FlakeWarden: Agentic Flaky-Test Triage

An agentic triage system for UiPath Test Cloud that decides whether a failing test is a real defect, a flaky test, or an environment problem, and routes each to the right action with a human approving every change.

Problem statement and proposed solution

Problem

Flaky tests are the most corrosive failure mode in CI. When a build goes red, an engineer often cannot tell a real regression from noise, so they either burn time triaging every failure or, worse, start ignoring red builds until a genuine regression ships. Google reported that about 16% of their tests showed flakiness and about 84% of pass to fail transitions came from flaky tests.

Solution

FlakeWarden looks at a failing test's history and evidence and classifies it as real defect, flaky, or environment. A deterministic flake-scorer handles the clear cases exactly and never guesses; a grounded UiPath Agent Builder classifier reasons over the ambiguous ones. UiPath Maestro orchestrates the flow, and every fix or quarantine passes through an Action Center human-review gate. Deterministic where it must be exact, generative where the context is messy.

Benefits and technologies used

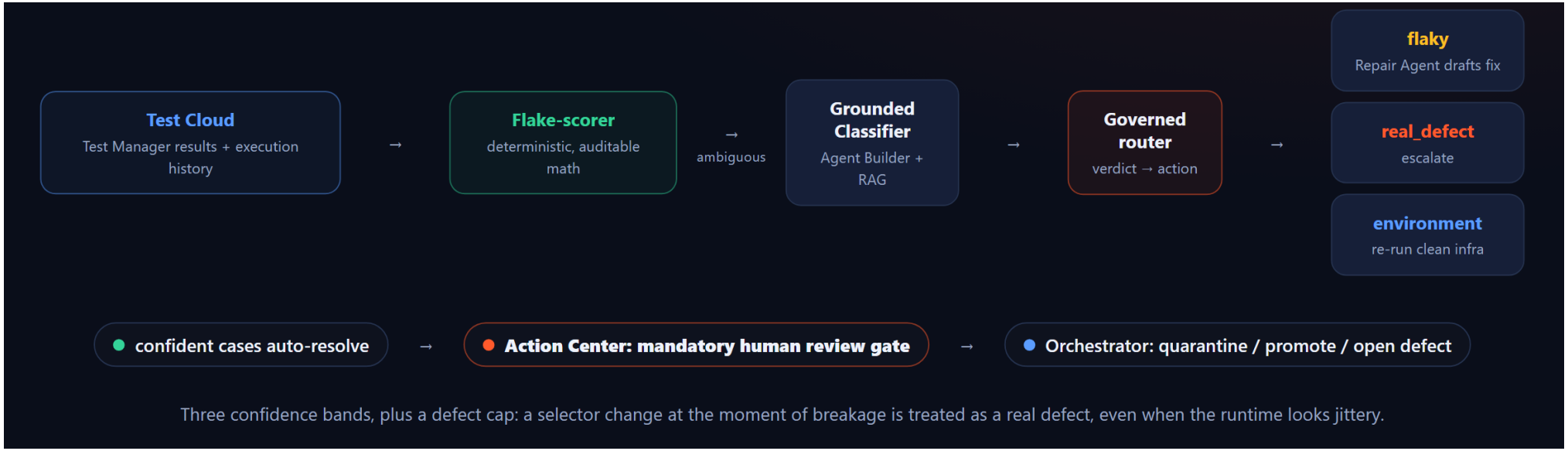
End-user	QA engineers, SDETs, test automation leads, release managers
User department	Quality Engineering, QA, Release Engineering, DevOps
Industries	Software, Financial Services, Healthcare, Enterprise SaaS
UiPath products used	Test Cloud, Test Manager, Agent Builder, Maestro, Action Center, Orchestrator, Context Grounding, AI Trust Layer, uip CLI
Other - integrations / APIs / technologies used	Python, Claude Code (Anthropic Claude), GitHub, Mermaid

Benefits, impact and outcomes

- Separates real defects from flaky tests, so teams stop wasting hours on noise and stop ignoring red builds.
- 90.7% accuracy and a 0% safety-direction false-positive rate on a 150-case corpus; a real defect is never hidden as flaky.
- Spends an LLM only on ambiguous failures; about a third resolve with exact, auditable math and no model call.
- Every fix, quarantine, or baseline change is a human-gated proposal, never an autonomous mutation.

Solution architecture

Test Cloud results feed a deterministic flake-scorer. Confident cases route straight to an action; ambiguous cases go to a grounded UiPath Agent Builder classifier. A governed router sends every verdict through an Action Center human-review gate before Orchestrator writes back the result. UiPath Maestro orchestrates the whole flow, and the entire build was driven by Claude Code through the uip CLI.
GitHub Repo: https://github.com/JonathanSolvesProblems/flakewarden



THANK YOU.

UiPath

FlakeWarden · github.com/JonathanSolvesProblems/flakewarden · Demo: youtube.com/watch?v=6md-vEuY_-0