

vLLM, $\lambda = 0.5$, NVMe

