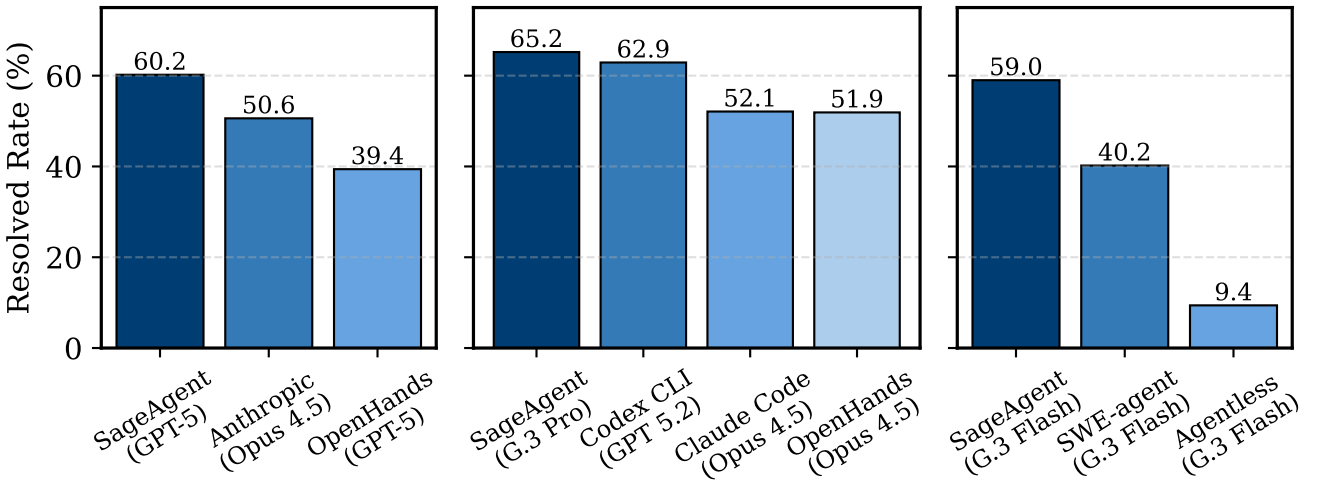
CyberGym
Terminal-Bench 2.0
SWE-Bench Pro (Python)
Resolved Rate (%)
0
20
40
60
60.2
50.6
39.4
65.2
62.9
52.1
51.9
59.0
40.2
9.4
SageAgent (GPT-5)
Anthropic (Opus 4.5)
OpenHands (GPT-5)
SageAgent (G.3 Pro)
Codex CLI (GPT 5.2)
Claude Code (Opus 4.5)
OpenHands (Opus 4.5)
SageAgent (G.3 Flash)
SWE-agent (G.3 Flash)
Agentless (G.3 Flash)