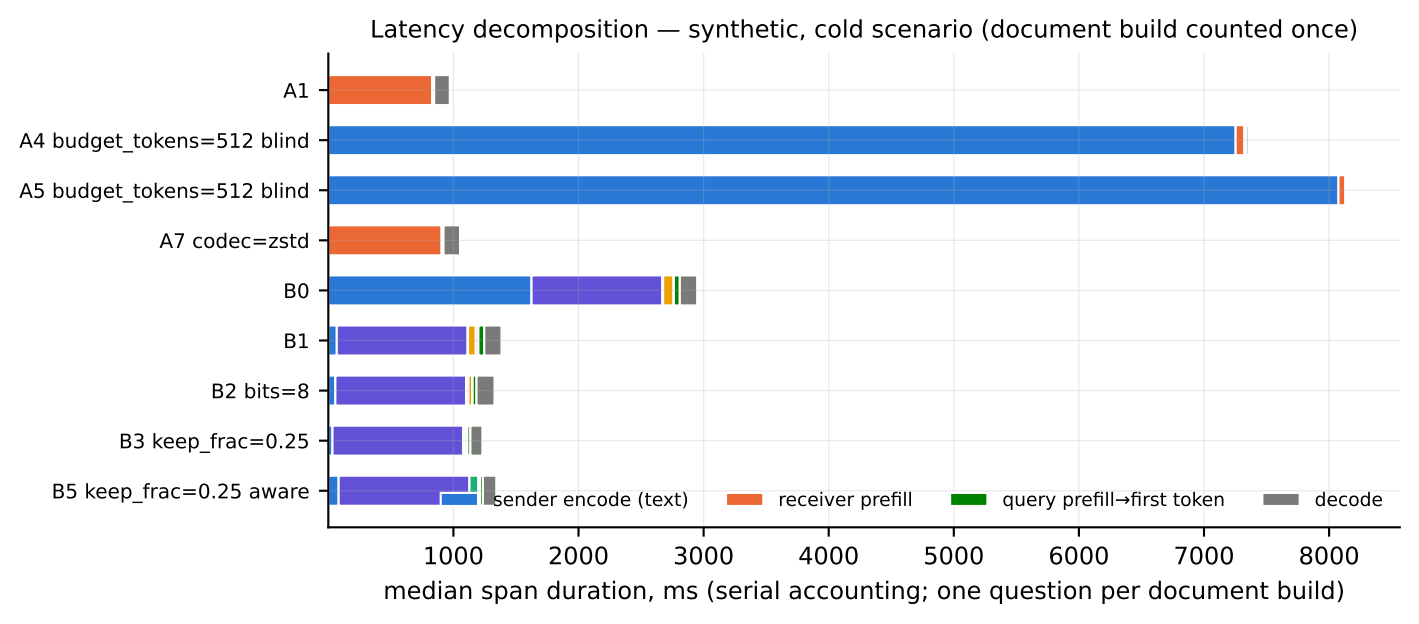

Latency decomposition — synthetic, cold scenario (document build counted once)
A1
A4 budget_tokens=512 blind
A5 budget_tokens=512 blind
A7 codec=zstd
B0
B1
B2 bits=8
B3 keep_frac=0.25
B5 keep_frac=0.25 aware
sender encode (text)
receiver prefill
query prefill→first token
decode
1000
2000
3000
4000
5000
6000
7000
8000
median span duration, ms (serial accounting; one question per document build)