



Models

- Llama-3.1-8B
- Qwen2.5-7B
- R1-Distill-Llama-8B
- R1-Distill-Qwen-7B

Techniques

- Baseline
- NF4
- Int8
- SnapKV
- PyramidKV
- StreamingLLM
- DuoAttn