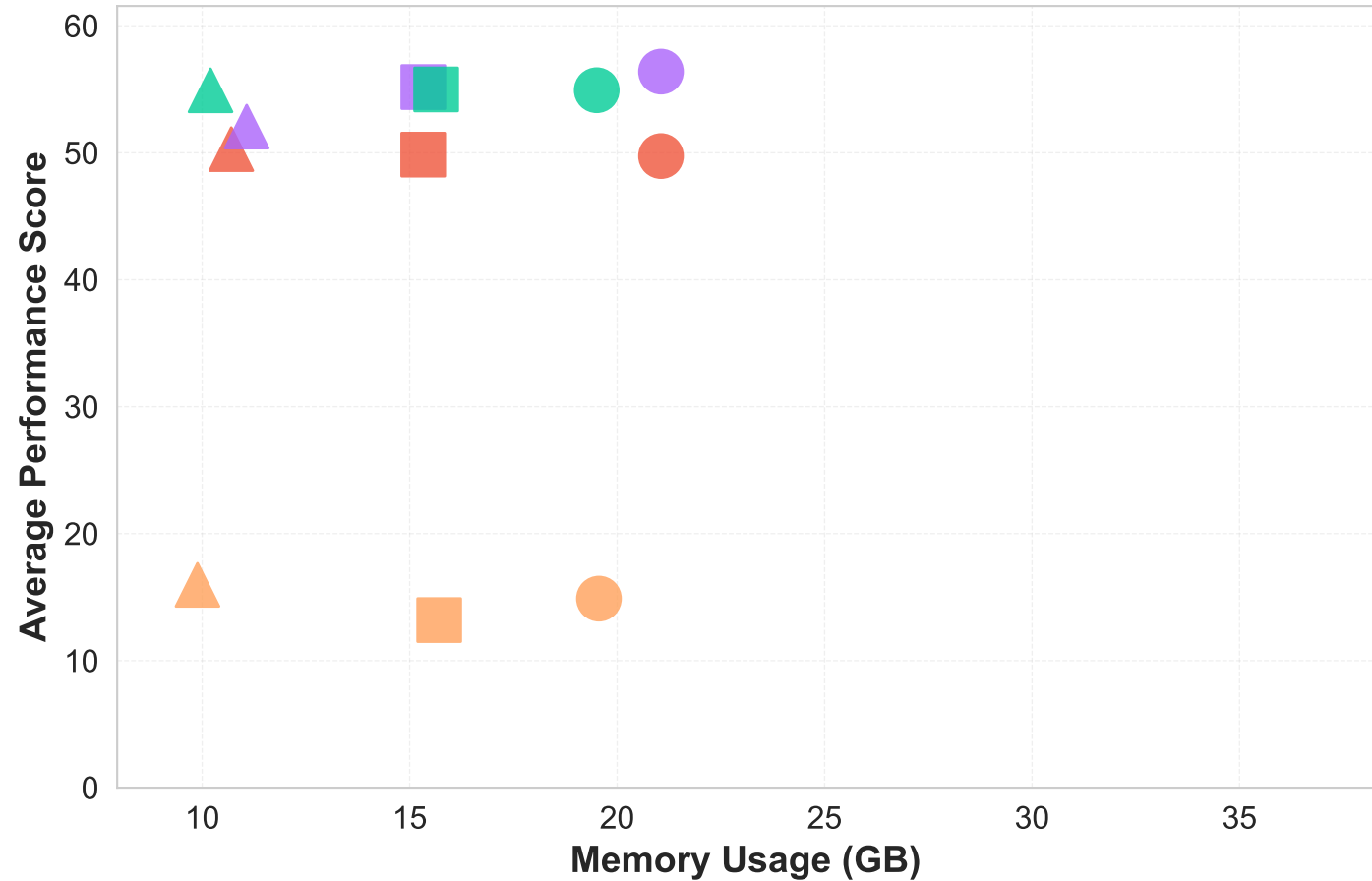


# Model Comparison: Average Performance vs Memory Usage

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

Quantization Methods



Token Eviction Methods

