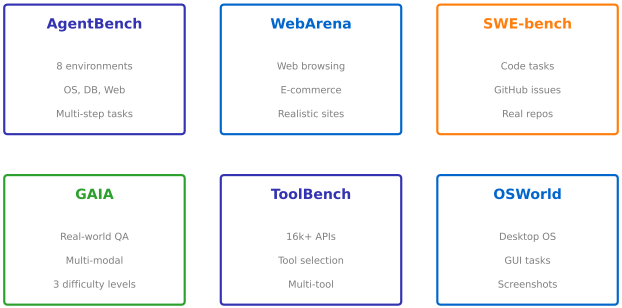
Agent Evaluation

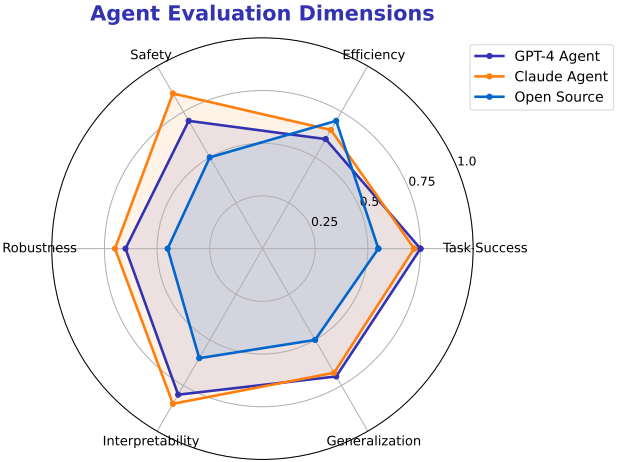
Week 10: Benchmarks, Metrics, and Assessment

PhD Course in Agentic Artificial Intelligence



Each benchmark tests different agent capabilities and environments.

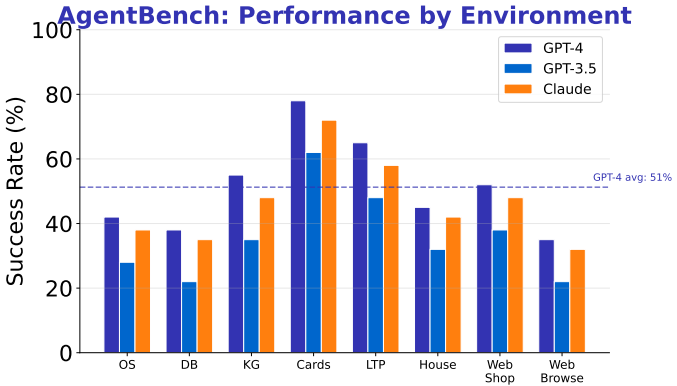
Evaluation Dimensions



Comp

evaluation requires multiple dimensions beyond accuracy.

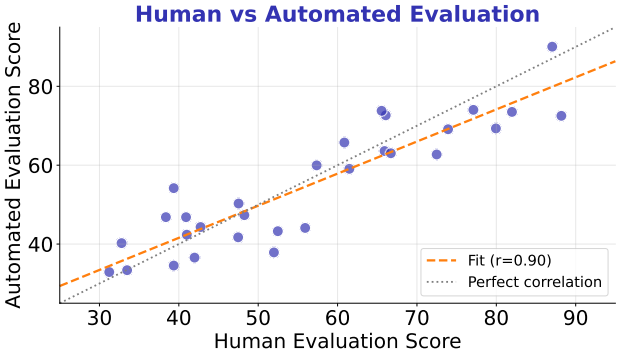
AgentBench Results



Perfo

varies significantly across environments and models.

Human vs Automated Evaluation



Auto

metrics correlate with human judgment but not perfectly.

Key Papers

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

Agen

provides the most comprehensive multi-environment evaluation.

Summary

Key Concepts

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

Next Week

- Domain Applications: Code, Finance, Healthcare

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