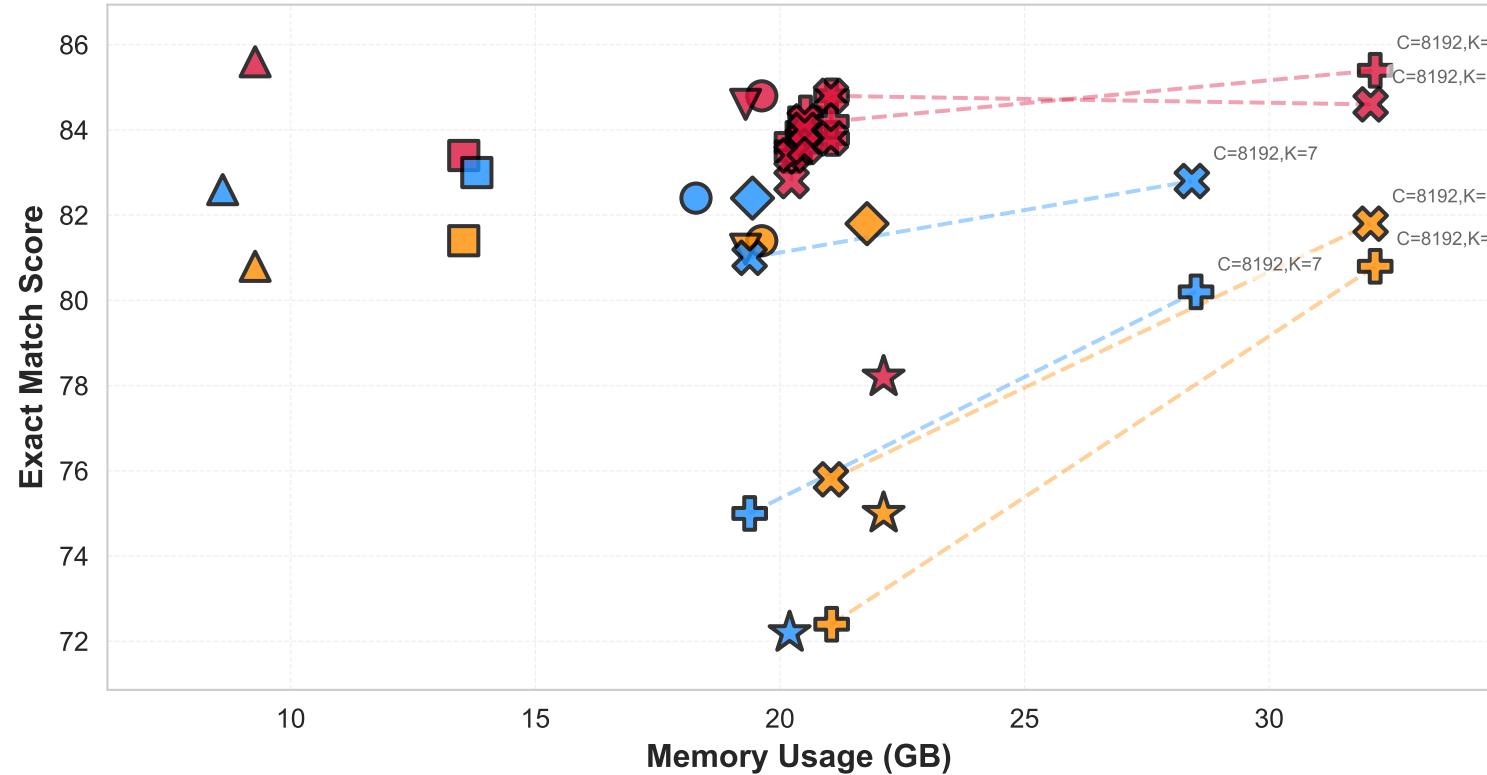


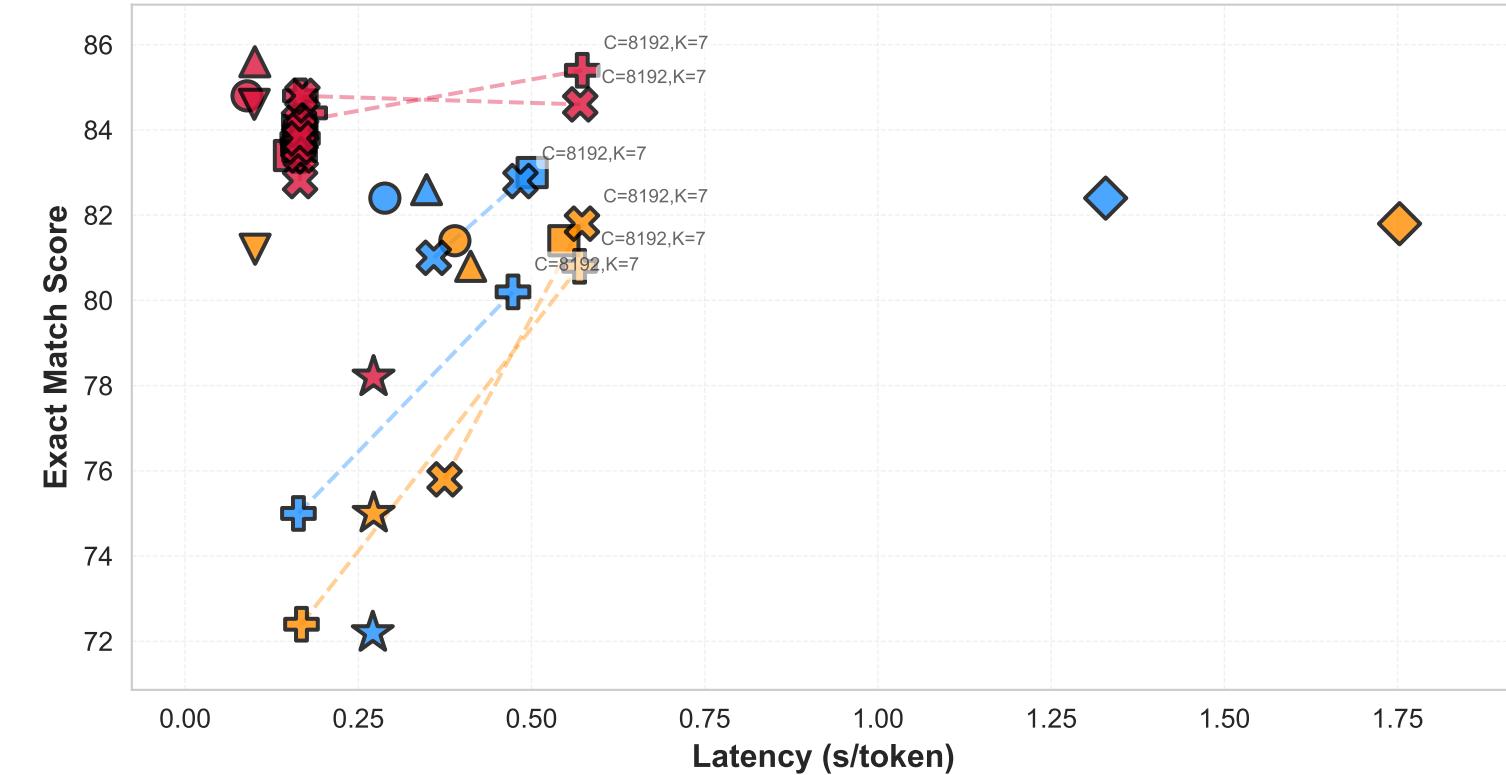
ICL Banking Task Performance Analysis (16K Context)

Memory and Latency vs Exact Match

Memory vs Performance



Throughput vs Performance



Models:

- Llama-3.1-8B-Instruct
- Qwen2.5-7B-Instruct

DeepSeek-R1-Distill-Llama-8B
DeepSeek-R1-Distill-Qwen-7B

- Qwen3-8B

Techniques:

- | | | | |
|---------------|----------|------------|--------------|
| Yarn-Qwen3-8B | baseline | minference | streamingllm |
| INT8 | INT4 | pyramidkv | duoattn |
| INT4 | snapkv | | |