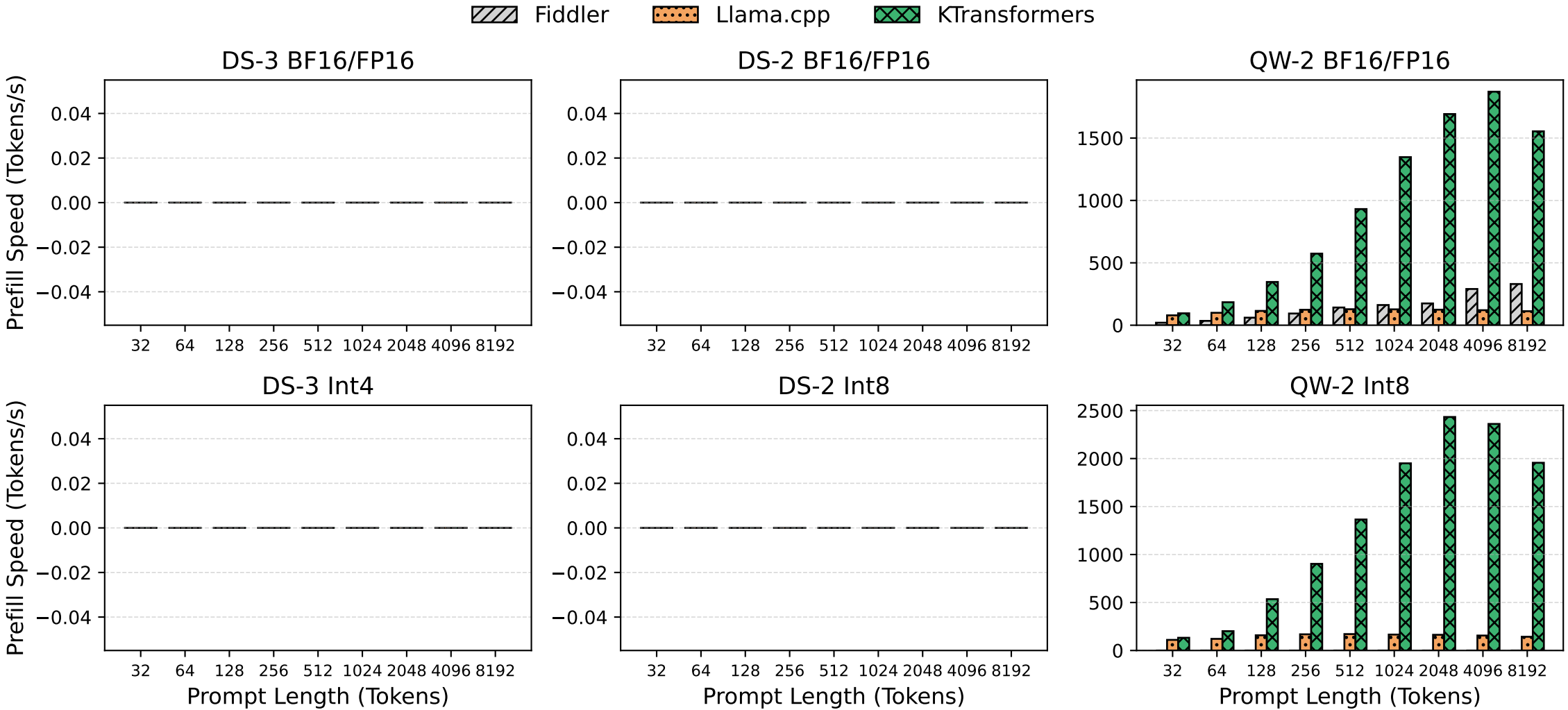
Fiddler
Llama.cpp
KTransformers
DS-3 BF16/FP16
DS-2 BF16/FP16
QW-2 BF16/FP16
DS-3 Int4
DS-2 Int8
QW-2 Int8
Prefill Speed (Tokens/s)
Prefill Speed (Tokens/s)
0.04
0.02
0.00
−0.02
−0.04
0
500
1000
1500
2000
2500
32
64
128
256
512
1024
2048
4096
8192
Prompt Length (Tokens)
Prompt Length (Tokens)
Prompt Length (Tokens)