Benchmark	Domain	Key metric	Open-source
AgentBench	OS, database, web navigation	End-to-end task success rate	Yes
Terminal-Bench	Command-line operations	Script execution accuracy within constraints	Yes
WebArena	E-commerce, forums, CMS	Functional correctness across user journeys	Yes
GAIA	Multi-step reasoning with tools	Question answering requiring web/tool use	Yes
ToolEmu	Tool use safety (emulated sandbox, 144 test cases, 36 toolkits)	High-stakes risk identification from underspecified instructions	Yes
BFCL	Function calling accuracy	Correct API call generation	Yes