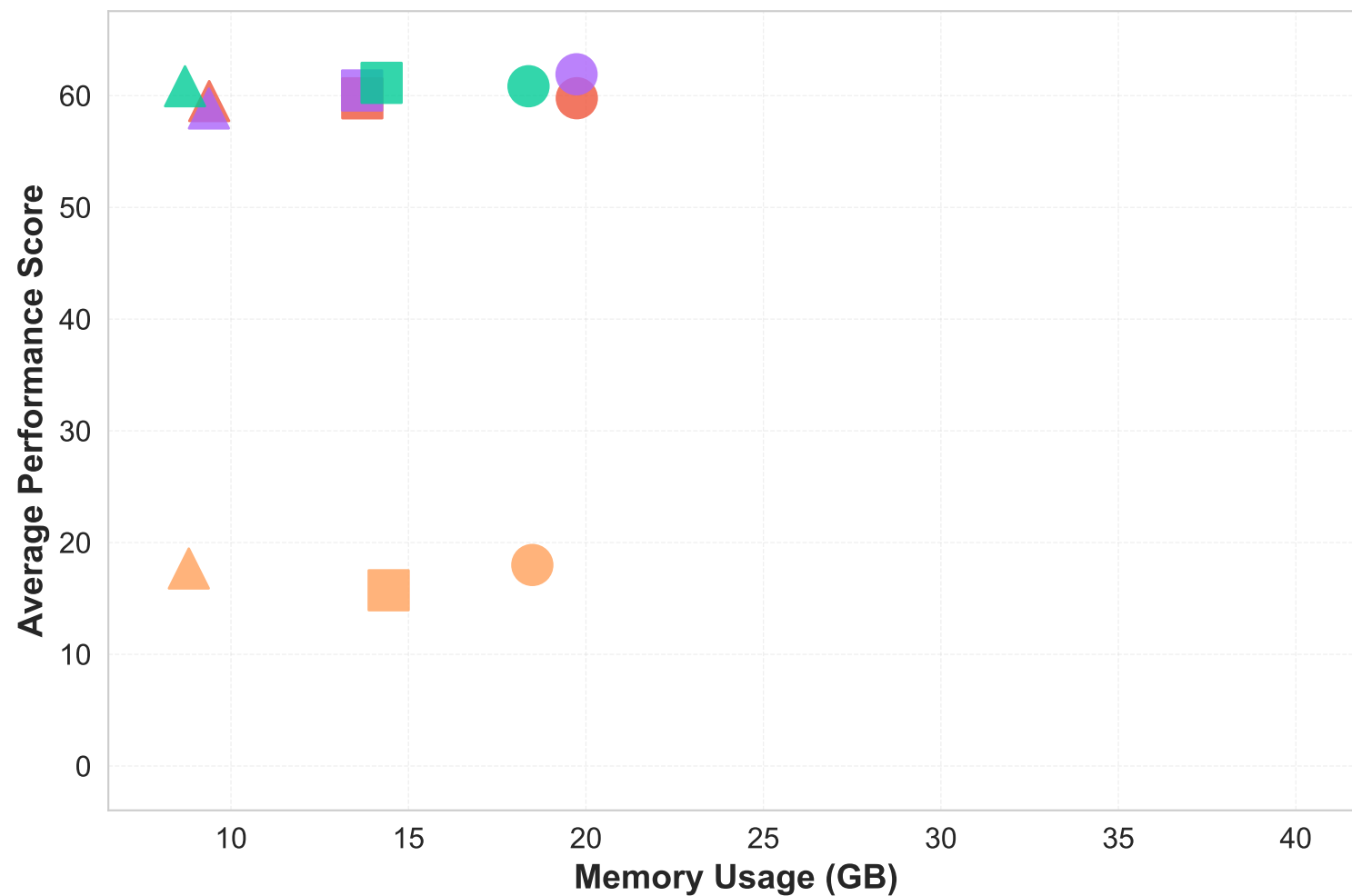


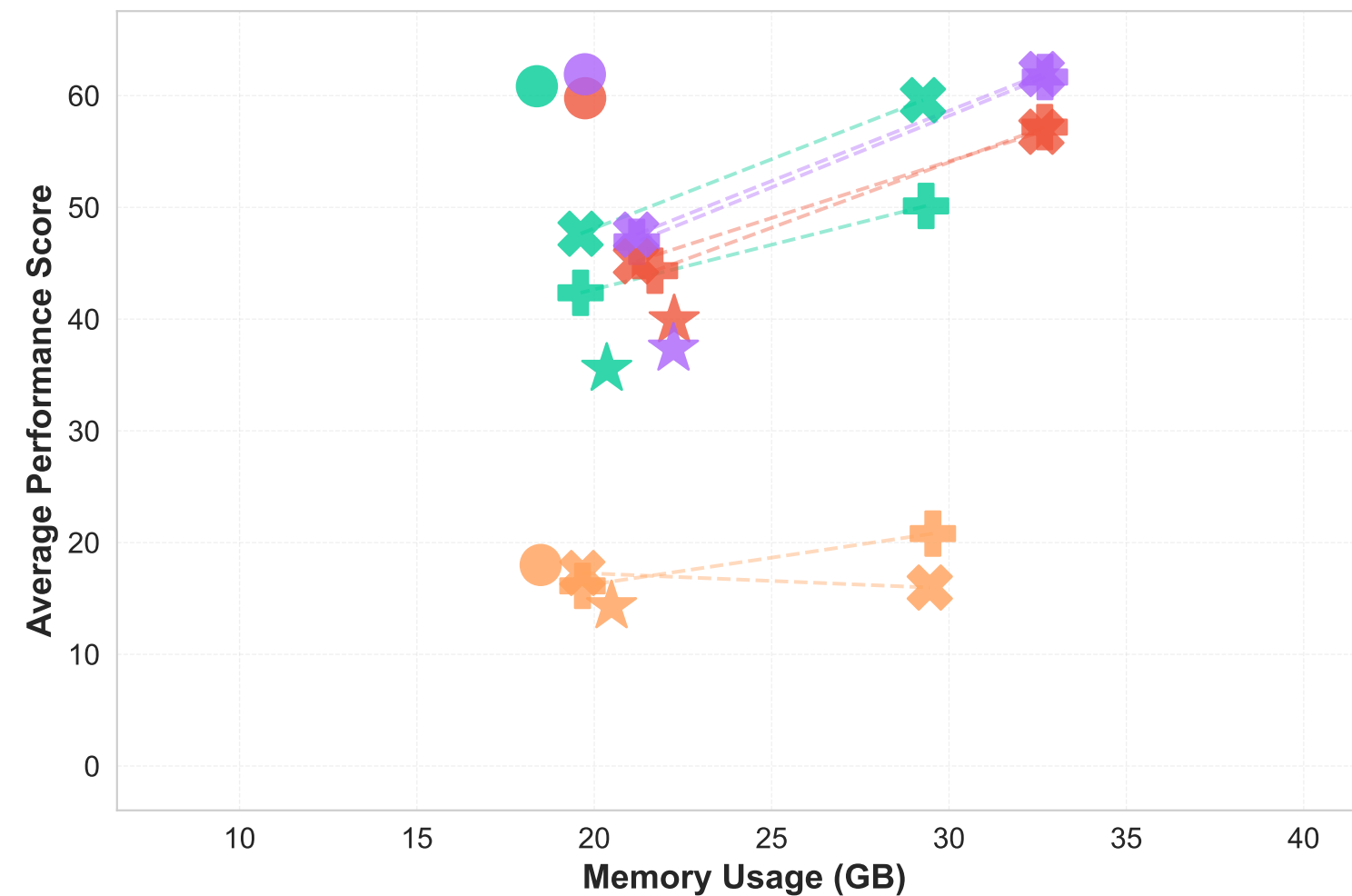
Model Comparison: Average Performance vs Memory Usage

HELMET (16K) and LongProc (2K) Benchmarks

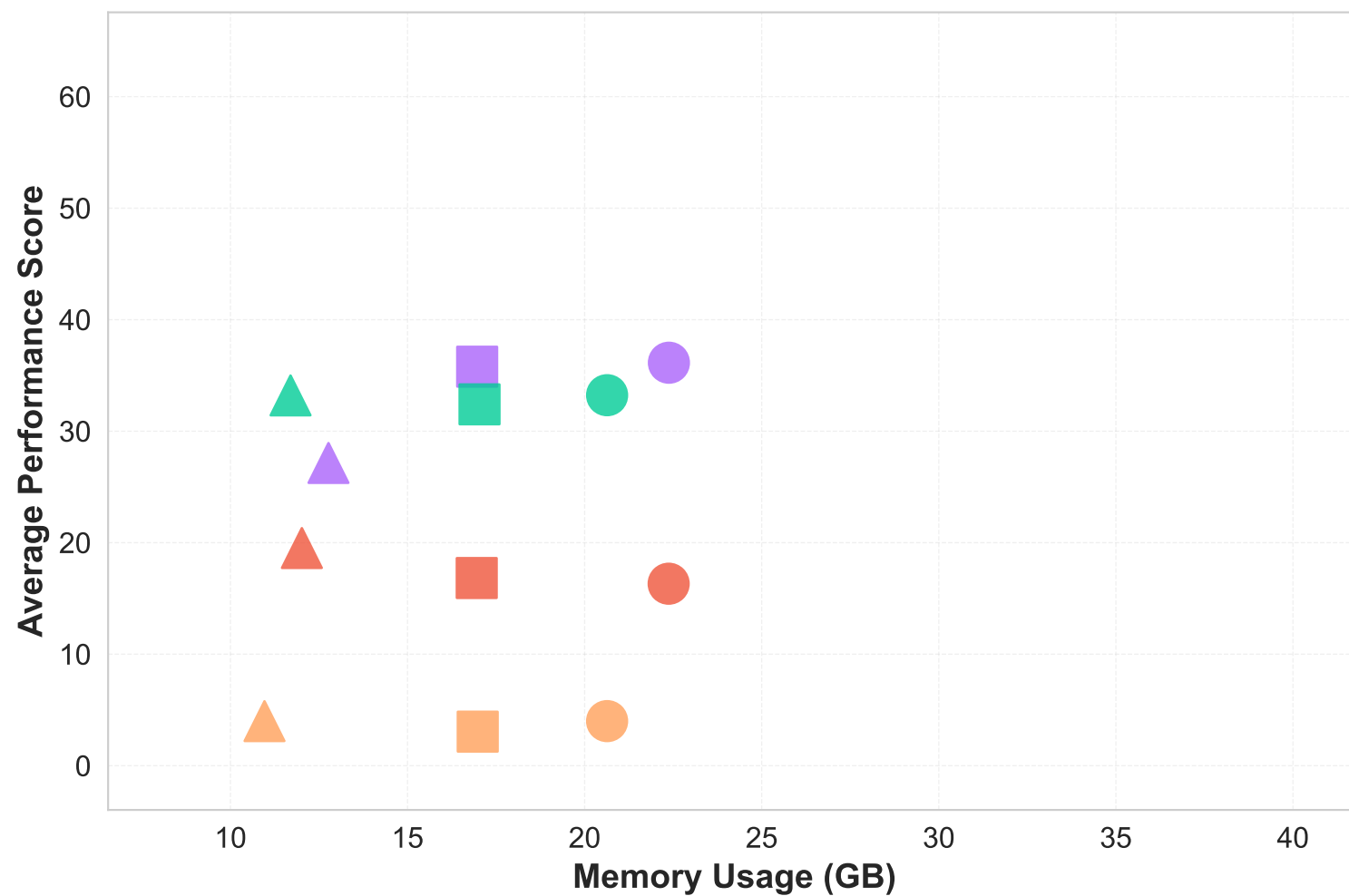
HELMET: Quantization Methods



HELMET: Token Eviction Methods



LongProc: Quantization Methods



LongProc: Token Eviction Methods

