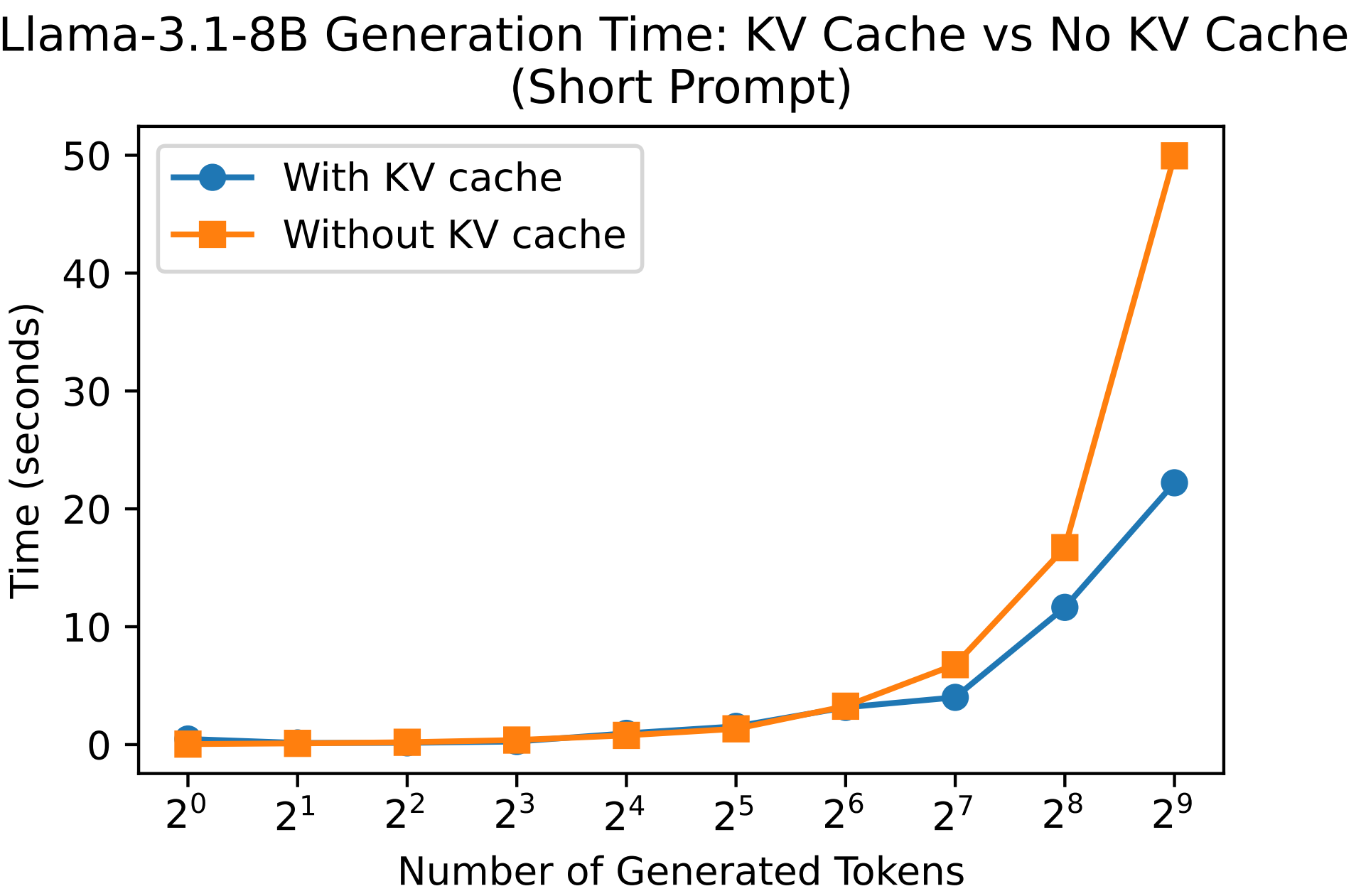
Llama-3.1-8B Generation Time: KV Cache vs No KV Cache
(Short Prompt)
With KV cache
Without KV cache
Time (seconds)
0
10
20
30
40
50
2^0
2^1
2^2
2^3
2^4
2^5
2^6
2^7
2^8
2^9
Number of Generated Tokens