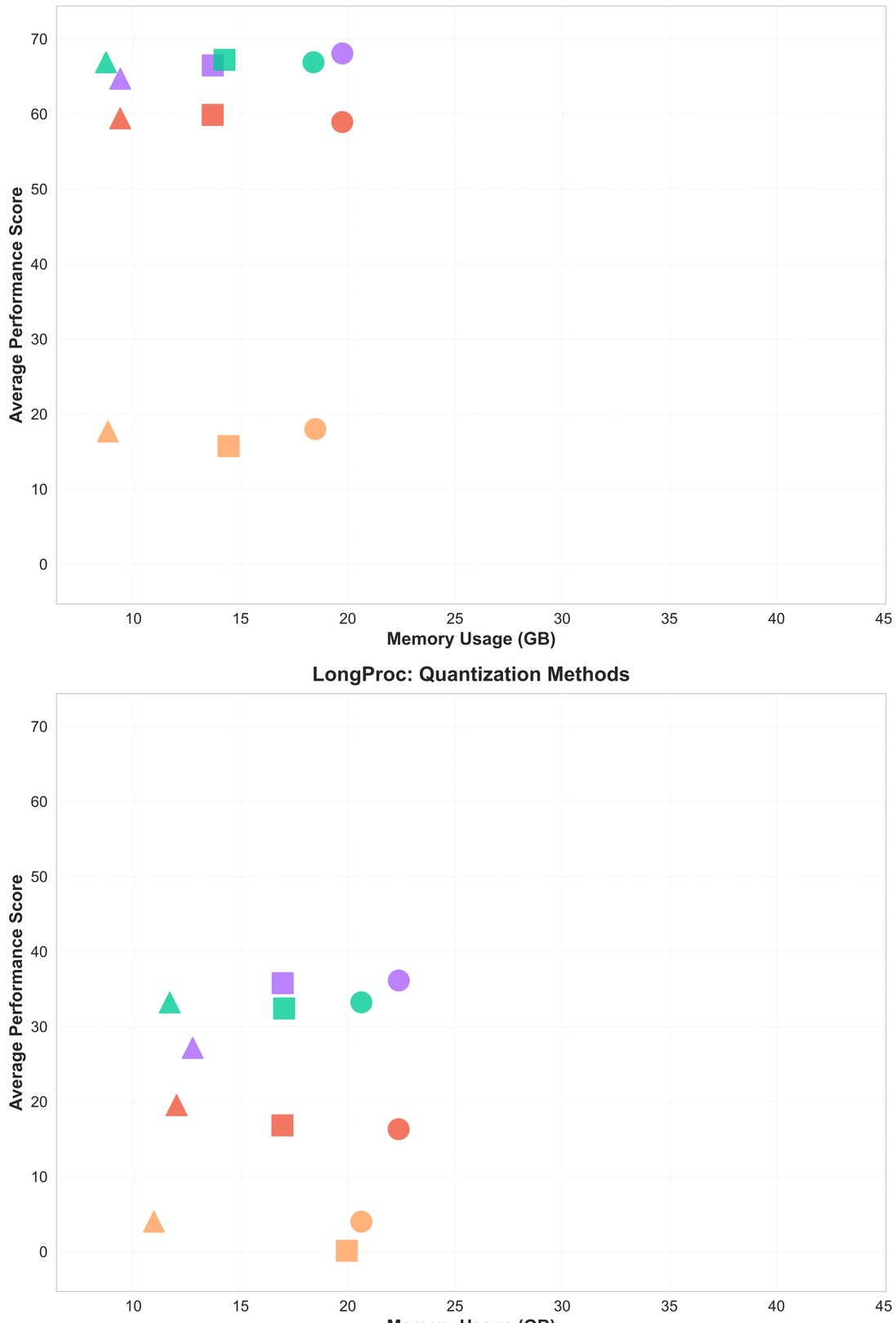
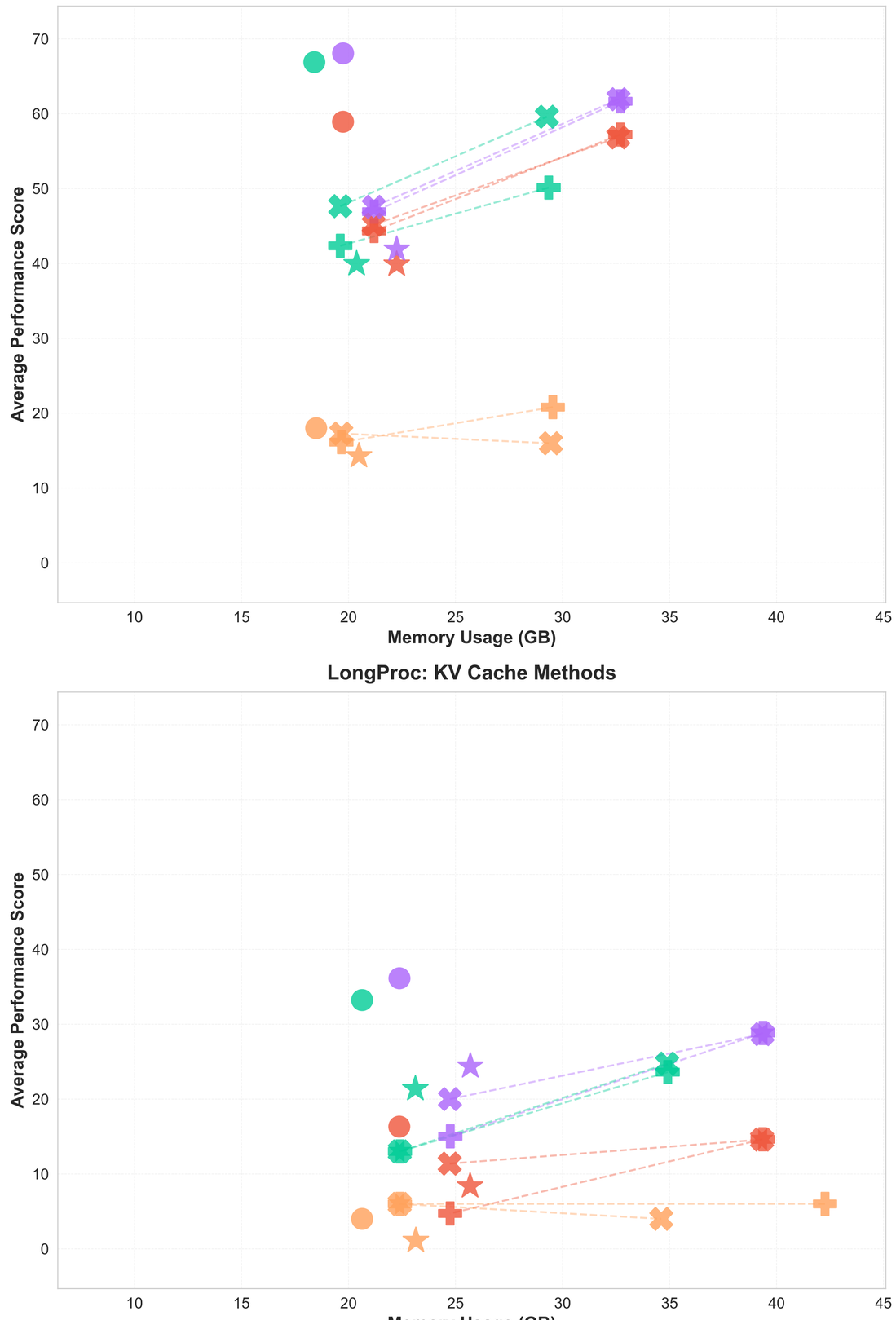
Model Comparison: Average Performance vs Memory Usage
HELMET (16K) and LongProc (2K) Benchmarks

HELMET: Quantization Methods



HELMET: KV Cache Methods



LongProc: Quantization Methods

LongProc: KV Cache Methods