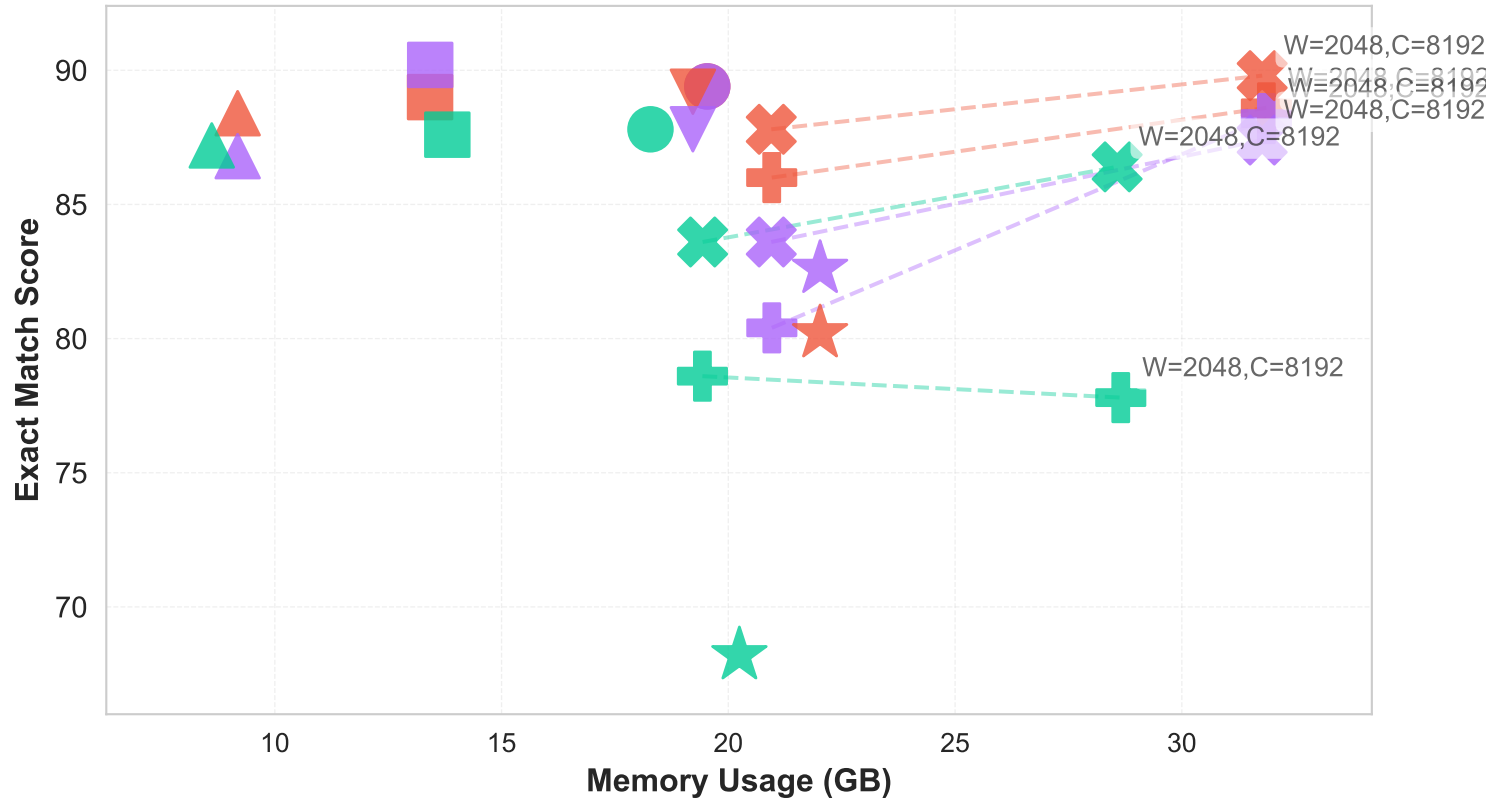


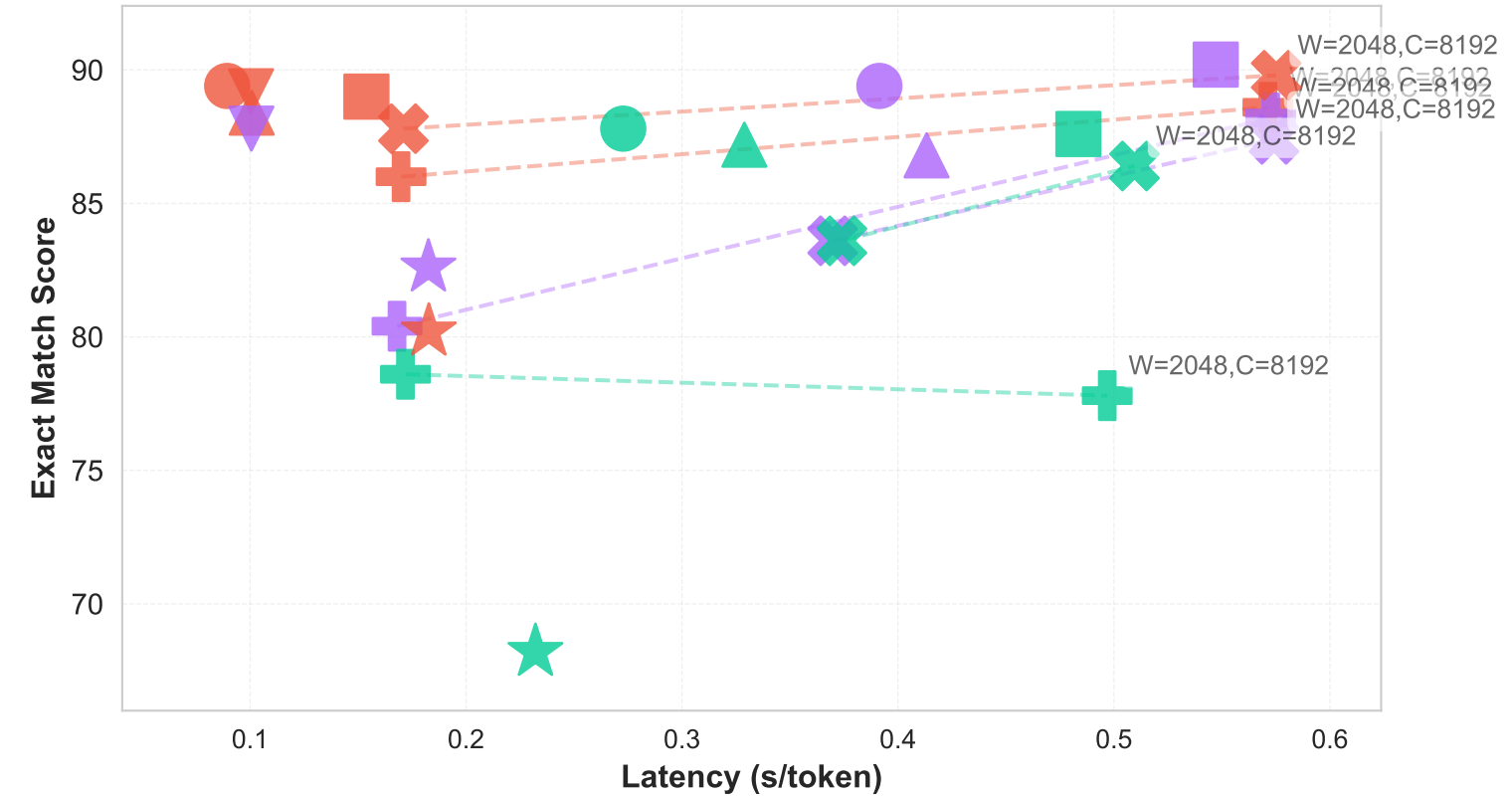
# Reasoning Models Recover Performance Degradation Compared to Non-Reasoning Models on In-Context Learning Tasks

## HELMET In-Context Learning Task: Clinic150 (16K Context)

### Performance vs Memory



### Performance vs Latency



Legend: Llama-3.1-8B-Instruct (Purple Circle), Qwen2.5-7B-Instruct (Teal Circle), DeepSeek-R1-Distill-Llama-8B (Red Circle), Baseline (Grey Circle), Int8 (Grey Square), NF4 (Teal Triangle), PyramidKV (Grey Plus), SnapKV (Grey X), StreamingLLM (Grey Star), DuoAttn (Grey Inverted Triangle)