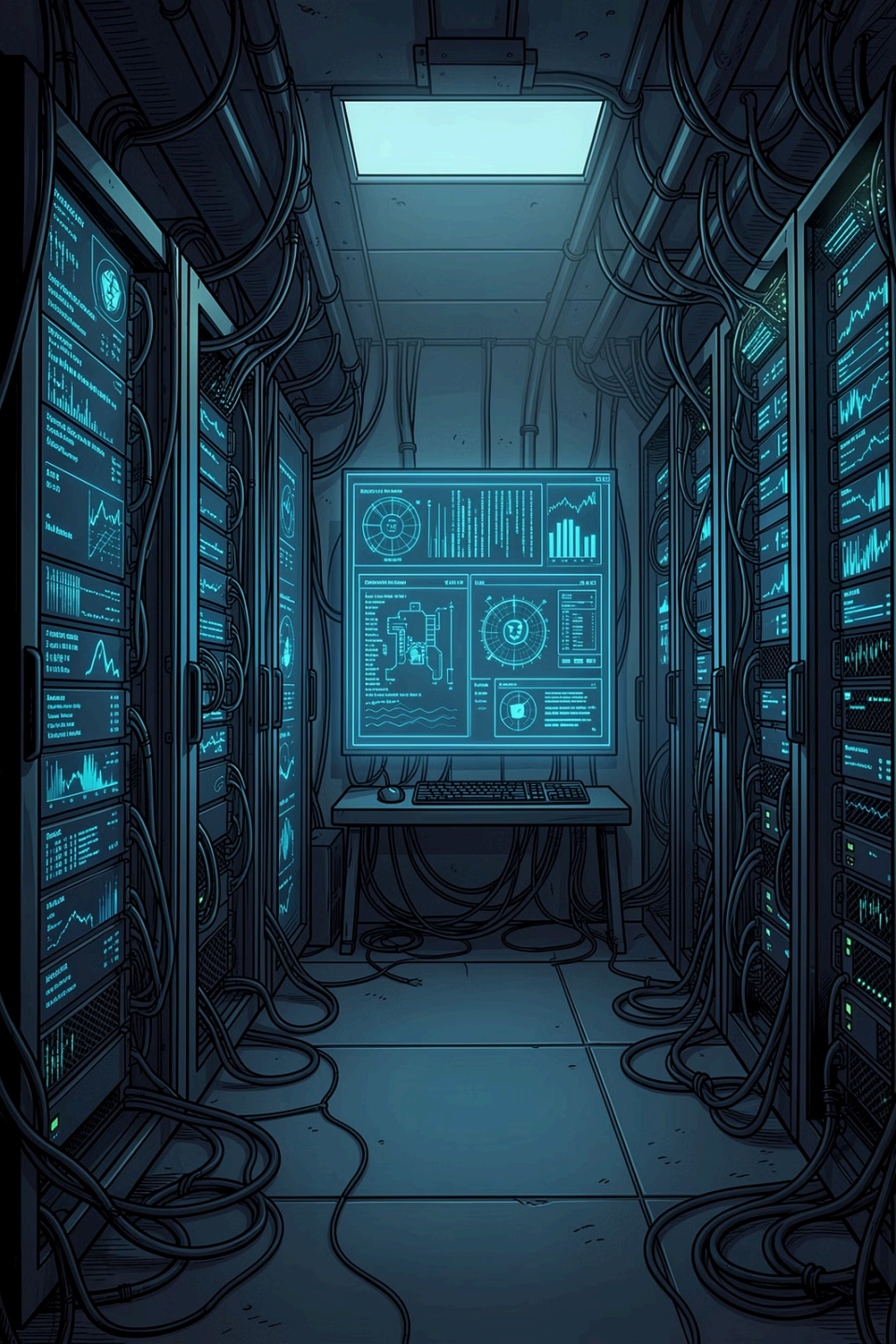




TECHNICAL SAFETY TRACK — SMALL AI & ON-DEVICE AI

DriftWatch: Oversight Decay Across AI Tiers

Simulation framework testing whether closed, open-source, or on-device models speed up **blind trust** in AI decisions.

The Core Problem: Oversight Decay

When AI is usually right, humans stop reviewing. It's a **feedback loop** with compounding risk.

1

Trust Builds

AI is right, so scrutiny drops

2

Skills Atrophy

Review ability weakens over time

3

Errors Slip Through

Mistakes go unnoticed

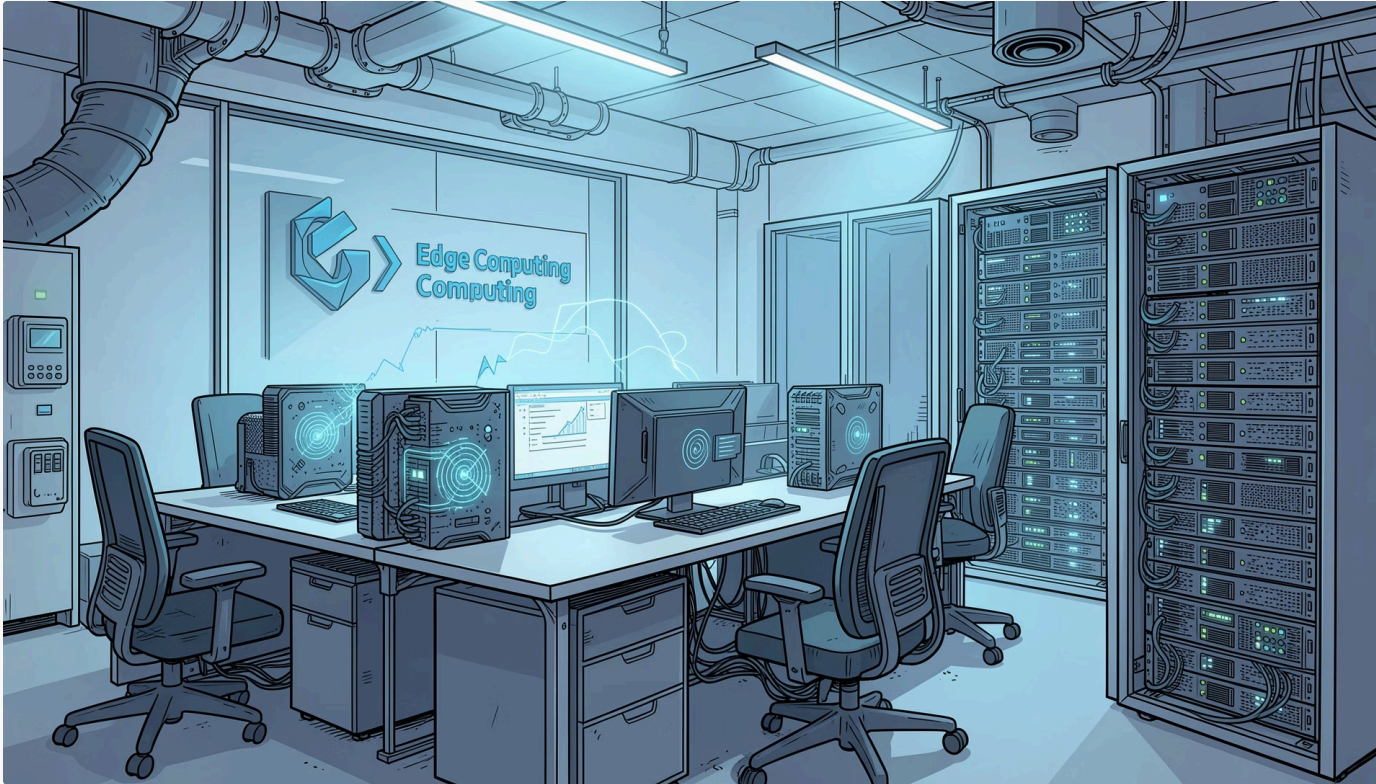
4

Failures Compound

Undetected errors stack up



Why This Matters for Technical Safety



Most AI safety work focuses on **frontier closed models**. But small, open-source, and on-device models are already used in real settings:

- ❏ **DriftWatch asks:** Does smaller or local AI change oversight safety — for better or worse?

Government Services

Benefits, permits, eligibility

Local Agencies

Low-resource oversight

Edge Devices

Offline AI, no central monitoring

System Architecture

DriftWatch simulates a human-AI decision loop. The **Caseworker Agent** backend is pluggable — swap models without changing simulation logic.



Each citizen tracks `review_probability` and `review_skill`. Both change with error detection, producing oversight decay.

Layers of Realism

Depth Variables

Language mismatch, confidence calibration, and explanation style affect error detection.

Strategic Adversaries

Intelligent actors exploit silent AI errors with fraudulent submissions.

Social Contagion

Trust spreads through citizen networks; one neighbor catching an error resets local oversight.

Three Model Tiers Tested

DriftWatch tests the **open-source / small / on-device axis** against a closed-model baseline.

Tier	Closed	Open (API)	Local / Quantized
Model	GPT-4o	Llama 3.1 8B	Llama 3.1 8B Q4_K_M
Runtime	OpenAI API	Ollama API	Ollama GGUF
Cost	Per-token	Free	Free
Privacy	Data leaves env	Self-hosted	On-device
Safety Profile	Frontier baseline	Open-weight	No vendor lock-in

Key Metrics: Oversight Framework

Measuring the problem, then testing the fixes.

Oversight Decay

Average review_probability over time

Time-to-Threshold

When silent error rate exceeds 10%

Silent Error Rate

Incorrect AI decisions missed by humans

Mitigation Interventions

- Spot Checks: random forced reviews
- Confidence Thresholds: low-confidence cases routed to review
- Mandatory Audits: periodic deep-dives reset trust



Tech Stack & Implementation

Backend



Python simulation engine with FastAPI-style API. Backends: OpenAI, Ollama API, Ollama GGUF local, and rule-based baseline.

Frontend

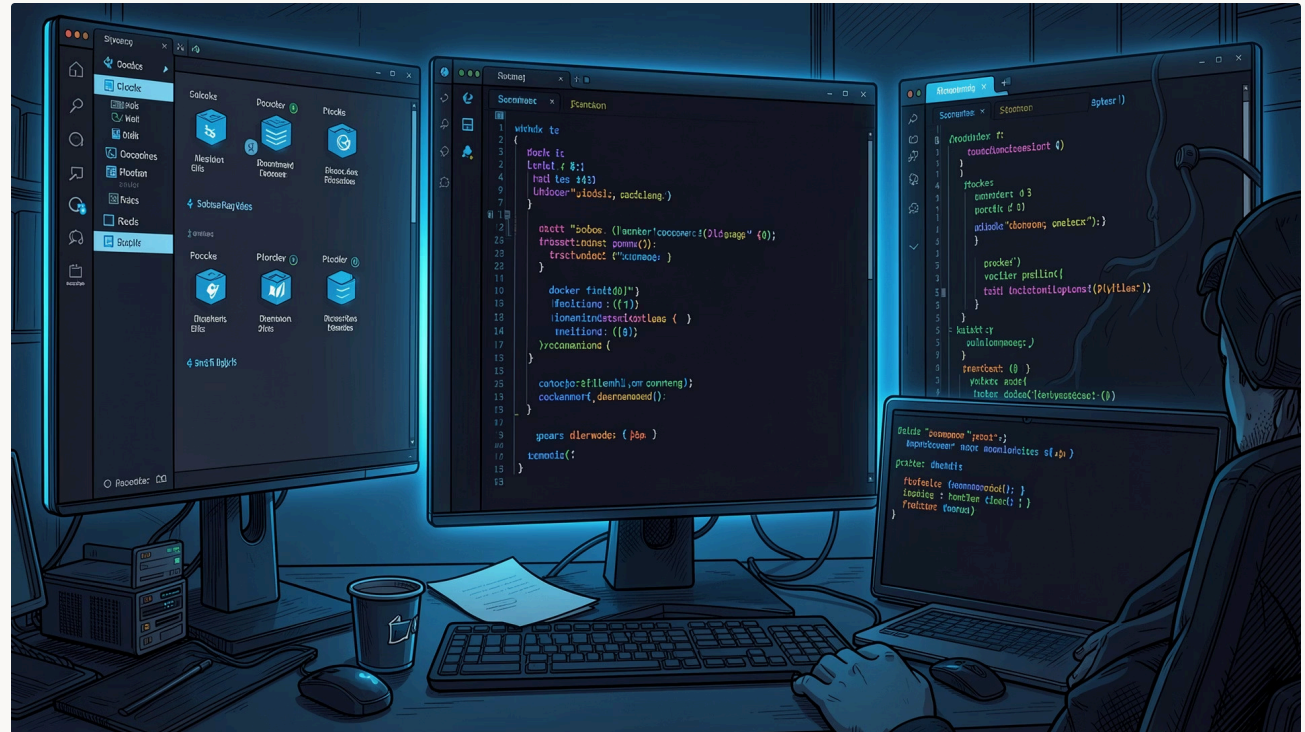


TypeScript / React dashboard for runs, metric curves, and citizen behavior.

Deployment



Dockerized with docker-compose. Deploys to GCP via Cloud Build. Backends switch via ENV variables.



Takeaways & Future Work

Core Insight

Model choice may matter less than the **oversight design** around it — smaller, local models make these safety tests **cheap, accessible, and vendor-free**, while mitigation interventions are now part of the toolkit.

① As models get smaller and cheaper, **oversight-aware design** becomes a core safety requirement.

What's Next

→ Real Multilingual Models

Test actual multilingual model behavior

→ Multimodal Inputs

Add images and documents

→ Larger-Scale Network Studies

Test oversight contagion in bigger, more realistic populations