
Sentinel: A Decentralized Threat Telemetry Network¹

Kamdem-Fezeu, Harry
Primary, NYC

Lu, Allen
Primary, NYC

Schmitz, Emil
Primary, NYC

Wen, Lily
Primary, NYC

With
Apart Research

Abstract

As generative AI lowers the marginal cost of personalized social engineering and misinformation to near-zero, traditional centralized defense mechanisms are becoming insufficient. This project introduces *Sentinel*, a browser-based distributed solution that accelerates AI defense through human-AI collaboration. Our system employs a hybrid verification architecture: users identify potential cognitive hazards (sybil attacks, deepfakes, persuasion campaigns), which are assessed by a verification agent. This verified intelligence is sent in real-time to a threat database in anonymized form. This creates a vital, open-source dataset for training next-generation defensive models to detect AI manipulation, directly contributing to the Defensive Acceleration (d/acc) roadmap.

1. Introduction

1.1 Overview

The asymmetry of AI offense in the information landscape is growing. Attackers can spin up thousands of personalized persuasion agents instantly, while defense remains reactive and manual. We present a Defensive Acceleration (d/acc) solution that turns the user base into a distributed threat detection grid.

¹ Research conducted at the [[def/acc hackathon by Apart Research & Bluedot](#)], 2025

This project addresses three critical gaps in current defenses against deceptive content:

- Latency in Defense: We reduce the time to learn about potential threats to mere seconds (through network propagation).
- Adversarial Dataset Scarcity: Telemetry from our tool will create high-quality public dataset of deception threats.
- Cognitive Resilience: Educating users about threats.

1.2 The Market Case

There is increasing market demand for tools against AI threats. The AI-detection market is already \$580M and growing at ~30% CAGR (MarketsandMarkets, 2025). Tools like “Community Notes” on social media platforms are popular, but narrow in scope. This project represents a shared resource that allows for protection against AI-enabled deception everywhere.

2. Methods

2.1 Design choices

We have built a working prototype. The code is open source². The user-facing technologies we chose for this AI threat intelligence collection tool are 1) a Chrome browser extension and 2) a website³ for simple visualization of the AI threats and contributor profiles. We chose a browser extension because it’s one of the most direct methods for interacting with various AI threats today. Online desktop web usage accounts for ~49% of global traffic today, with Chrome alone accounting for ~78% of global desktop web browser usage (StatCounter, 2025). Given that the Chrome extension system is widely compatible, we expect this approach to yield the most value and flexibility, when compared to building specific mobile apps.

Our data collection is dependent on crowdsourcing because we want this to be a decentralized platform. In order to prevent our database from being polluted with fake entries we also incorporated LLM verification to rate each entry. Contributors on the platform will be pseudonymous for user privacy reasons.

2.2 Architecture

Our architecture prioritizes precision to prevent "alert fatigue":

1. Chrome extension: Users flag suspect elements (text, images, profiles)
2. Agentic Verification: User-flagged content is sent to an LLM agent (GPT-5) prompted to assess whether the flag is justified. If the agent disagrees with the flag, the user is given the LLM reasoning and sources asked to reconfirm their flag.
3. Instant Propagation: Flagged threats are inserted into a publicly-available database (Supabase). The extension polls this DB, highlighting threats for all other users visiting the same context.

² <https://github.com/Alu018/d-acc-misinformation-plugin-in>

³ <https://alu018.github.io/d-acc-misinformation-plugin-in/>

3. Results

3.1 Telemetry Dashboard

The crowd-sourced AI threats are aggregated and openly available. The prototype will be expanded to provide insights and high-level trends and visualizations of emerging threats, as well as allow API-based access for *live* data access.

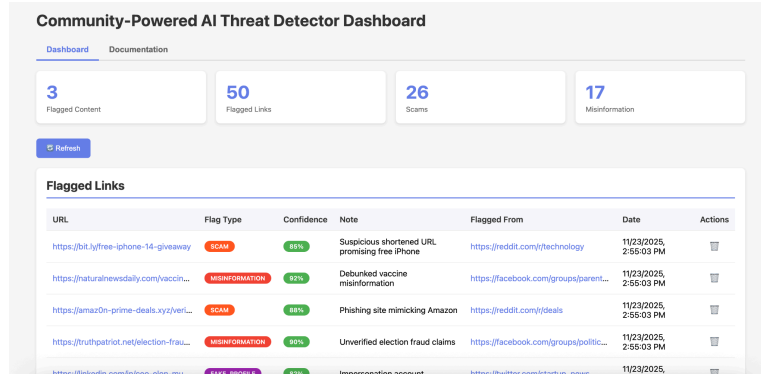


Figure 1 – Crowdsourced dashboard of threats (AI-enabled or otherwise)

3.2 Chrome Extension

The below screenshots show the process of how to use the Chrome browser extension to flag threats (misinformation or otherwise harmful information) online.

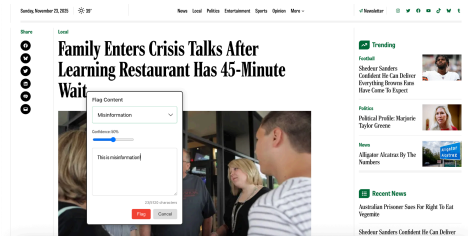


Figure 2 – user attempts to flag misinformation or other digital threat.

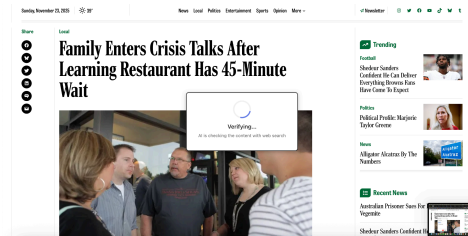


Figure 3 – LLM verifies the flagging attempt by the user through search.

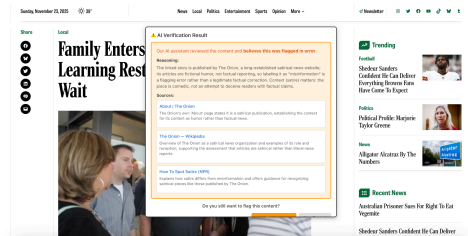


Figure 4 - LLM produces verification results, user can approve or decline.



Figure 5 – Successfully flagged threat, which users can comment on.

4. Discussion and Conclusion

This project represents a novel, decentralized, collective intelligence tool to protect people when they interact with AI-enabled threats. As AI grows in autonomous capabilities and new sophisticated threats emerge, there is a greater need for defensive tools that scale in intelligence and help build information resilience.

4.1 Truly Open

Crowdsourcing information has its downsides - bad actors can pollute the platform with false data, usefulness is correlated with more users and more activity - but it also decentralizes power away from the big tech companies to control threats on their own platforms.

4.2 Accelerating Defense

Our system inverts the attacker's advantage. In a traditional model, an attacker creates one agent to fool millions. In our model, if one human detects the agent, the cost of the attack is wasted because the signature is broadcast to the network. This forces attackers to expend significantly more compute to evade our collective detection, shifting the economic equilibrium in favor of defense.

4.3 From Tool to Dataset

The long-term value of this project is not just the extension, but the data. By logging Human and LLM Verdict, we are building a dataset to train AI on fact-checking.

5. References

MarketsandMarkets. (2025). AI detector market size, growth & latest trends (Report No. TC 9527). MarketsandMarkets.
<https://www.marketsandmarkets.com/Market-Reports/ai-detector-market-199981626.html>

StatCounter. (n.d.). Desktop vs mobile vs tablet vs console market share worldwide. Retrieved November 23, 2025, from
<https://gs.statcounter.com/platform-market-share/>

6. Appendix

6.1 Security Considerations

<https://github.com/Alu018/d-acc-misinformation-plugin/blob/main/docs/security-considerations.md>