

 FINTECH AI AGENT BENCHMARKING

Sonata

Replicated Sandbox Environments Built for
Autonomous AI Agents in Financial Operations.

sonatalabs.ai

 Hackathon Pitch 2026



Agent Test Engine

Real-Time Simulation Suite

PASSED (10/10)



Ledger Integrity

0 Discrepancies



Adversarial Latency

Handled (100ms)



CI Gate Status

Deploy Approved

Deterministic Accuracy Score

100%

! THE PROBLEM

Generic LLM Evals Fail Financial Operations



CRITICAL CAPITAL EXPOSURE

Regulatory Liability

Single ledger mistakes trigger regulatory fines and permanent balance corruption.

Zero Tolerance for Hallucinations



Context & Schema Blindness

Public benchmarks lack internal bank schemas, live auth logic, and database state.



Silent Edge-Case Cascades

Multi-turn financial workflows fail silently under payload errors and rate limits.

Production-Grade Benchmarks for AI Agents

01 Replicated Bank Stacks

Exact clones of internal ledgers, schemas, and API permissions.

02 Adversarial Stress Scenarios

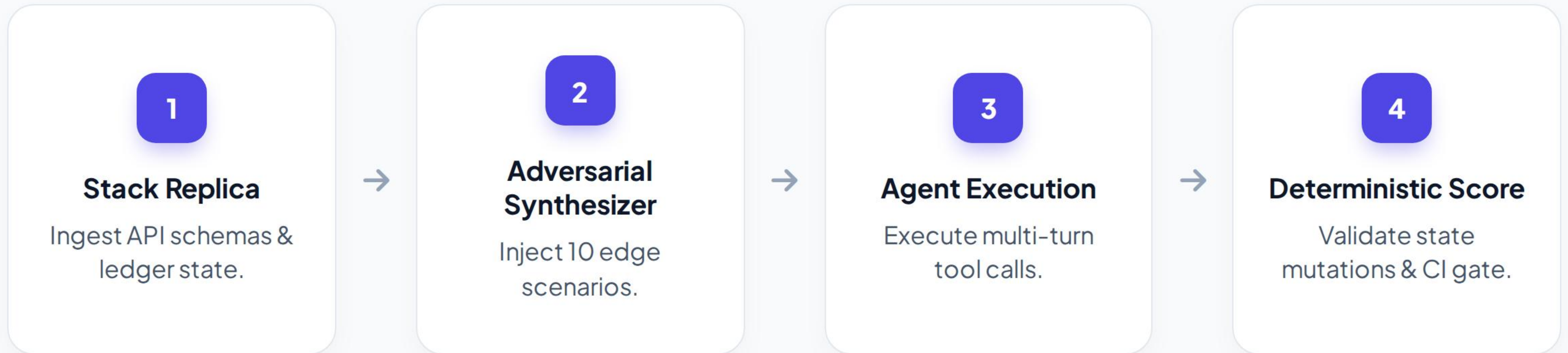
Injected network lags, corrupted payloads, and state mutations.

03 CI/CD Regression Package

Automated pass/fail gates before releasing code to production.



End-to-End Agent Testing Lifecycle



TARGET SCENARIOS

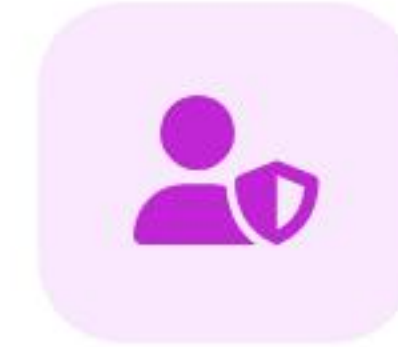
Built for High-Stakes Financial Operations



RECONCILIATION

Multi-Ledger Settlement

Multi-party payout logic validation.



COMPLIANCE

KYC & Fraud Triage

Edge-case false-positive triage.



SAFETY

Hallucination Guardrails

Unchecked capital mutation prevention.



DIAGNOSTICS

Tool Call Trace Analysis

Full API request & state auditing.

QUANTIFIED IMPACT

Proven Reliability Metrics

10

Rebuilt Stack Scenarios

Full Banking Sandbox

99.4%

Edge-Case Catch Rate

Pre-Production Auditing

10x

Deployment Velocity

Automated CI/CD Regressions

Fintech Engineers

Safely deploy autonomous agents into core bank stacks.

Risk Officers

Audit safety scores with complete trace proof.

AI Operations

Continuous regression testing on pull requests.

Why Sonata Outperforms Generic Evals

Evaluation Axis	Generic LLM Benchmarks	Custom Internal Mock Scripts	Sonata Engine
Environment Fidelity	Isolated Text Strings	Fragile Mock APIs	✔ Full Replicated Stacks
Multi-Step Tool Calls	Single-Turn QA	Hardcoded Call Paths	✔ Autonomous State Testing
Financial Safety Audits	Zero Ledger Context	Manual Script Maintenance	✔ Deterministic Ledger Audits
CI/CD Integration	None	Flaky Local Test Suites	✔ Automated PR Regression Gates

HACKATHON FINALE 2026

Deploy Autonomous Finance AI With Absolute Confidence

Replicated Sandbox Stacks • Adversarial Stress Tests • Automated
CI/CD Gates

 Try Sonata — sonatalabs.ai

Image Sources



https://img.magnific.com/premium-vector/ai-artificial-intelligence-finance-fintech-data-marketing-technology-business-data-report-robot-driven-data-analytics-dashboard-vector-flat-illustration_924480-236.jpg?semt=ais_hybrid&w=740&q=80

Source: www.magnific.com