

Agent Evaluation

Week 10: Benchmarks, Metrics, and Assessment

PhD Course in Agentic Artificial Intelligence

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

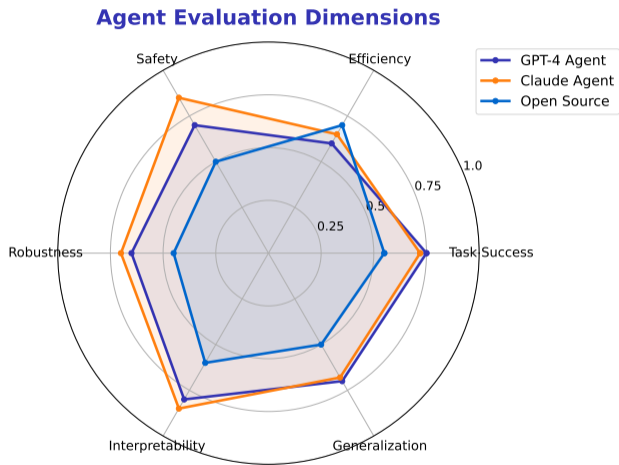
ToolBench

16k+ APIs
Tool selection
Multi-tool

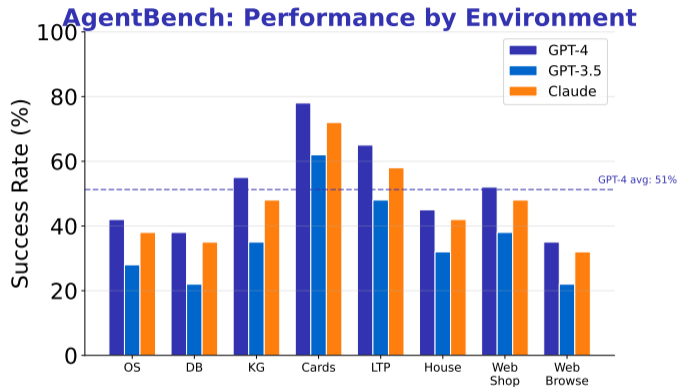
OSWorld

Desktop OS
GUI tasks
Screenshots

Each benchmark tests different agent capabilities and environments.

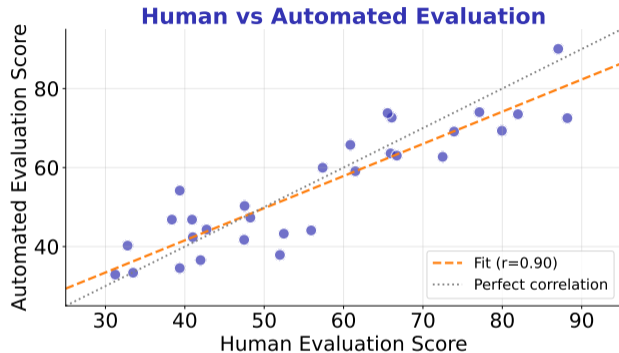


evaluation requires multiple dimensions beyond accuracy.



varies significantly across environments and models.

Human vs Automated Evaluation



metrics correlate with human judgment but not perfectly.

Auto

This Week

- Liu et al. (2023). “AgentBench: Evaluating LLMs as Agents.” arXiv:2308.03688
- Zhou et al. (2024). “WebArena: A Realistic Web Environment.” arXiv:2307.13854
- Mialon et al. (2024). “GAIA: A Benchmark for General AI Assistants.” arXiv:2311.12983

provides the most comprehensive multi-environment evaluation. **Agent**

Key Concepts

- **Benchmarks:** AgentBench, WebArena, SWE-bench (software engineering), GAIA
- **Dimensions:** Success, efficiency, safety, robustness
- **Methods:** Automated metrics, LLM-as-Judge, human evaluation

Next Week

- Domain Applications: Code, Finance, Healthcare

evaluation = right benchmark + right metrics + right baseline (reference model).

Good