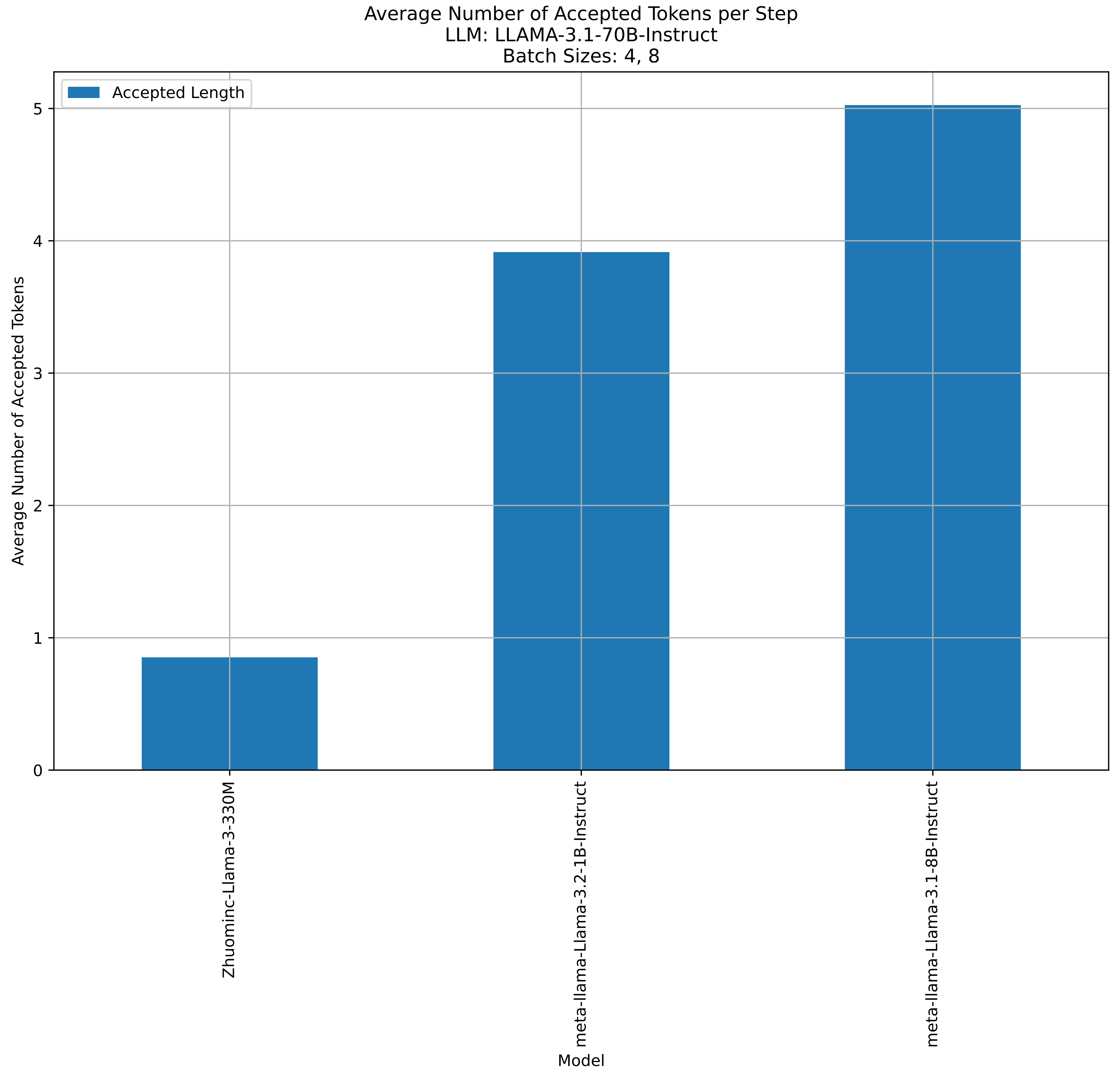
Average Number of Accepted Tokens per Step
LLM: LLAMA-3.1-70B-Instruct
Batch Sizes: 4, 8
Accepted Length
Average Number of Accepted Tokens
0
1
2
3
4
5
Zhuominc-Llama-3-330M
meta-llama-Llama-3.2-1B-Instruct
meta-llama-Llama-3.1-8B-Instruct
Model