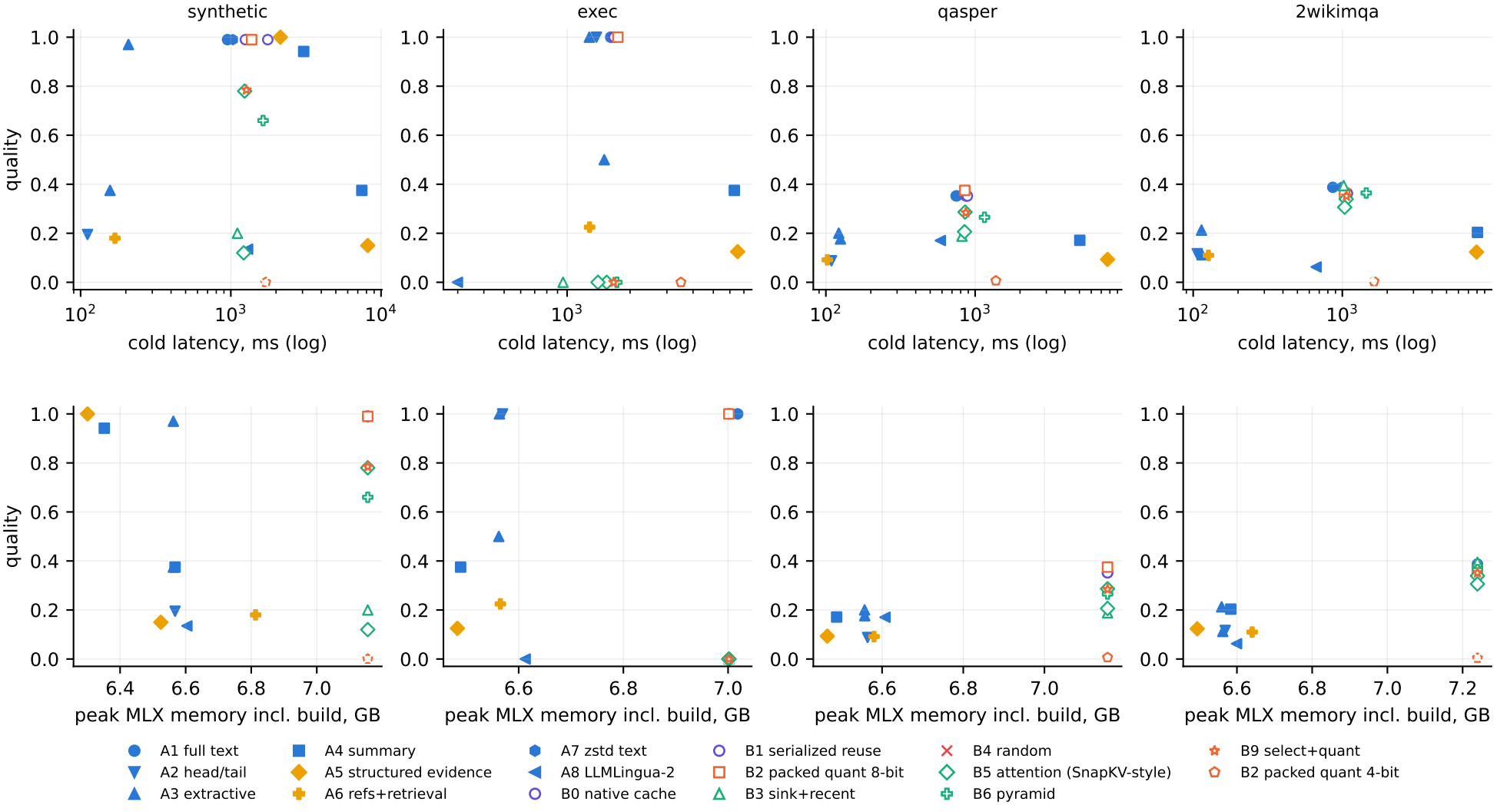

Quality vs cold latency (top) and peak MLX memory (bottom) - Qwen2.5-3B pair on the laptop (models resident: ~6.5 GB baseline)
synthetic
exec
qasper
2wikimqa
quality
cold latency, ms (log)
peak MLX memory incl. build, GB
A1 full text
A2 head/tail
A3 extractive
A4 summary
A5 structured evidence
A6 refs+retrieval
A7 zstd text
A8 LLMLingua-2
B0 native cache
B1 serialized reuse
B2 packed quant 8-bit
B3 sink+recent
B4 random
B5 attention (SnapKV-style)
B6 pyramid
B9 select+quant
B2 packed quant 4-bit