Agentic Cloud Benchmark: Towards Fully Autonomous Cloud Software Workflows

- Mentor: Minghua Ma minghuama@microsoft.com

- Opensource

Project Overview:
Modern cloud software development involves complex, repetitive workflows: building code, testing across multiple environments, deploying updates, and repairing failures. Despite advances in automation, orchestrating these steps end-to-end remains challenging, and evaluating autonomous systems in realistic cloud settings lacks standardized benchmarks.

The **Agentic Cloud Benchmark (ACBench)** addresses this gap by providing a standardized framework to **evaluate end-to-end autonomous software workflows in the cloud**. ACBench tests agents across realistic cloud scenarios, assessing their ability to build, deploy, detect failures, and automatically repair systems. By capturing both efficiency and reliability metrics, it enables researchers and practitioners to measure, compare, and improve the capabilities of agentic cloud systems.

Learning Objectives:
Students completing this course will be able to:

1. Build a cloud-based system that automates software build, test, and repair workflows.
2. Collect and analyze observability data to evaluate agent performance.
3. Define agent skills and create reproducible benchmarking scenarios.
4. Demonstrate agent performance using standardized evaluation metrics.

Key Deliverables:

- Automated build, test, and repair benchmark framework
- Observability module for metrics, logs, and results
- Agent skill catalog and benchmark scenarios
- Benchmark demonstration and analysis
- Documentation and usage guide

Technologies:

- Cloud platforms (AWS, Azure, GCP)

- CI/CD tools (GitHub Actions)
- Containers (Docker)
- Observability tools (Prometheus, Grafana)
- Python/TypeScript for orchestration

Related works:

SWE-bench Leaderboards

SWE-bench-Live Leaderboard

Build-bench

AIOPSLAB