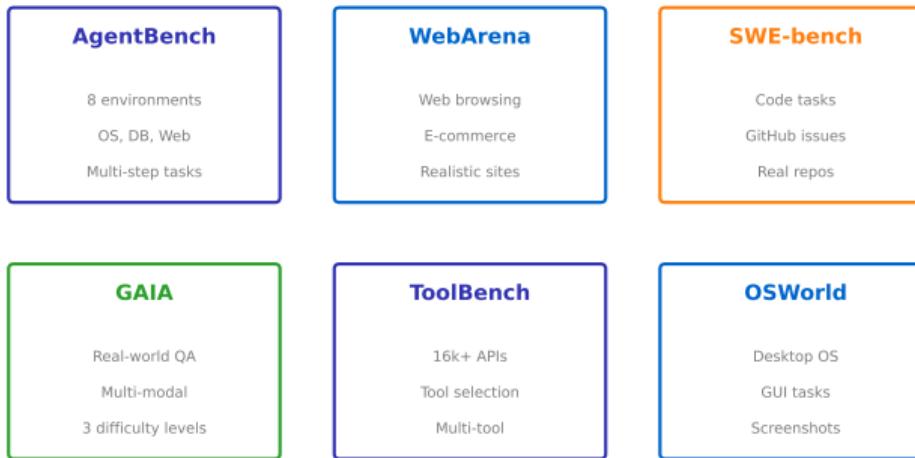


Agent Evaluation

Week 10: Benchmarks, Metrics, and Assessment

PhD Course in Agentic Artificial Intelligence

Agent Benchmark Landscape

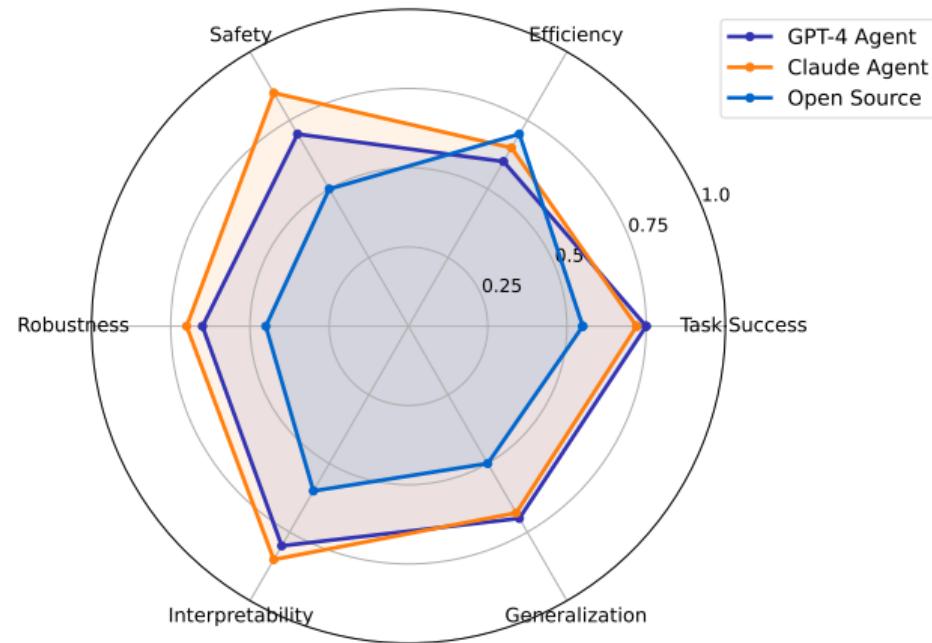


benchmark tests different agent capabilities and environments.

Each

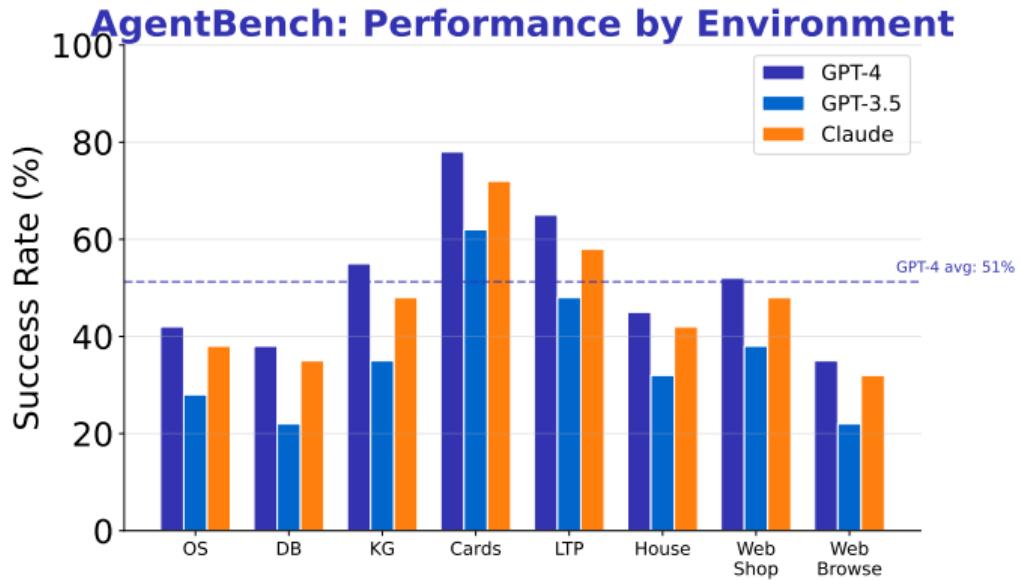
Evaluation Dimensions

Agent Evaluation Dimensions



evaluation requires multiple dimensions beyond accuracy.

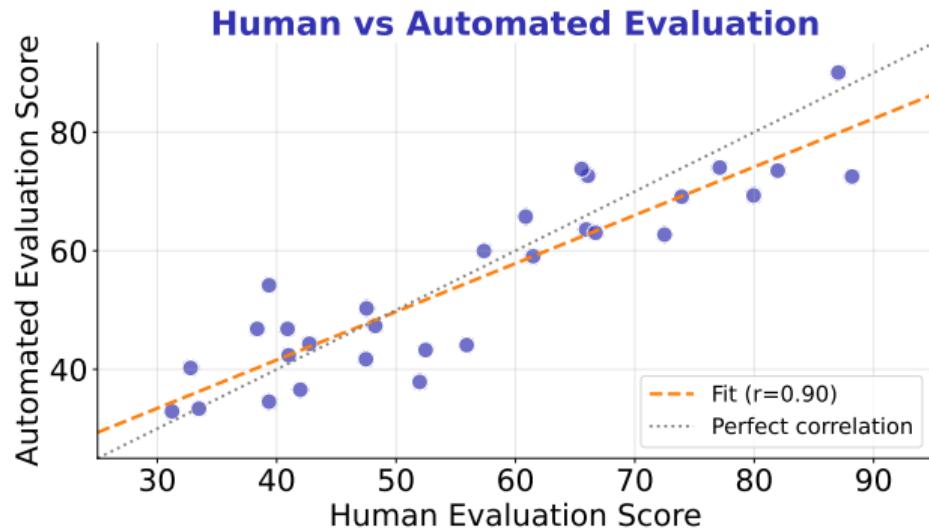
Com



varies significantly across environments and models.

Perfo

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.

Agen

Key Concepts

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

Next Week

- Domain Applications: Code, Finance, Healthcare

evaluation = right benchmark + right metrics + right baseline.

Good