Agent Benchmark Landscape

AgentBench

8 environments

OS, DB, Web

Multi-step tasks

WebArena

Web browsing

E-commerce

Realistic sites

SWE-bench

Code tasks

GitHub issues

Real repos

GAIA

Real-world QA

Multi-modal

3 difficulty levels

ToolBench

16k+ APIs

Tool selection

Multi-tool

OSWorld

Desktop OS

GUI tasks

Screenshots