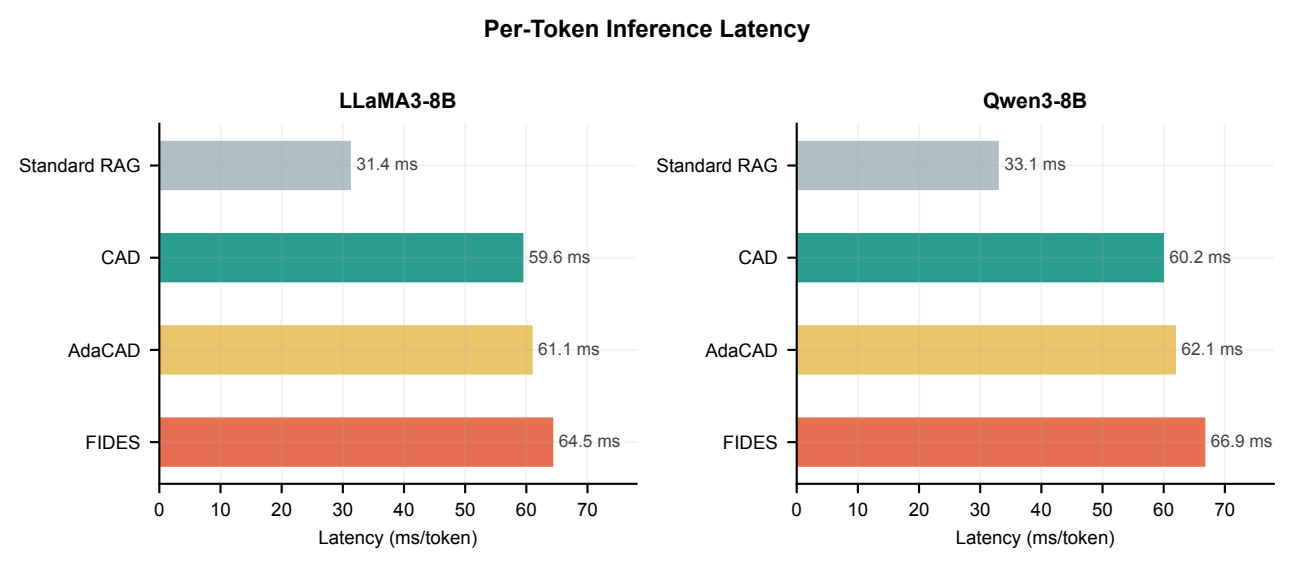
Per-Token Inference Latency
LLaMA3-8B
Standard RAG
CAD
AdaCAD
FIDES
31.4 ms
59.6 ms
61.1 ms
64.5 ms
0
10
20
30
40
50
60
70
Latency (ms/token)
Qwen3-8B
Standard RAG
CAD
AdaCAD
FIDES
33.1 ms
60.2 ms
62.1 ms
66.9 ms
0
10
20
30
40
50
60
70
Latency (ms/token)