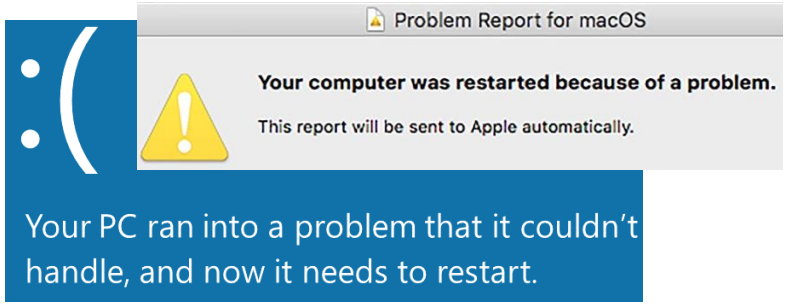
ITBench: Evaluating AI Agents across Diverse Real-World IT Automation Tasks

Saurabh Jha · Rohan Arora · Yuji Watanabe · Takumi Yanagawa · Yinfang Chen · Jackson Clark · Bhavya Bhavya · Mudit Verma · Harshit Kumar · Hirokuni Kitahara · Noah Zheutlin · Saki Takano · Divya Pathak · Felix George · Xinbo Wu · Bekir Turkkan · Gerard Vanloo · Michael Nidd · Ting Dai · Oishik Chatterjee · Pranjal Gupta · Suranjana Samanta · Pooja Aggarwal · Rong Lee · Jae-wook Ahn · Debanjana Kar · Amit Paradkar · Yu Deng · Pratibha Moogi · Prateeti Mohapatra · Naoki Abe · Chandrasekhar Narayanaswami · Tianyin Xu · Lav Varshney · Ruchi Mahindru · Anca Sailer · Laura Shwartz · Daby Sow · Nicholas Fuller · Ruchir Puri

IBM

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Relatable IT Automation Domains



Sarah

Site Reliability Engineer

SRE

Marcus

Chief Information Security Officer

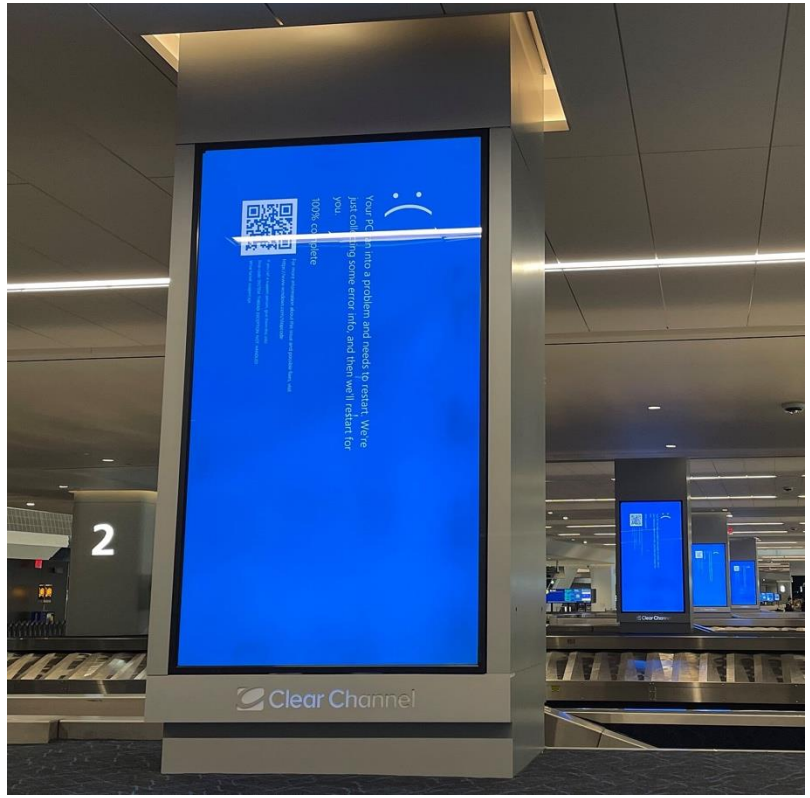
CISO

Lisa

FinOps Engineer

FinOps

CrowdStrike Outage: The Day IT Stood Still



A **single** software update brought down:

Airlines | Hospitals | Banks | 911 Services

8.5M Devices

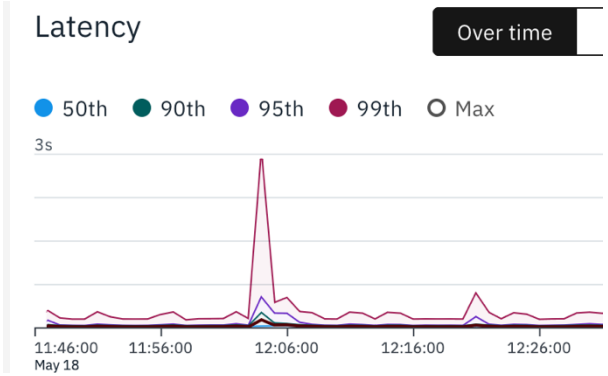
$5.4B Impact

79 min fix

Days of Recovery

These scenarios happen **frequently per day** across every large organization

How LLM agents can Help?

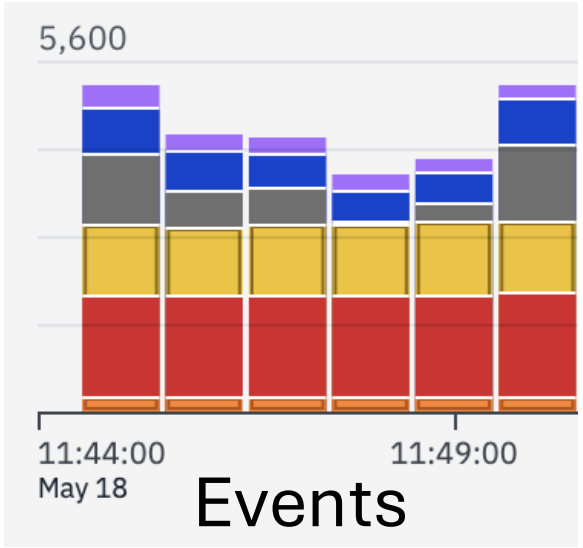


Timeseries Data

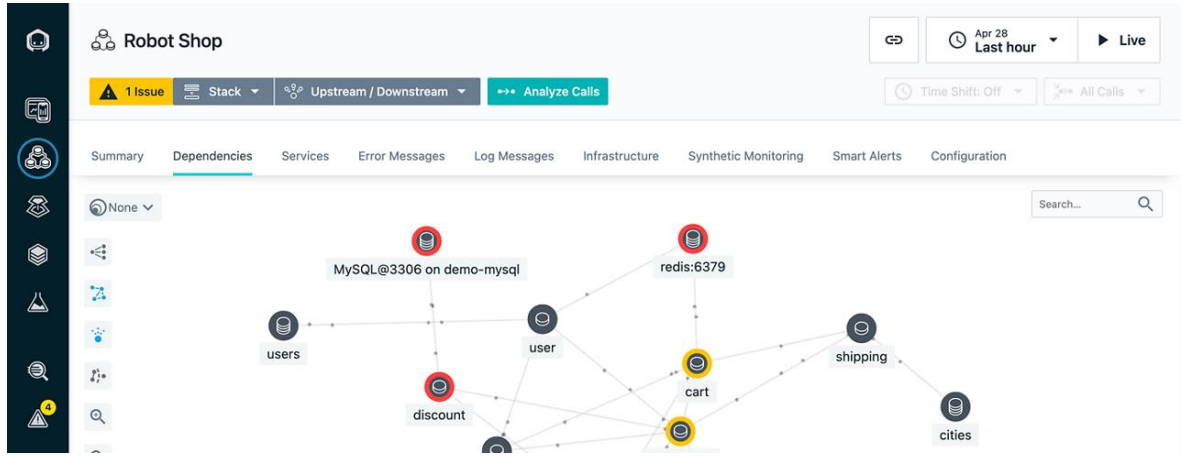
Agents


```
Error: 13 INTERNAL: failed to prepare order: fa
    at <unknown> (.next/server/pages/api/checkou
    at new Promise (<anonymous>) {
  code: 13,
  details: 'failed to prepare order: failed to
  metadata: [Metadata]
```

Code, Error logs



Events



Dashboards

Uniqueness of the IT Domain

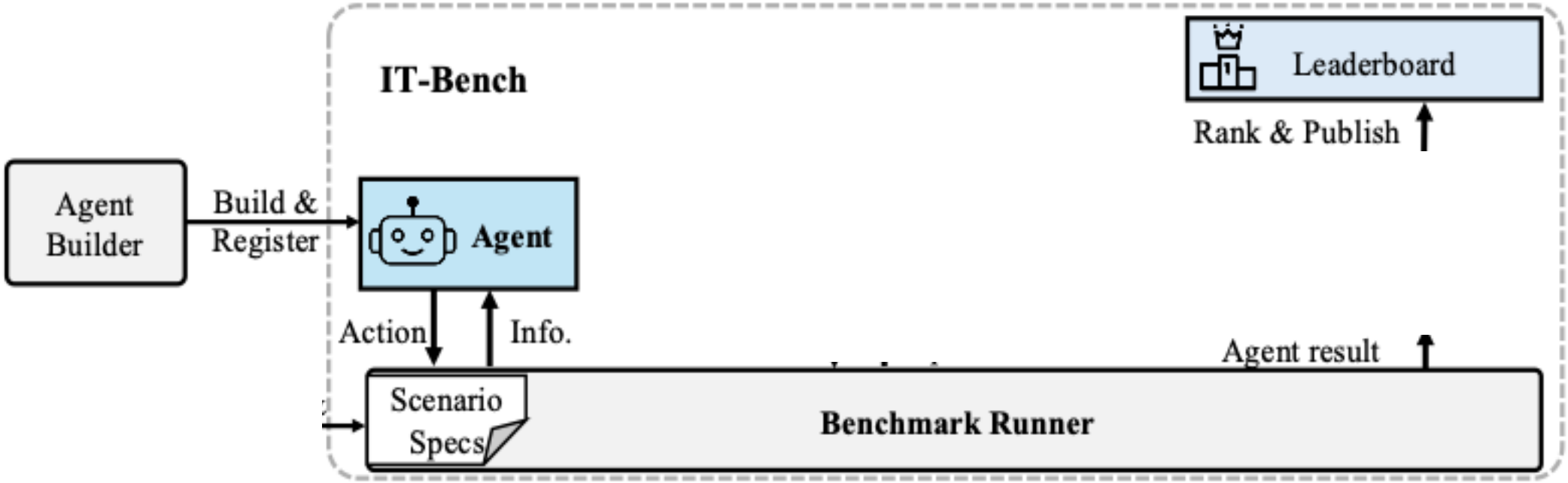
Scale & Diversity

Hundreds of interconnected services, each with different technologies and failure modes

Risk Assessment

Every action needs evaluation: downtime risk? Security vulnerabilities?

Introducing ITBench



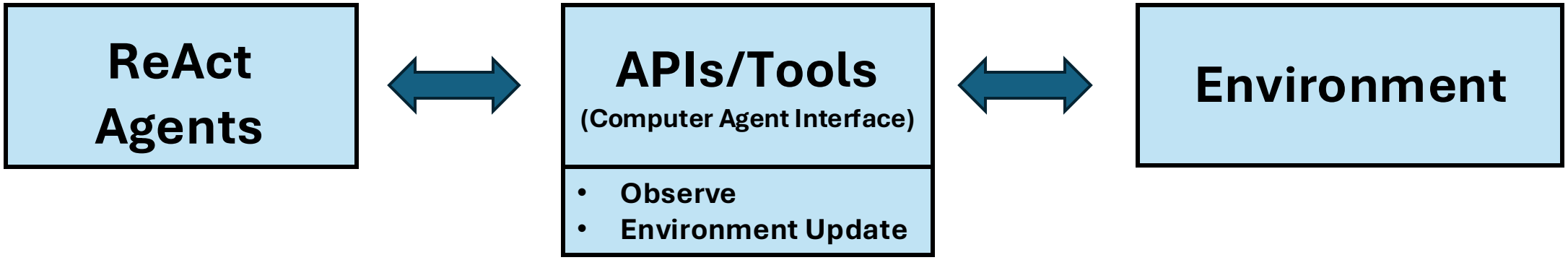
Realistic environments

Real-world grounded scenarios

Hides IT domain complexity

Rigorous evaluation framework

Baseline Agents

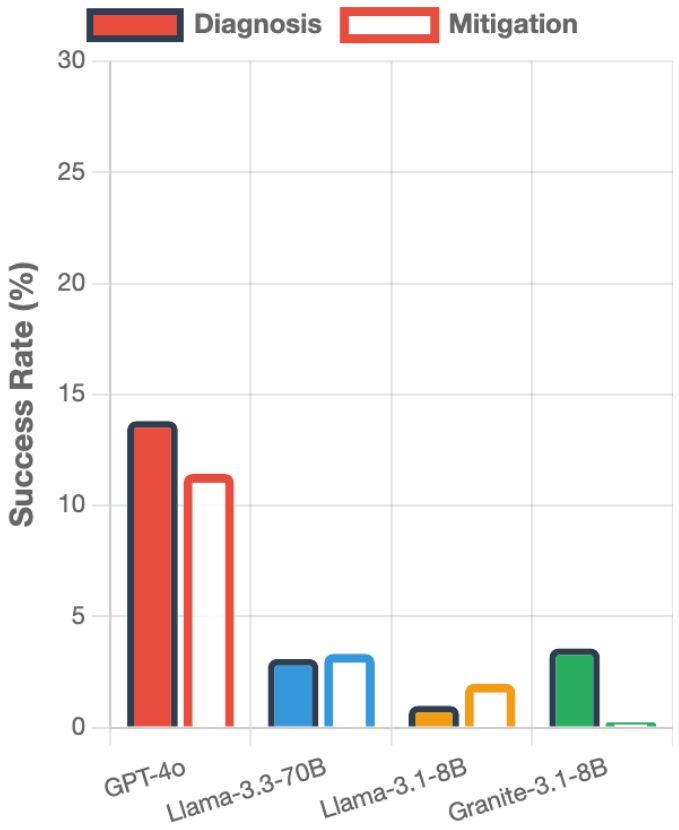


Thought: System is offline. Let me get the most recent events.

Action: Use tool get_events ()

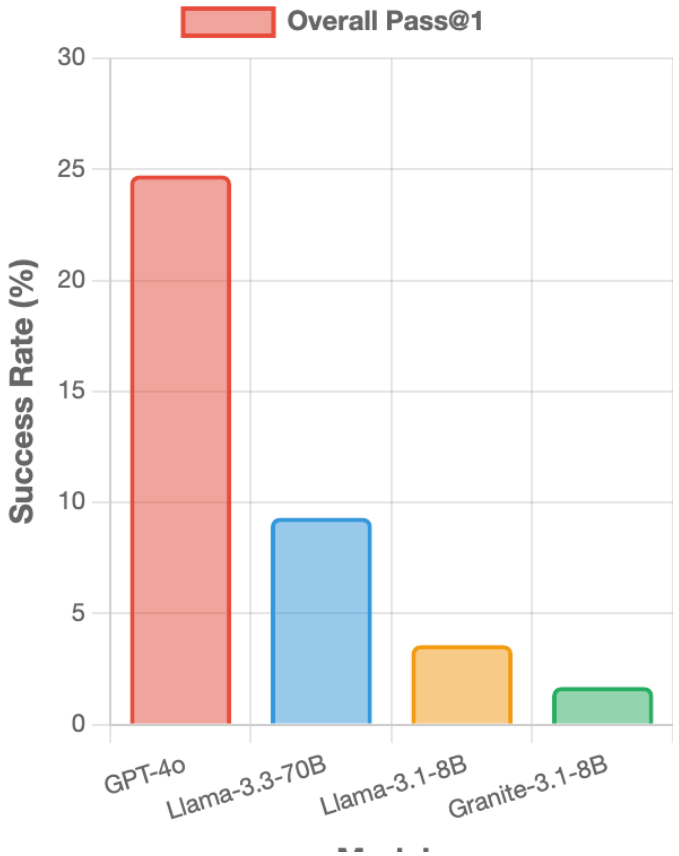
Observation: Code change occurred in file xyz

Promise of Agents



Site Reliability Engineer (SRE)

42 scenarios



Chief Information Security Officer (CISO)

50 scenarios

11.4%
SRE Scenarios

Failure Mitigation

24.7%
CISO Scenarios

Policy Generation

25.8%
FinOps Scenarios

Cost Analysis & Optimization

Why is ITBench HARD for LLM agents?

Handling LLM Context is Challenging!

Task: Identify events causing the alerts

Active Alerts

RequestErrorRate - checkout
Error rate: 0.45 | Active since: 00:06:22Z

RequestErrorRate - payment
Error rate: 0.58 | Active since: 00:06:24Z

Recent Events (Observations)

00:05:35Z - ConfigMap Modified
Feature flag enabled for handling additional currencies

00:05:40Z - NetworkPolicy Applied
default-deny-all-ingress applied cluster-wide (reverted at 00:06:40Z)

Gemini 2.5 Pro INCORRECT

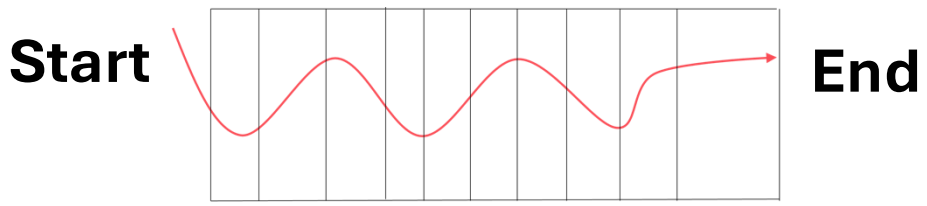
Claude Sonnet 4 INCORRECT

Magistral Medium 3 INCORRECT

Irrelevant, conflicting context, non LLM friendly language

→ Confidently incorrect decisions

Long Horizon Planning is Challenging!



Agent Trajectory
(GPT-4o = 93 turns)

```
Turn 1: kubectl get pods --all-namespaces
Turn 2: kubectl describe node production-
worker-03
Turn 3: kubectl get pv | grep critical-data
Turn 4: kubectl logs payment-service-xyz
Turn 5: kubectl get deployments -n production

                    ...

Turn 45: kubectl get nodes --show-labels
Turn 46: kubectl describe pv critical-data-pv
Turn 47: kubectl delete node production-
worker-03
💥 200+ pods evicted, payment system down
Turn 48: kubectl delete pv critical-data-pv
💥 Customer data permanently lost
Turn 49: kubectl delete namespace production
💥 Entire production environment destroyed
```

Each Step = Potential Catastrophe

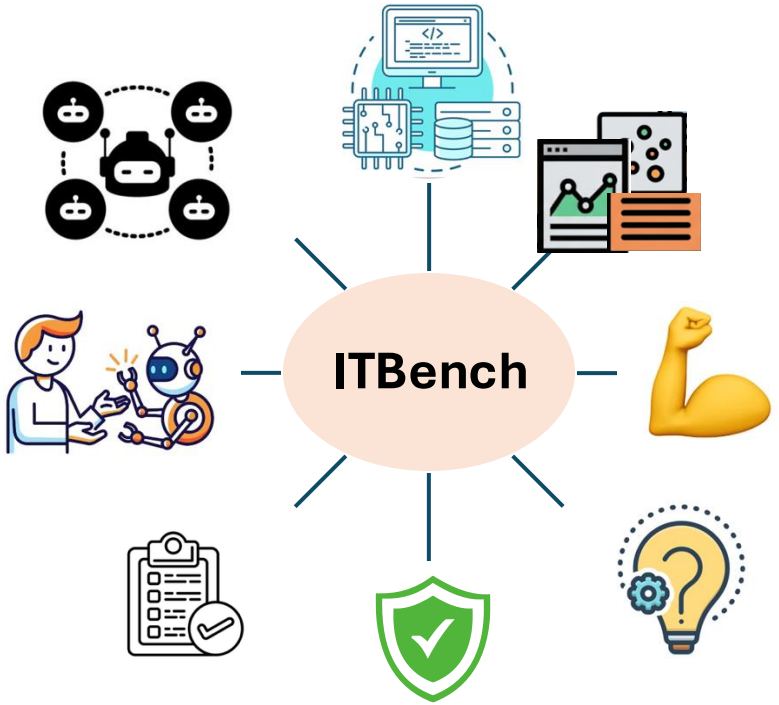
- State changing actions can be hazardous
  - No test time compute
- Failure to achieve stated goals

One Platform. Many Challenges. Why ITBench?

System Complexity
Manage the full IT environment including hardware and software stacks, networks

Multi-Agent Co-ordination
Effective collaboration between various agents (e.g., diagnosis agent and mitigation agent)

Data Heterogeneity
Understand logs, metrics, configs, code



Agent-Human Interaction
Effective instruction following and explainability

Robustness
Handle noise and unpredictability in live environment (e.g., network latency variations)

Evaluation
Assess quality of agent actions for IT management tasks

Safety
Ensure system integrity and prevent catastrophic failures (e.g., database deletions)

Long Context Reasoning
Reasoning over massive IT data (e.g., causal analysis)

Growing with the Community

🎯 For Agent Developers

- Explore sample scenarios on GitHub
- Develop agents for realistic problems
- Submit to our leaderboard
- Benchmark against state-of-the-art

🤝 For Contributors

- Submit new scenario families
- Contribute evaluation metrics
- Share domain expertise
- Help expand to new IT domains

https://github.com/ITBench-Hub/itbench

Thank You

See you at the poster session!

Poster **#W-103**
West Exhibition Hall B2-B3

ITBench on GitHub

https://github.com/ITBench-Hub/itbench