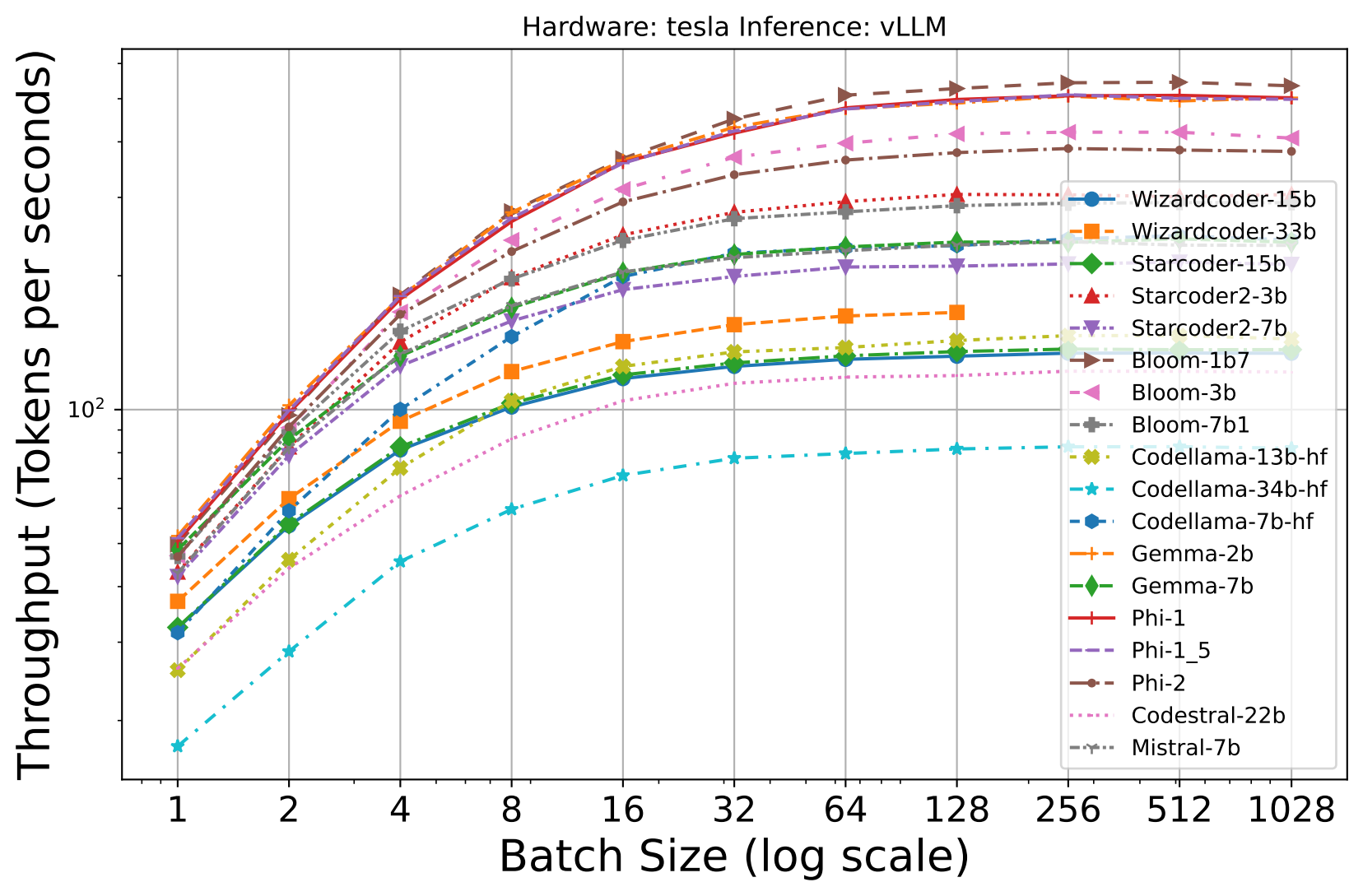
Hardware: tesla Inference: vLLM
Throughput (Tokens per seconds)
10^2
1
2
4
8
16
32
64
128
256
512
1028
Batch Size (log scale)
Wizardcoder-15b
Wizardcoder-33b
Starcoder-15b
Starcoder2-3b
Starcoder2-7b
Bloom-1b7
Bloom-3b
Bloom-7b1
Codellama-13b-hf
Codellama-34b-hf
Codellama-7b-hf
Gemma-2b
Gemma-7b
Phi-1
Phi-1_5
Phi-2
Codestral-22b
Mistral-7b