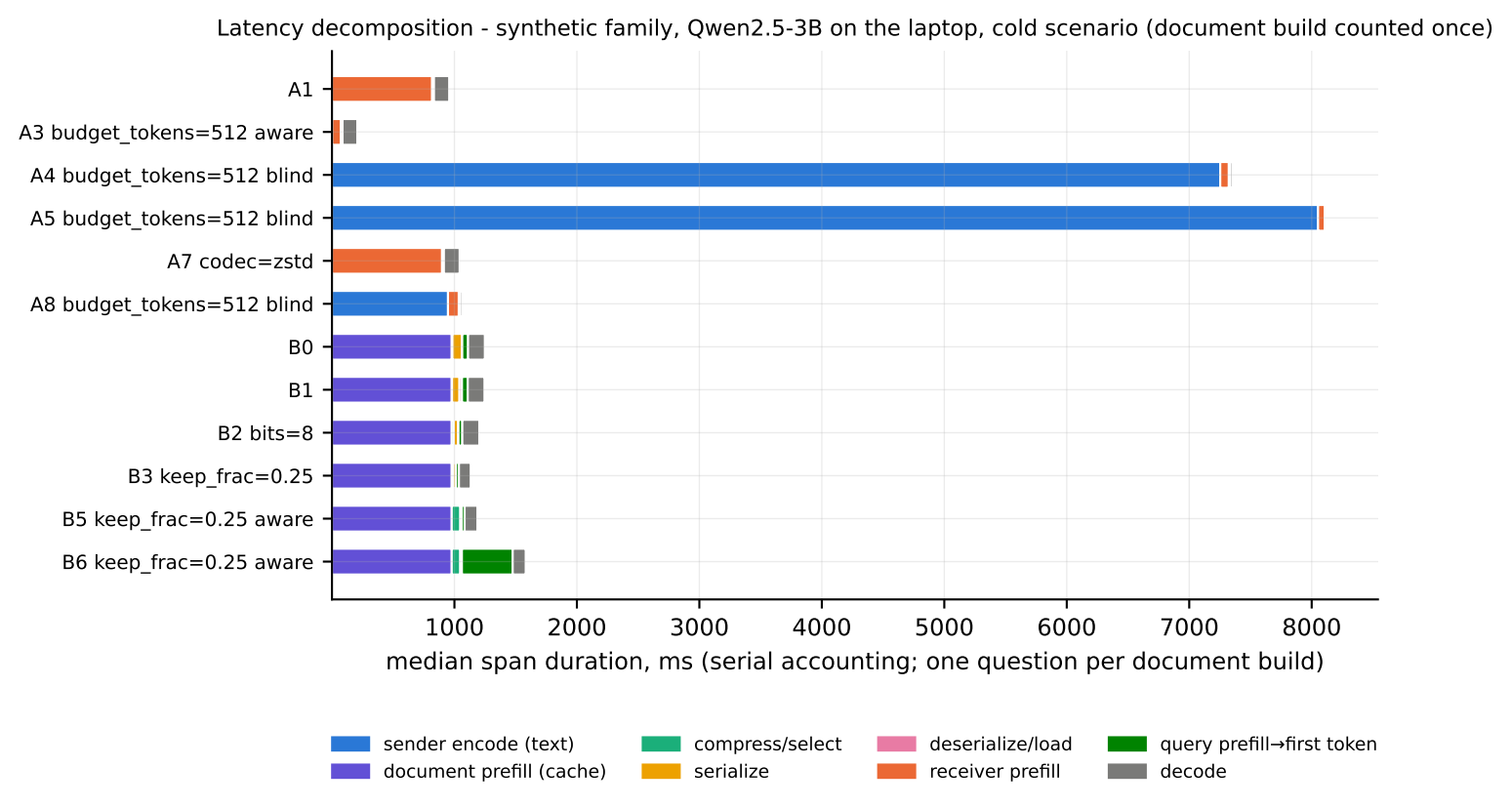

Latency decomposition - synthetic family, Qwen2.5-3B on the laptop, cold scenario (document build counted once)
A1
A3 budget_tokens=512 aware
A4 budget_tokens=512 blind
A5 budget_tokens=512 blind
A7 codec=zstd
A8 budget_tokens=512 blind
B0
B1
B2 bits=8
B3 keep_frac=0.25
B5 keep_frac=0.25 aware
B6 keep_frac=0.25 aware
1000
2000
3000
4000
5000
6000
7000
8000
median span duration, ms (serial accounting; one question per document build)
sender encode (text)
document prefill (cache)
compress/select
serialize
deserialize/load
receiver prefill
query prefill→first token
decode