Llama-3.2-1B Generation Time: KV Cache vs No KV Cache (Short Prompt)

Legend:
- With KV cache
- Without KV cache

X-axis: Number of Generated Tokens (2^0 to 2^9)
Y-axis: Time (seconds)