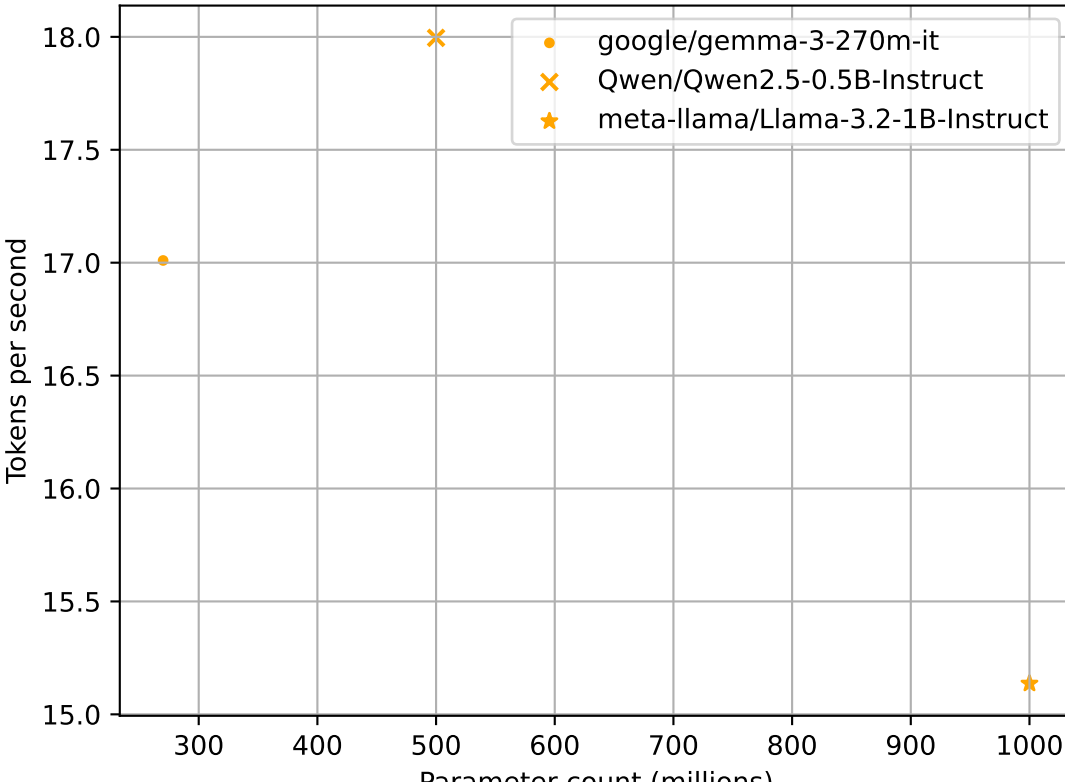
Tokens per second vs. parameter count (millions)
google/gemma-3-270m-it
Qwen/Qwen2.5-0.5B-Instruct
meta-llama/Llama-3.2-1B-Instruct
Tokens per second
Parameter count (millions)
18.0
17.5
17.0
16.5
16.0
15.5
15.0
300
400
500
600
700
800
900
1000