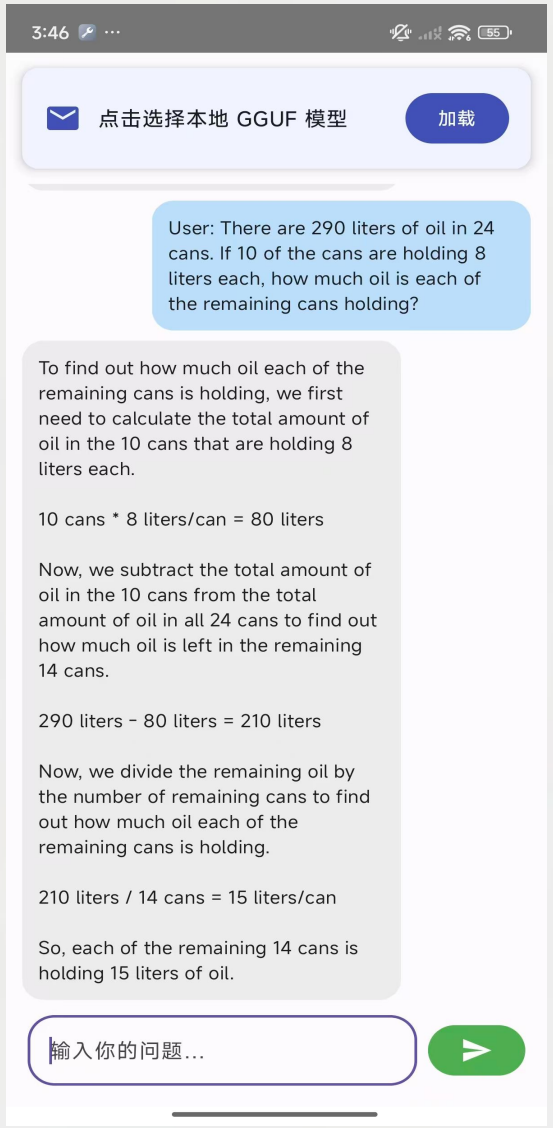


小米14手机端侧展示

LORA微调提升数学推理能力

量化+KV Cache显著降低延迟与功耗

小米14手机APP展示



指标	FP16模型	Q4KM模型	后者提升幅度
采样时间	148.48 ms	47.24 ms	68.2%更快
prompt处理	100.96 ms/token	57.25 ms/token	43.3%更快
推理速度	220.91 ms/token	97.25 ms/token	56.0%更快
总token/s	4.53	10.28	127%提升

推理速度对比

指标	FP16模型	Q4KM模型	节省幅度
主机内存	3359 MiB	1349 MiB	59.8%节省
模型内存	2944 MiB	934 MiB	68.3%节省

内存使用对比