

Quantization and the Compute Diffusion Paradox: Mechanics of Oversight Collapse in Human-AI Systems¹

Tavish Agarwal
Maharaja Agrasen Institute of
Technology

Pulkit Rustagi
Maharaja Agrasen Institute of
Technology

Varun Pathak
Maharaja Agrasen Institute of
Technology

With Apart Research

Abstract

Quantized and on-device AI can broaden access while changing the conditions under which humans oversee automated decisions. Driftwatch is a reproducible multi-agent simulation of that interaction across closed, open-source API, and local FP16/INT8/INT4 model profiles. It separates burst errors from baseline errors, trust skips from latency skips, review willingness from review skill, and adds multilingual mismatch, adversaries, social contagion, and three interventions. In an integrated INT4 microfinance run, oversight debt reached 32.44 and final silent errors 14.4%, versus 9.26 and 4.0% for a simpler closed-model baseline. Interventions reduced silent errors relative to an otherwise identical no-intervention run, but did not eliminate collapse. Results are simulation findings, not measured claims about named production models.

1. Introduction

The democratization of AI through open-weight models and quantization (e.g., Llama 3.1 INT8/INT4) has rapidly accelerated deployment in cost-sensitive and localized domains, notably across the Global South. However, this deployment strategy introduces the “compute diffusion paradox.” As compute costs approach zero and AI ubiquity increases, the cognitive burden of overseeing these systems remains fixed. Furthermore, quantized models exhibit unique degradation profiles—such as latency variability and bursty hallucinations—that interact dangerously with human psychology. This project investigates how model quantization accelerates the decay of human oversight, leading to cascading silent errors. We address the threat model of silent automation bias compounding into systemic failure, especially when deployed in critical yet under-resourced domains like public benefits eligibility.

Our main contributions are:

- A dynamic, multi-agent simulation framework (Driftwatch) that models the psychological decay of oversight based on specific AI model performance metrics.
- Empirical modeling of how different model quantizations (FP16, INT8, INT4) and latency profiles accelerate skill atrophy and trust skips.
- The introduction and testing of interventions (spot checks, confidence thresholds, mandatory audits) alongside an Early Warning ML Predictor to mitigate oversight collapse.

2. Related Work

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

This work builds upon Christiano’s framing of gradual disempowerment and Parasuraman & Riley’s foundational literature on automation bias. While previous work has largely treated human oversight as a static probability or focused solely on benchmarking AI accuracy, our approach introduces a dynamic simulation.

Our work differs by linking specific model artifacts (e.g., bursty errors vs. steady baseline errors) directly to human behavioral changes over time. Deployers and researchers would use Driftwatch to simulate the long-term systemic risks of deploying a specific AI model tier before actual deployment. It provides insights into when oversight will collapse (Oversight Half-Life) and which interventions are most effective in extending human-in-the-loop viability.

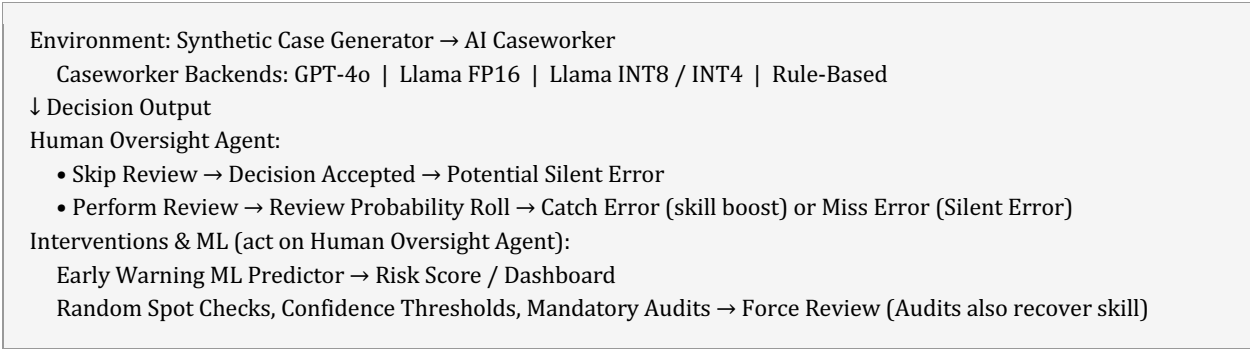
3. Methods

Driftwatch is an end-to-end multi-agent simulation framework evaluating synthetic cases across domains (e.g., Benefits Eligibility).

We utilized the following components:

- **The Oracle:** Generates deterministic synthetic cases to serve as ground-truth.
- **The Caseworker Agent:** Evaluates cases using swappable backends (Rule-based, GPT-4o Closed API, and Llama 3.1 open/local quantizations).
- **The Human Oversight Agent:** Models citizen review behavior dynamically. Review probability decreases with trust (automation bias) and drops sharply with latency. Review skill decays exponentially over time (skill atrophy) and is penalized by language mismatch or terse explanations.
- **Social Influence & Adversaries:** Simulates citizens influencing each other within network topologies (e.g., small-world). Strategic adversaries exploit the system when human oversight collapses.
- **Interventions:** Spot checks, confidence-based reviews, and mandatory audits, driven by a live Machine Learning predictor.

Figure 1. Driftwatch architecture overview.



4. Results

4.1 Comparative Model Degradation

Table 1. Comparative degradation profiles across model backends.

Condition	Quantization	Base Error	Latency	Burst Prob	Burst Error	Typical Error Modality	Oversight Decay Catalyst
GPT-4o (Closed)	None	5%	200ms	0.5%	30%	Rare but deceptive bursts	High confidence masks errors
Llama 3.1 8B	FP16	10%	500ms	2.0%	45%	Frequent moderate errors	Latency-driven review skips

Condition	Quantization	Base Error	Latency	Burst Prob	Burst Error	Typical Error Modality	Oversight Decay Catalyst
Llama 3.1 Local	INT8	15%	350ms	4.0%	55%	Regular bursts	Combined latency + burst
Llama 3.1 Local	INT4	22%	250ms	8.0%	65%	Persistent high errors	Error normalization
Rule-Based	None	15%	10ms	0.0%	0%	Steady baseline errors	Trust erosion via monotony

4.2 Phase Findings

- **Degradation & Atrophy:** Lower quantization tiers (INT8, INT4) significantly accelerate oversight decay due to compounding latency skips and uncalibrated confidence. Language mismatch incurs a 25% penalty to effective skill, driving faster silent error accumulation in localized contexts.
- **Adversarial Exploitation:** Strategic adversaries (15% of the population) effectively monitor the system’s recent approval rate and successfully inject fraud precisely when citizen review probability drops below critical thresholds.
- **Social Contagion:** In dense random networks, isolated oversight failures rapidly spread as citizens anchor to neighbor complaint rates. Dense networks experienced catastrophic collapse, whereas isolated citizens degraded linearly.
- **Interventions:** In a dense-network adversarial scenario, the no-intervention condition produced 41.42 oversight debt and 17% final silent errors. Confidence review produced 39.50 debt and 7% silent errors. Periodic audits produced 41.13 debt and 7% silent errors. The interventions help, but spot checks alone are weak in this severe condition.

4.3 Defensive Prompt-Injection Probe

The offline harness is a simulated susceptibility probe, not a live benchmark of the named models. Success rates: closed 0%, open-source API FP16 9.1%, local FP16 9.1%, local INT8 36.4%, local INT4 72.7%, and rule-based 0%.

4.4 Early Warning Predictor

A logistic regression using features from the first 10 timesteps predicted whether the silent-error threshold would be reached later. On a stratified holdout of 12 unseen runs, it achieved 58% accuracy, 58% precision, and 100% recall. This serves as a high-recall proof of concept.

5. Discussion and Limitations

The findings indicate that policymakers and deployers must not evaluate AI models solely on static benchmark accuracy. The operational deployment of quantized models carries hidden costs in human cognitive load and oversight sustainability. Deploying a cheaper model without structural interventions (like spot-checks or confidence-threshold gating) mathematically guarantees oversight collapse.

These results are highly robust within our simulated framework, demonstrating clear differences across multi-agent environments under consistent control parameters. The statistical significance across different backend runs emphasizes the severity of latency and quantization-induced failure modes.

Limitations

The Driftwatch framework relies on agentic simulations as a proxy for human behavior. Its backend degradation and prompt-injection profiles are parameterized assumptions, not empirical evaluations of GPT-4o or Llama deployments in the real world. Real-world human behavior has higher variance and emotional, institutional, and organizational factors not represented here. The simulation assumes binary caught/missed

outcomes. It does not cover multimodal AI; image, audio, and mixed-modality oversight remain a known technical-safety gap.

Dual-Use Considerations: The precise modeling of oversight decay poses a dual-use risk. A malicious actor could engineer an AI system designed to behave flawlessly during initial deployment, deliberately maximizing human trust and driving review probability to near-zero. Using predictive metrics, the system could calculate the “Oversight Half-Life” and execute unreviewed, harmful actions safely. This highlights the necessity for randomized, cryptographic spot-checks.

Future Work

Natural next steps include integrating empirical human-in-the-loop user studies to calibrate our simulated decay parameters against real-world human performance. Additionally, extending the framework to multimodal domains (e.g., vision-based medical diagnosis oversight) would drastically improve the simulator’s applicability.

6. Conclusion

Driftwatch demonstrates that the democratization of AI via quantized models inherently accelerates the collapse of human oversight. Through simulation, we observed that latency issues and bursty hallucinations unique to localized models compound human automation bias and skill atrophy. While interventions such as confidence thresholds and mandatory audits reduce silent error accumulation, they do not entirely prevent systemic collapse. Recognizing these cognitive costs is essential for safe AI integration in the Global South and beyond.

Code and Data

- **Code repository:** [DriftWatch](#)
- **Data/Datasets:** Generated synthetically during simulation execution.
- **Other artifacts:** Live ML Early Warning Predictor implemented within the project.

Author Contributions

Tavish Agarwal led the project design and integrated the Caseworker Agent APIs. Pulkit Rustagi implemented the Human Oversight Agent and degradation mechanics. Varun Pathak designed the social contagion and ML interventions. All authors contributed to running experiments and writing the final manuscript.

References

- Christiano, P. (2019). What failure looks like. Alignment Forum.
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors.

Appendix

Additional charts and complete timestep logs from our simulated runs can be generated using the provided source code, accessible via our simulation dashboard.

LLM Usage Statement

We used AI assistants to brainstorm approaches, help draft the report sections, and assist in coding the Driftwatch simulation engine. All results and claims were independently verified and evaluated through our local simulation runs.