Inference speed of 1B model

Legend:
- SpikingBrain-1B (prefill: 273.31 tokens/s)
- Llama3.2-1B (prefill: 239.42 tokens/s)
- SpikingBrain/Llama

X-axis: Output Length (K tokens)
Y-axis (left): Decode Speed (tokens/s)
Y-axis (right): Speed Ratio (SpikingBrain/Llama)

Data labels: 48.71, 48.79, 20.10, 3.17, 15.39x, 7.52x, 4.04x, 2.42x