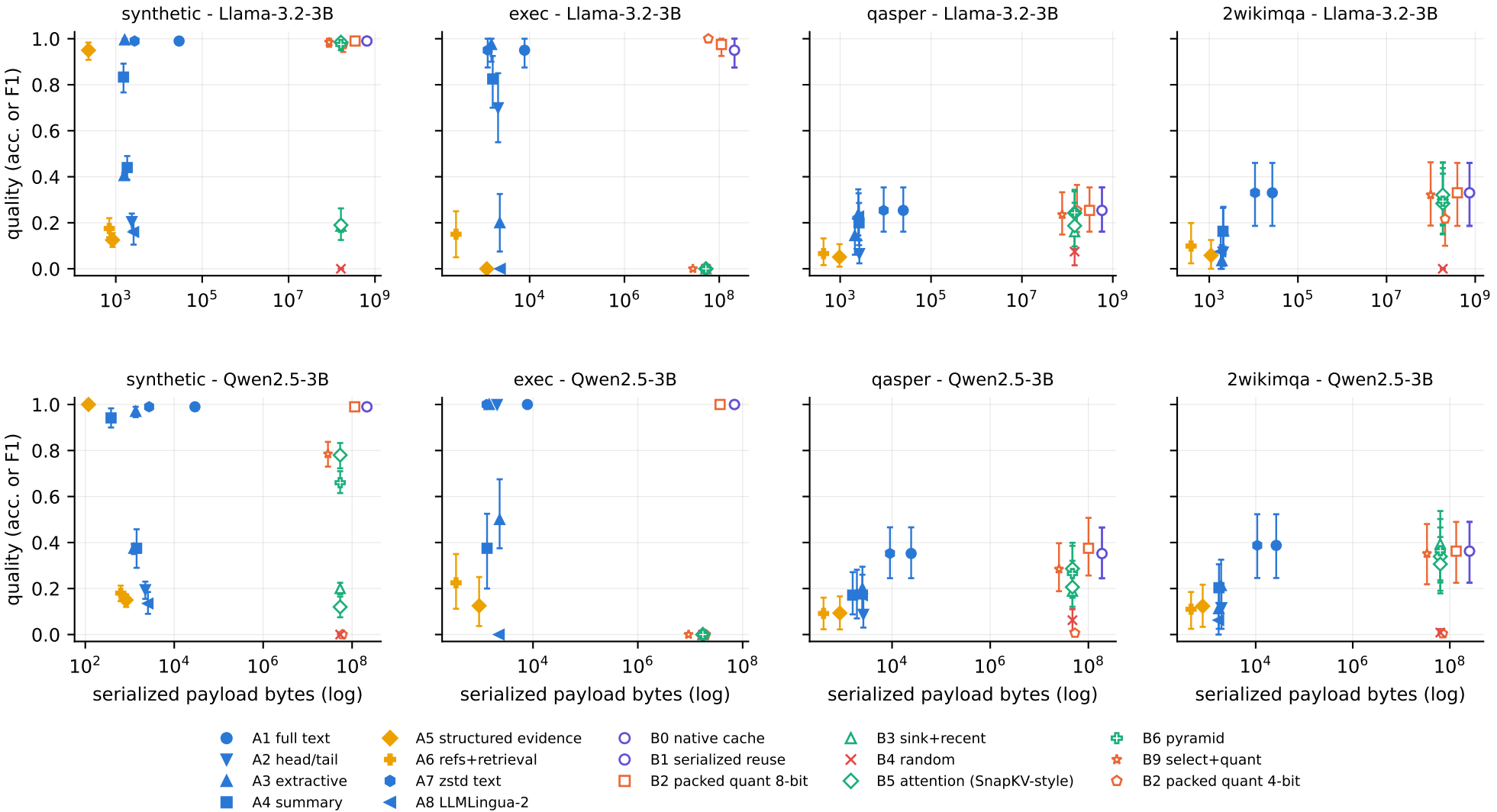

Quality vs measured serialized payload bytes (medians; 95% cluster-bootstrap CIs)
synthetic - Llama-3.2-3B
exec - Llama-3.2-3B
qasper - Llama-3.2-3B
2wikimqa - Llama-3.2-3B
synthetic - Qwen2.5-3B
exec - Qwen2.5-3B
qasper - Qwen2.5-3B
2wikimqa - Qwen2.5-3B
quality (acc. or F1)
serialized payload bytes (log)
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