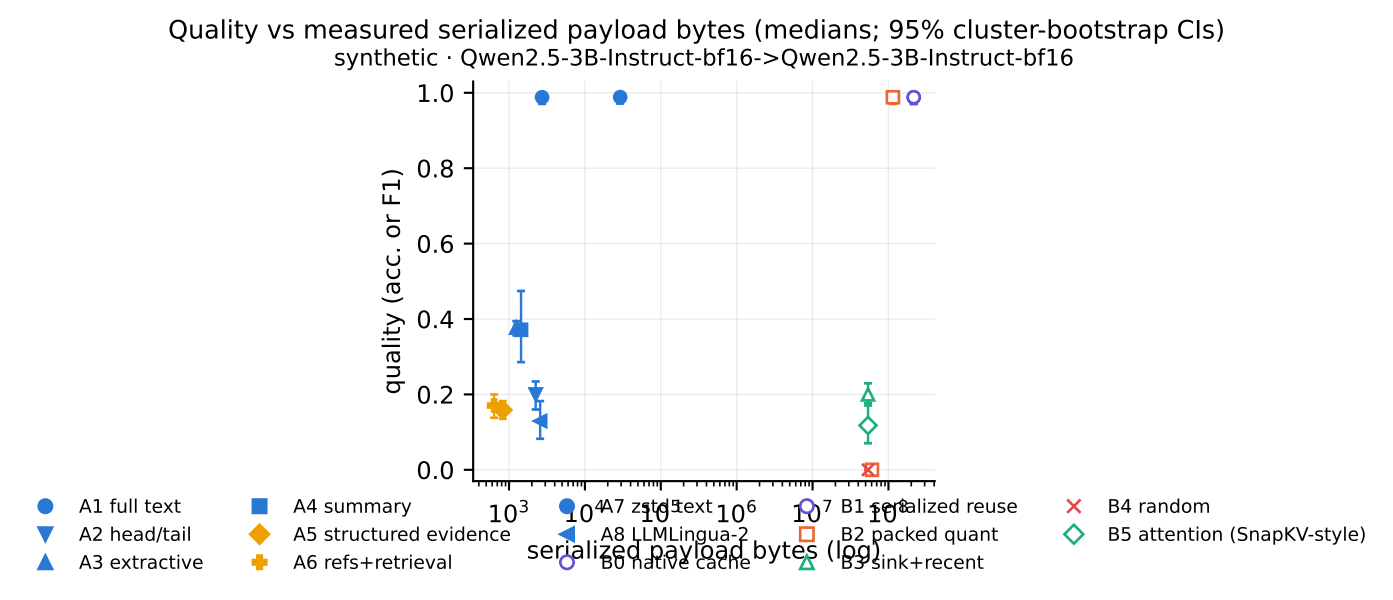

Quality vs measured serialized payload bytes (medians; 95% cluster-bootstrap CIs)
synthetic · Qwen2.5-3B-Instruct-bf16->Qwen2.5-3B-Instruct-bf16
quality (acc. or F1)
1.0
0.8
0.6
0.4
0.2
0.0
10^3
10^4
10^5
10^6
10^7
10^8
serialized payload bytes (log)
A1 full text
A2 head/tail
A3 extractive
A4 summary
A5 structured evidence
A6 refs+retrieval
A7 zstd-text
A8 LLMLingua-2
B0 native cache
B1 serialized reuse
B2 packed quant
B3 sink+recent
B4 random
B5 attention (SnapKV-style)