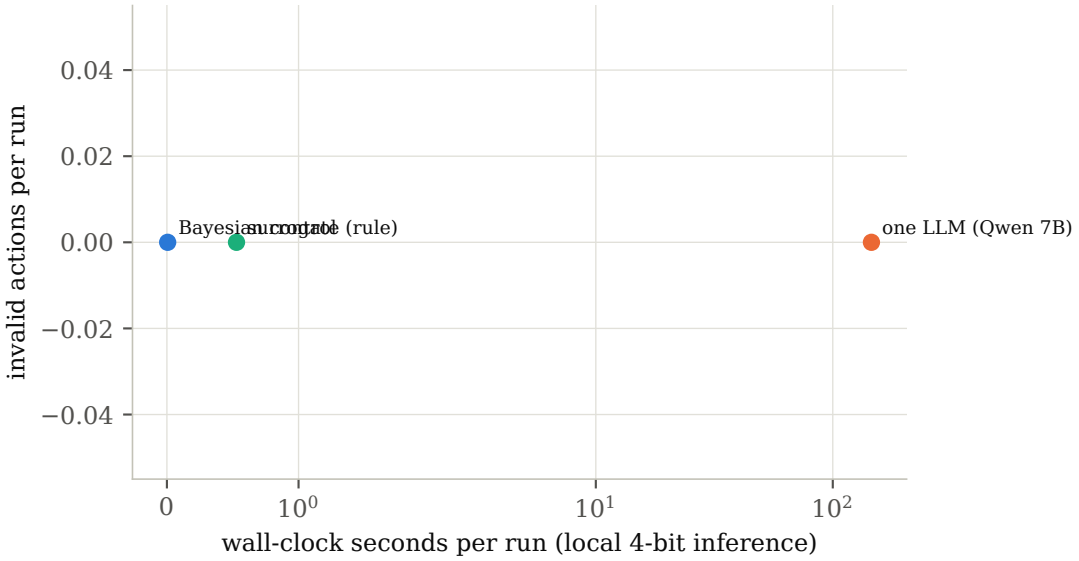

Inference cost frontier
0.04
0.02
0.00
−0.02
−0.04
invalid actions per run
Bayesian surrogate (rule)
one LLM (Qwen 7B)
0
10^0
10^1
10^2
wall-clock seconds per run (local 4-bit inference)