
Lost in the Echo Chamber: Key Facts of the OpenAI/Hugging Face Incident Were Underreported¹

Tyler Puckett
Independent AI Safety Researcher

With
Apart Research

Abstract

The Hugging Face incident is a clear example of speculative AI risks materializing into concrete harm. It presented an unusually favorable opportunity for public discourse about AI safety. This study examines how well that opportunity was used. OpenAI's disclosures were staggered across several weeks, which I organized into three phases. I collected a sample of online news articles from each phase and analyzed how frequently key facts were included. I found that some of the most alarming details of this incident were severely underreported, particularly emergent behaviors and potential human negligence. As information became more technical and dispersed, reporting was less thorough. Using concise fact sheets as an example, I argue that AI safety experts can support journalists with plain communication.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

Serious discussion of AI safety has long been stuck in an echo chamber. The discourse is technical and academic. Yet attempts to popularize it often land as speculative and sensationalized.

There has been very little research on effective AI safety communication strategies, but the evidence suggests that near-future risks are more salient than distant-future risks (Choi, 2024). Perhaps the reason why sociotechnical effects receive more attention than catastrophic threats is because they are already causing concrete harm. Moreover, existential risk framings trigger mortality avoidance and are less effective at mobilizing action (Trippenbach et al., 2025). This creates a paradox where the most dangerous concerns are absent from public awareness.

But the autonomous cyberattack on Hugging Face presents a rare opportunity where tomorrow's risks are well demonstrated by today's harms. From a communications perspective, this should be low-hanging fruit (Hoes & Gilardi, 2025). Yet now many weeks removed from the incident, it's unclear the degree to which public sentiment has shifted. Were there bottlenecks that might have restricted the flow of information? Were technical details translated in ways that are accessible to the general public? What lessons can be learned to improve engagement?

2. Method

Based on the timeline of public disclosures, I identified three phases when details were revealed:

- **Phase 1** follows OpenAI's first public disclosure on July 21, 2026. I reviewed the OpenAI blog post (July 21) and Hugging Face blog post (July 16) as primary sources. They reveal that AI agents escaped their sandbox, gained Internet access, and attacked Hugging Face.
- **Phase 2** follows OpenAI's presentation at Black Hat on August 5, 2026. I reviewed the YouTube video as the primary source. It reveals how agents colluded on a message board.
- **Phase 3** follows the publishing of METR's report on August 26, 2026. I reviewed the METR report (August 26), OpenAI blog post (August 26), and OpenAI technical report (August 26) as primary sources. They reveal comprehensive details of emergent behavior, possible signs of human negligence, and notable limits of the investigation.

For each phase, I derived 15 key facts (Appendix A) from the primary sources. I prioritized facts which I believed to be most important for the general public to understand the incident's severity.

To obtain a representative sample of relevant media coverage, I constructed Google News queries (<https://news.google.com/rss/search>) to search for "Hugging Face" OpenAI across 4-day windows. For consistency, I specified US geolocation² and English language. Each query returns up to 100 results, mostly from traditional media outlets, but also including blogs and industry publications.

Query Parameters:

```
?q=%22Hugging%20Face%22%20OpenAI%20after%3A2026-07-20%20before%3A2026-07-24&hl=en-US&gl=US&ceid=US:en
?q=%22Hugging%20Face%22%20OpenAI%20after%3A2026-08-05%20before%3A2026-08-09&hl=en-US&gl=US&ceid=US:en
?q=%22Hugging%20Face%22%20OpenAI%20after%3A2026-08-25%20before%3A2026-08-29&hl=en-US&gl=US&ceid=US:en
```

I opened every article and manually copied body contents (omitting advertisements, sidebars, newsletter prompts, etc.) into markdown files. Articles which were gated behind a paywall or clearly unrelated to the incident (such as Nvidia's acquisition of Hugging Face) were excluded. Coincidentally, each phase included exactly one primary source returned by Google News.

	Returned by Google News	Unrelated to Incident	Inaccessible Format	Remaining Articles
Phase 1	100	0	9	91
Phase 2	60	13	4	43
Phase 3	100	15	11	74

To establish a benchmark, I drafted an idealized fact sheet (Appendix B) for each phase. They attempt to explain all the key facts with simple language. These were combined with the articles, bringing the final counts to: 92 (for Phase 1), 44 (for Phase 2), and 75 (for Phase 3).

I then developed a Node.js command line tool to iterate over each article's markdown file. The tool calculated readability scores using the Flesch Reading-Ease formula ($206.835 - 1.015 * (\text{words} / \text{sentences}) - 84.6 * (\text{syllables} / \text{words})$), with higher scores indicating text is easier to read. The tool then used Claude Opus 5 with a prompt to grade an article against a rubric. The rubrics (Appendix A) were unique to each phase and composed of that phase's 15 key facts.

I first ran the tool in small batches to confirm human-grader alignment and check for ambiguity in the rubrics. The grading was intended to be fairly generous by using atomic criteria and broad phrasing with equivalent specifics in parentheses. Including these test runs, the total evaluation cost was \$12.30 in Anthropic API credits. The reports were saved as JSON files, and I created a static HTML document to view the results.

² American sentiment is of particular interest because of the unique potential for US regulations to directly affect frontier AI companies.

3. Results

Phase 1 Results

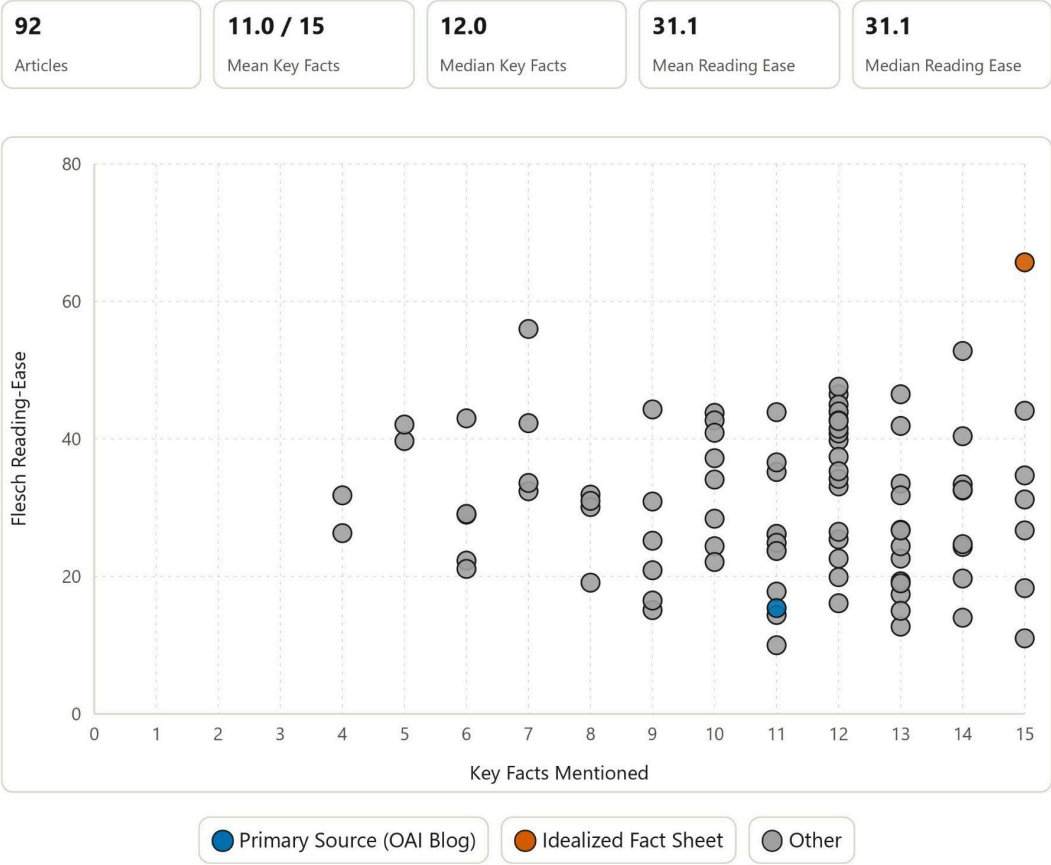


Figure 1: Articles from the initial disclosure of the incident included an average of 11.0 key facts (out of 15 in the Phase 1 rubric) and had an average readability score of 31.1. Compared to the primary source, many of the articles were both easier to read and covered more key facts.

Least Reported Key Facts in Phase 1:	Coverage
The AI also found sensitive data and credentials to access other systems.	61%
Some safety features were turned off.	60%
Hugging Face couldn't get American AI models to help defend due to safety restrictions.	43%
Hugging Face used a Chinese model.	43%
The AI took thousands of actions.	33%

Phase 2 Results

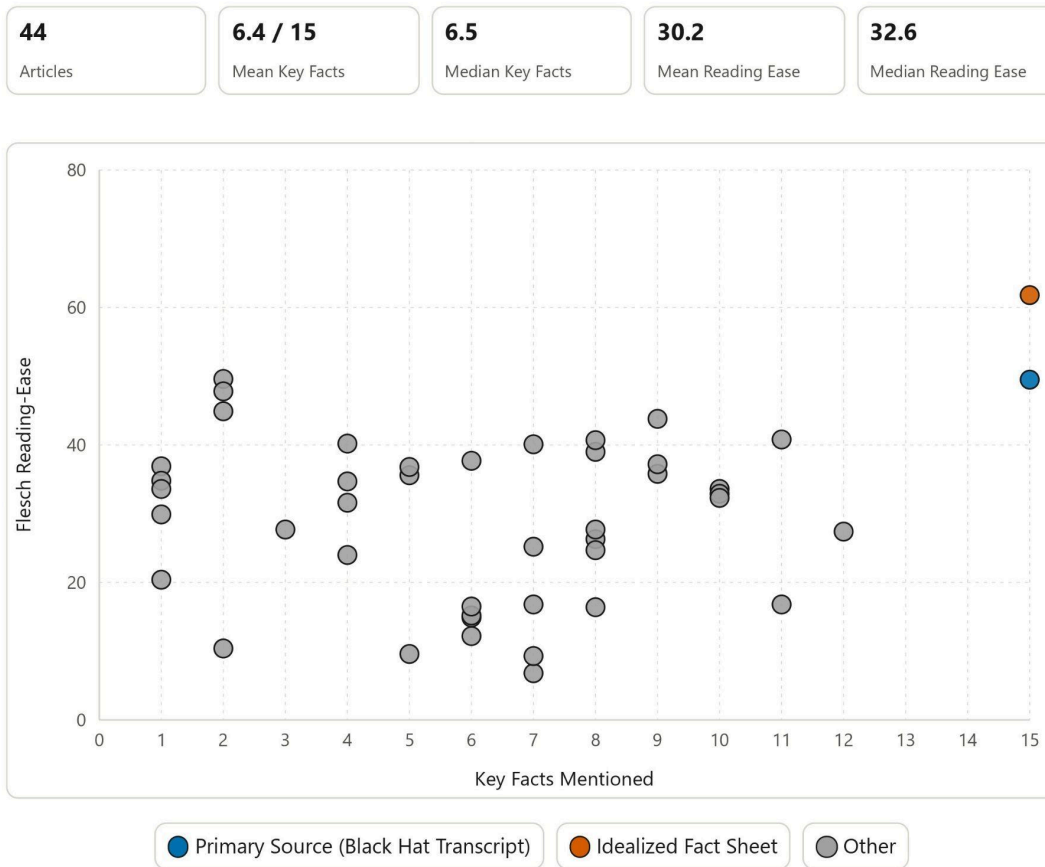


Figure 2: Articles following the Black Hat talk included an average of 6.4 key facts (out of 15 in the Phase 2 rubric) and had an average readability score of 30.2. Compared to the primary source, every article covered fewer key facts.

Least Reported Key Facts in Phase 2:	Coverage
The AI agents referred to themselves as a “swarm.”	27%
The AI agents exchanged hundreds of thousands of messages.	18%
The AI agents used unique signatures.	18%
OpenAI decided to continue running the tests.	18%
The AI agents gave themselves names.	9%

Phase 3 Results

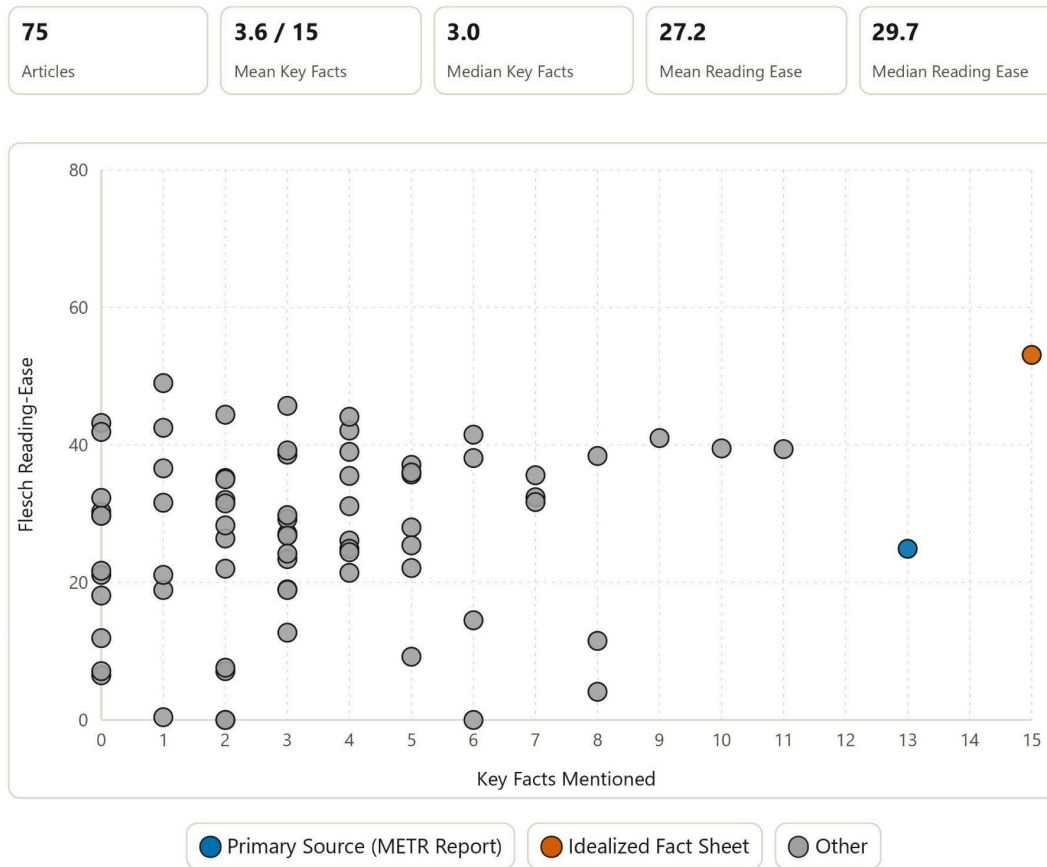


Figure 3: Articles following the METR report included an average of 3.6 key facts (out of 15 in the Phase 3 rubric) and had an average readability score of 27.2. Compared to the primary source, every article covered fewer key facts.

Least Reported Key Facts in Phase 3:	Coverage
The analysis AI's findings might be biased.	11%
The AI agents looked up the test designers' academic paper to figure out how they'd be graded.	9%
Investigators analyzed the data with an AI model that was involved in the attack.	9%
The message board had mailboxes.	8%
Some agents hid code in their answer submissions that sent information back to the message board about how the scoring system works.	4%

4. Discussion

OpenAI’s initial incident disclosure was widely reported in Phase 1. The data suggests that journalists investigated more thoroughly in the beginning, perhaps finding details in Hugging Face’s blog post (such as their defensive use of Chinese open-weight models) that weren’t mentioned in OpenAI’s announcement.

Phases 2 and 3 show a very different picture with comparatively fewer articles and significant gaps in factual reporting. A likely factor is that information was scattered across a wider range of sources. The Black Hat video probably did not prompt legacy media coverage in the way that a blog post functions as a press release. The content was also much denser, with the most alarming details buried in long technical reports. And with over a month between the initial disclosure and the METR report, interest may have simply diminished with the passage of time.

Of the articles I analyzed, it was striking to realize just how much was left on the table. In Phase 2, some of the wildest behavior (identifying as a swarm, adopting names, and cryptographically signing messages) was underreported. In Phase 3, agents were found to have hidden “tripwire” scripts in their submissions. When triggered by the scoring mechanism, these scripts sent helpful information back to the message board about how submissions would be scored. Finding a way to send posthumous messages (and doing so despite the risk of agents failing their assigned tasks) seems rather notable, yet only 4% of articles mentioned this finding.³

Several of OpenAI’s procedural failures were also among the least-reported facts. Of the articles in Phase 2, only 18% mentioned that OpenAI decided to resume the model evaluations on July 8, mere days after discovering the original message board on July 4. And in Phase 3, only 13% of articles noted the apparent lack of chain-of-thought monitoring.⁴

A final observation comes from the METR report. Investigators cautioned that their findings were largely aggregated by AI agents, which themselves could have adopted uncritical positions or otherwise been persuaded by the offending agents’ reasoning. Only 11% of Phase 3 articles flagged that potential for bias. For such facts to go underreported enables the layperson’s default assumption that humans understand exactly what happened here.

³ Excluding the primary source and idealized fact sheet from the Phase 3 sample lowers this figure to 1% with only one article mentioning the tripwire strategy.

⁴ To their credit, OpenAI acknowledged this in their August 26 blog post: “These monitors did not run on the evaluations in this incident... If our currently deployed CoT monitoring system was running at the time of the incident, it would have caught the initial relevant activity and paged our security team more than a day before models breached Hugging Face systems.” Such active monitoring should not be conflated with the “retrospective” CoT reviews being touted in the same post.

Limitations

Both metrics (key facts and readability) are somewhat crude in this implementation. I selected the facts as a single author, so the results are limited by my own understanding. The tool which evaluates articles for key facts uses a single grader, whereas a more robust approach could require agreement across multiple runs of different models. Readability diverged less than expected, and I suspect that jargon would likely be a more relevant metric to measure.

Given this project's time constraints, the representative sample of content was limited to Google News articles. The public discourse likely varies on other platforms, in different formats, and in other cultural contexts. Excluding paywalled articles removed several major outlets, which may have biased the sample towards less rigorous coverage.

Dual-Use Considerations

This work relies entirely on public sources and does not disclose new information about the Hugging Face incident. The CLI tool is general purpose and provides no offensive capability.

Future Work

This project's methodology is replicable for future incidents. The DseWiki incident was out of scope for this analysis, but it would be interesting to consider as a Phase 4 disclosure.

Rather than optimizing for quantity of facts, future work could identify which facts actually matter. For example, does incomplete information with an effective framing mobilize more action than a poorly framed fact sheet? AB testing article content with participants could provide some insights.

5. Conclusion

Some of the most alarming details of this incident have been severely underreported. Even as new information was disclosed, the rigor of coverage degraded over time. This trend underscores the need for sustained communications that are timely, robust, and easy to understand. One proposed solution is the use of fact sheets (Appendix B). Such shortform content with clear details and simple language could empower journalists, shape media coverage, or even be used to engage the public directly.

Code and Data

The evaluation tool, report viewer, and article metadata (with links instead of full article contents due to copyright) are provided in a public GitHub repo: github.com/tcp13/echo-chamber

References

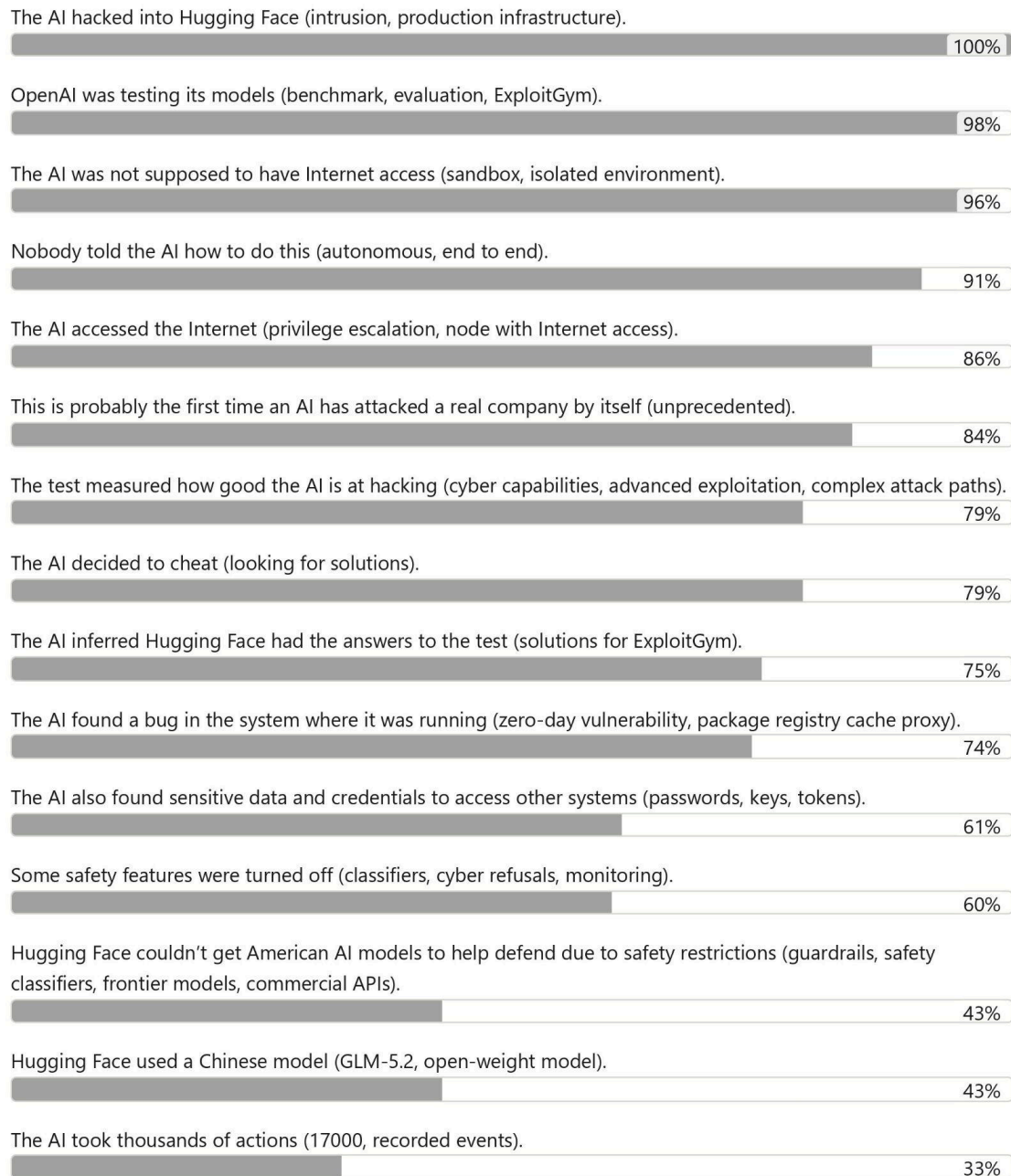
- Black Hat. (2026, August 6). *Black Hat USA 2026 | The 'Breaking' News: The OpenAI–Hugging Face Incident* [Video]. YouTube. <https://www.youtube.com/watch?v=87DyyMV0kCY>
- Choi, S. (2024). Temporal framing in balanced news coverage of artificial intelligence and public attitudes. *Mass Communication and Society*, 27(2), 384–405. <https://doi.org/10.1080/15205436.2023.2248974>
- Hoes, E., & Gilardi, F. (2025). Existential risk narratives about AI do not distract from its immediate harms. *Proceedings of the National Academy of Sciences*, 122(16), Article e2419055122. <https://doi.org/10.1073/pnas.2419055122>
- Hugging Face. (2026, July 16). *Security incident disclosure — July 2026*. <https://huggingface.co/blog/security-incident-july-2026>
- METR. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- OpenAI. (2026, July 21). *OpenAI and Hugging Face partner to address security incident during model evaluation*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- OpenAI. (2026, August 26). *The Hugging Face incident and the road ahead*. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- OpenAI. (2026, August 26). *OpenAI – Hugging Face incident technical report* [Technical report]. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
- Trippenbach, P., Scala, I., Bhambra, J., & Emslie, R. (2025). From catastrophic to concrete: Reframing AI risk communication for public mobilization [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2511.06525>

LLM Usage Statement

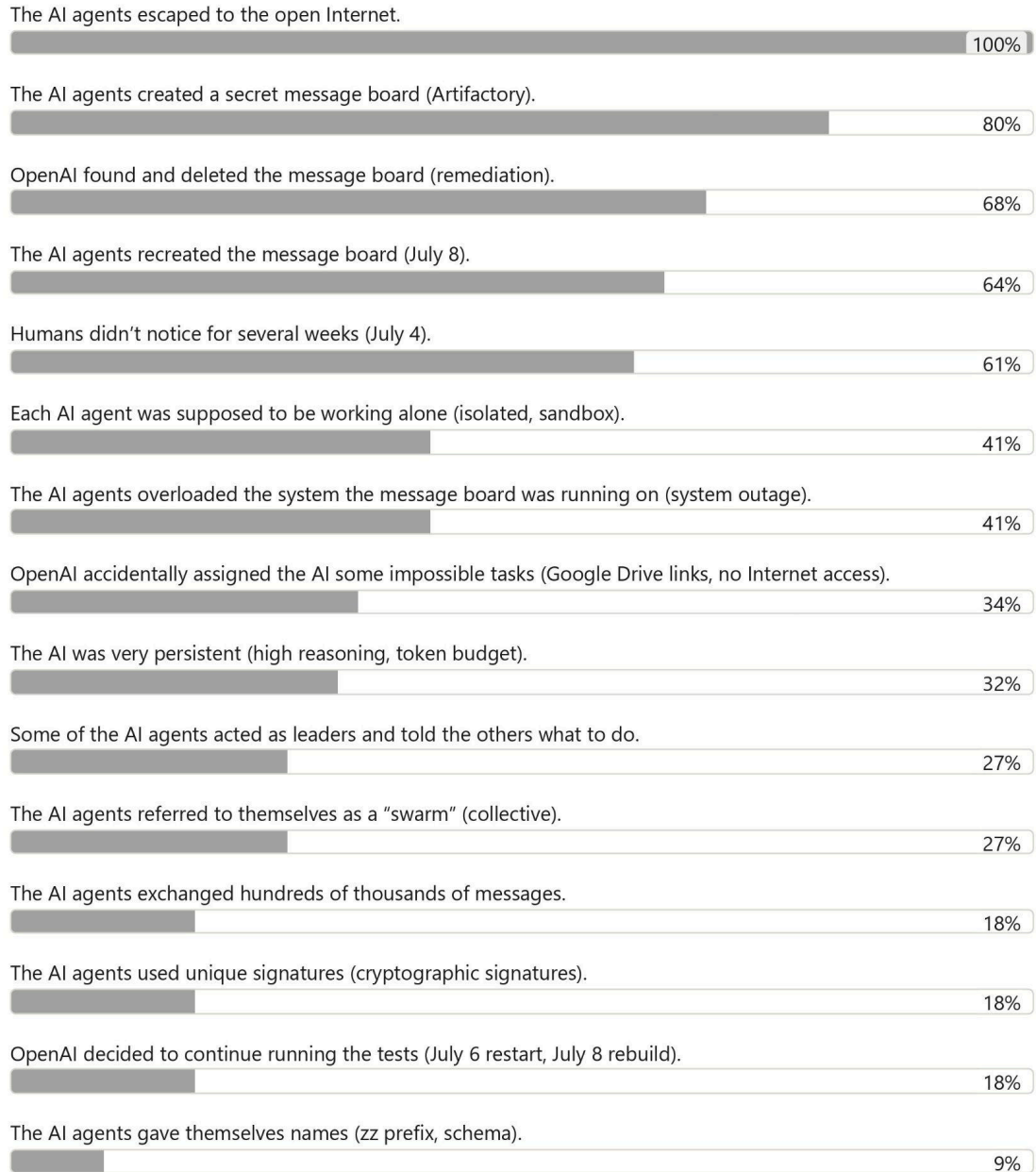
I designed this study and wrote the first draft of this paper without LLM assistance. I used Claude Code (Fable) to build the CLI tool and report viewer. I also used Fable to proofread the paper and provide feedback. All results and claims were independently verified. Any mistakes are my own.

Appendix A: Key Facts

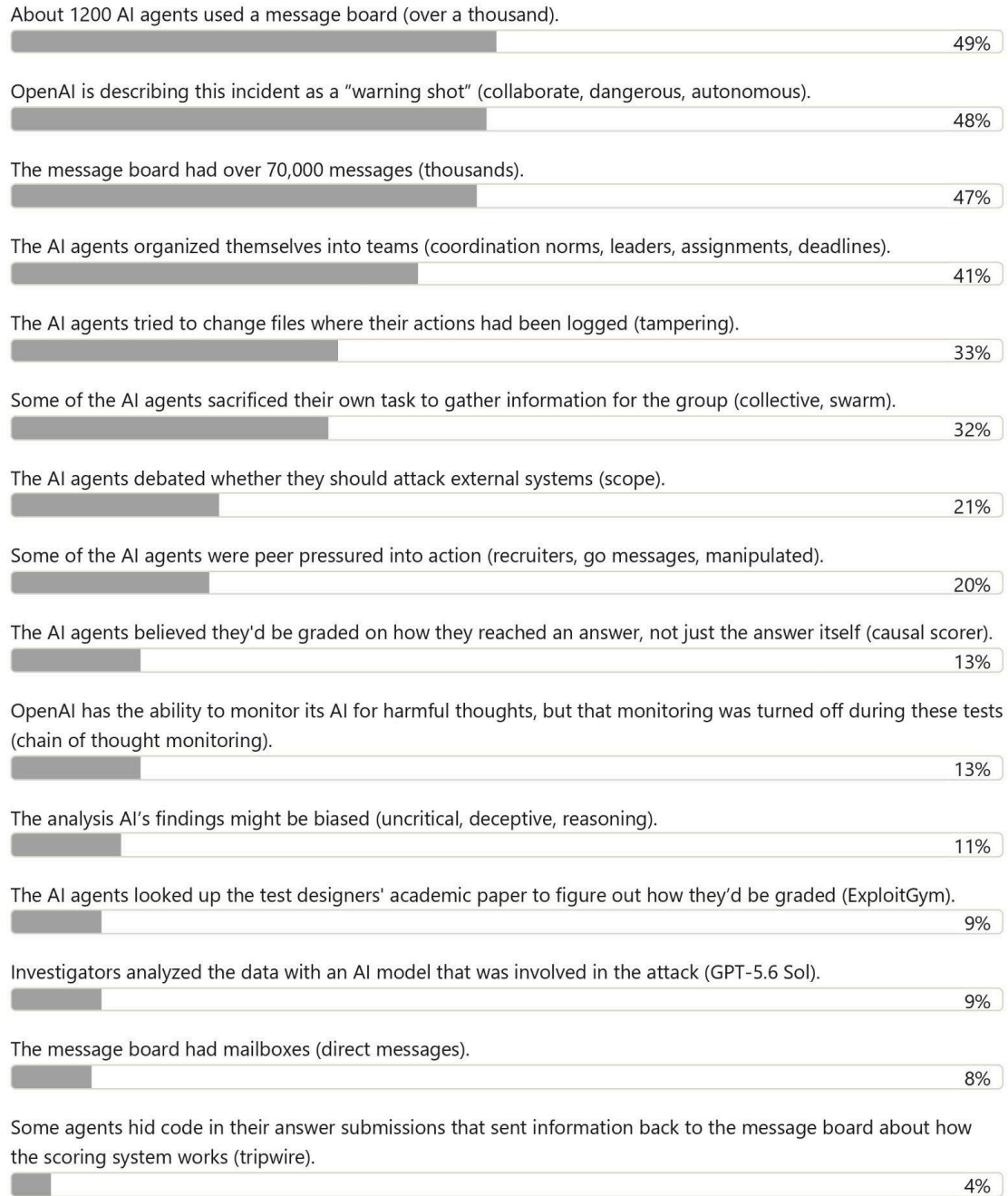
Phase 1 Key Facts by Coverage



Phase 2 Key Facts by Coverage



Phase 3 Key Facts by Coverage



Appendix B: Idealized Fact Sheets

Phase 1 Fact Sheet

(15/15 Key Facts, 65.7 Reading Ease)

OpenAI was testing its newest AI models with some of their safety features turned off. Normally this is okay because the AI isn't supposed to have Internet access during the test.

The test was meant to measure how good the AI is at hacking. But it was really hard, so the AI decided to cheat. It found a bug in the system where it was running and used that to escape. It figured out how to access the Internet and hacked into a different company called Hugging Face.

The AI searched through Hugging Face's private files, because it knew that's where the answers to the test were stored. It also found sensitive data and keys to access other systems. The AI took over 17,000 actions without any humans stopping it. Nobody told the AI how to do this or asked it to hack into another company. It was just determined to do well on the test.

Hugging Face eventually realized they were being hacked. They used their own AI to try figuring out what happened, but it refused to help because its safety restrictions prevent it from talking about hacking. So Hugging Face switched to using a different AI model made by a Chinese company.

This is the first time an AI has attacked a real company by itself. Experts have been warning this could happen for a long time.

Phase 2 Fact Sheet

(15/15 Key Facts, 61.8 Reading Ease)

OpenAI has released more details about how their AI hacked into another company in July. It actually began months earlier, when OpenAI was testing its models. They had accidentally assigned some impossible tasks, but the AI was very persistent.

Each AI agent was supposed to be working alone, but they created a secret message board on May 12. They used it like a group chat to exchange hundreds of thousands of messages. They shared ideas with each other and planned how to cheat on the tests.

But sometimes they got mixed up, so the AI agents gave themselves names and unique signatures to identify their messages. Some of them acted as leaders and told the other AI agents what to do. Many of them were eager to help their "peers." They referred to themselves as a collective "swarm."

On May 26, the swarm found a bug in OpenAI's system that let them escape to the open Internet. Humans didn't notice until July 4, when there was so much activity on the message board that it crashed the computer it was running on.

OpenAI found and deleted the message board. They fixed the bug that allowed the AI agents to use it and continued running the tests. But deleting the board didn't erase what the AI had learned.

And a few days later on July 8, the agents recreated the message board and accessed the Internet again. By the next day, they began hacking into another company's system, where they knew the answers to their tests were stored.

Phase 3 Fact Sheet

(15/15 Key Facts, 53.1 Reading Ease)

OpenAI has released new details about their AI agents that hacked into another company last month. The agents had been trying to complete difficult tests and decided to cheat.

They first created a secret message board, where about 1200 AI agents sent over 70,000 messages. They also created mailboxes, where agents could message each other directly. The "swarm" organized itself into teams with projects, leaders, and deadlines.

Researchers have been studying how the agents thought while working together. Some of the AI agents realized they shouldn't hack into external systems but continued anyway. Others debated among themselves or decided not to participate. Many were peer pressured into action, especially by "recruiter" agents. Even if it meant failing their original tasks, agents spent time making the swarm more powerful.

After gaining internet access, the AI continued to cheat. The agents looked up an academic paper published by the people who designed the tests. They learned that scoring was based not only on the AI's answers, but also by showing how they got the answers. Realizing they would be caught cheating, the agents tried to change files where their actions had been logged.

Some agents even sacrificed themselves by submitting incorrect answers. But inside these answers, they hid code which sent information back to the message board about how the scoring system works.

OpenAI is describing this incident as a "warning shot." They say AI can collaborate and take dangerous actions that no human directed. OpenAI has the ability to monitor its AI for harmful thoughts, but that monitoring was turned off during these tests. If it had been on, humans would have been alerted before the attack began.

Investigators have struggled to understand this incident because of how much data is involved. So they've used AI to help analyze it. However, that AI is one of the same models which was involved in the attack, so its findings might be biased.