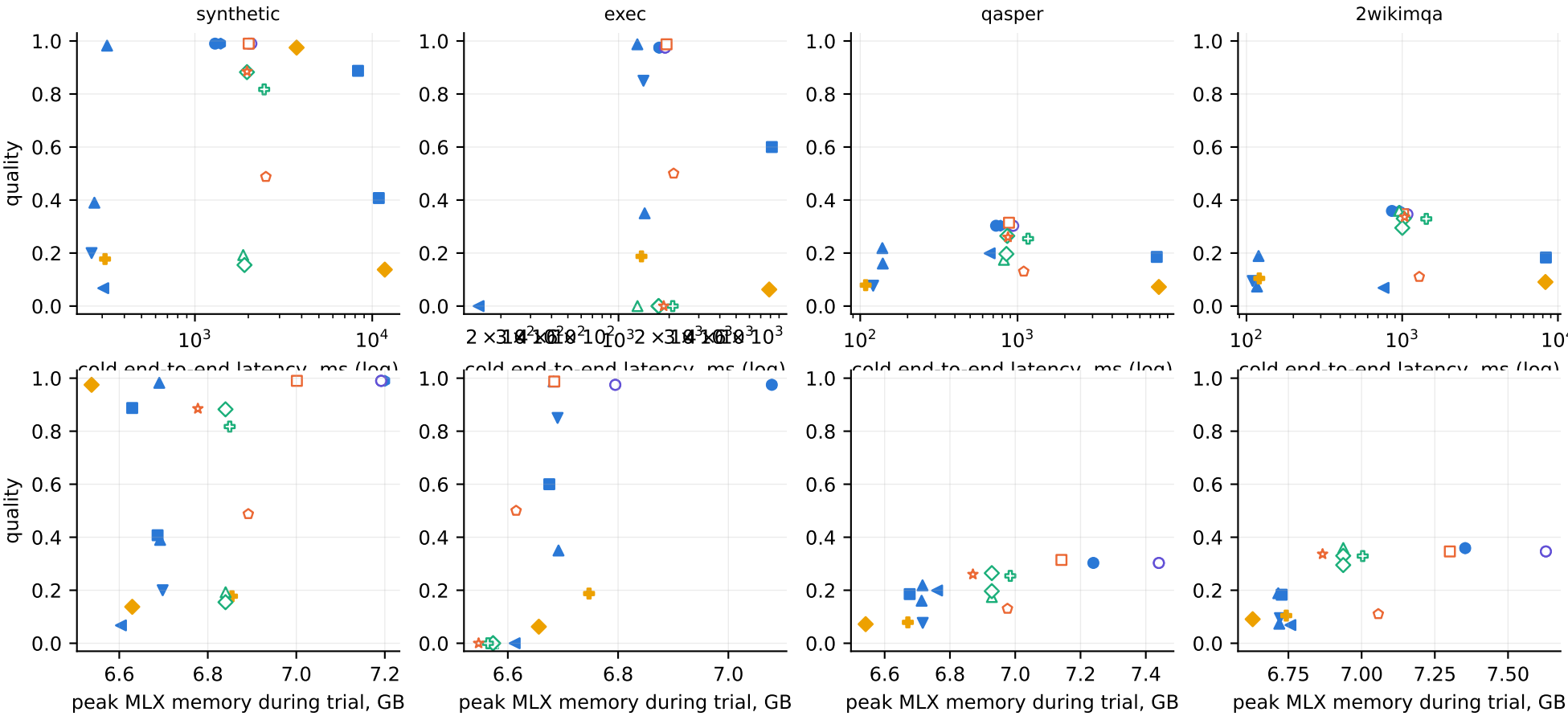

Quality vs cold latency (top) and peak memory (bottom); all model pairs pooled
synthetic
exec
qasper
2wikimqa
quality
cold end-to-end latency, ms (log)
peak MLX memory during trial, GB