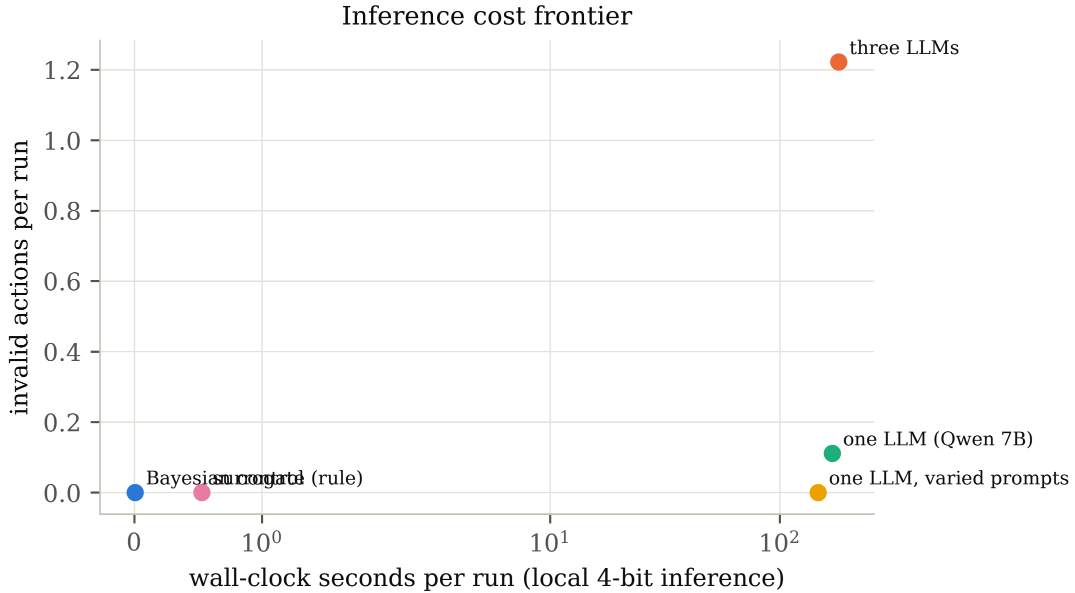
Inference cost frontier
invalid actions per run
1.2
1.0
0.8
0.6
0.4
0.2
0.0
three LLMs
one LLM (Qwen 7B)
Bayesian control
surrogate (rule)
one LLM, varied prompts
0
10^0
10^1
10^2
wall-clock seconds per run (local 4-bit inference)