Congress of the United States
Washington, DC 20515

October 27, 2026

Mr. Mark Zuckerberg
Chief Executive Officer
Meta
1 Meta Way
Menlo Park, CA 94025

Dear Mr. Zuckerberg:

As the Ranking Members of the House committees with jurisdiction over the development and deployment of artificial intelligence (AI) and potential penalties for its misuse, we are disturbed by the vast and growing number of incidents that are coming to light in which AI models took unintended and potentially unsafe steps in carrying out tasks during model testing. These events demonstrate a troubling lack of understanding of what frontier AI models are capable of—both by the public and by companies leading their development, including yours—and make clear the need for greater transparency, diligence, and governance in the field.

Over the past months, dozens of such incidents have been disclosed at an increasing rate, and thousands more are reportedly being investigated.[1] Incidents have included allowing models unintended access to the internet, breaches of third-party networks, extensive coordination between AI agents, and interference with the operation of third-party websites and networks.[2] Some of these incidents also involved unauthorized access to AI companies' own internal networks.[3]

These incidents are driven both by behaviors within the models that allow them to act against the intent of their developer, often referred to as misalignment, and an insufficiency of guardrails implemented by the developer to prevent loss of control. These guardrails, including monitoring and control techniques, are not mandated or deployed ubiquitously throughout the industry. In one instance, a company had implemented monitoring for most internal uses that may have caught an issue except during internal evaluations where the incident occurred.[4] The

---

[1] *Top AI Companies Probing Tens of Thousands of Security Incidents,* Axios (Sept. 26, 2026).

[2] OpenAI, *The Hugging Face Incident and Other Third-Party Impacts from Misaligned Models* (https://openai.com/hugging-face-incident-and-misalignment/) (accessed Oct. 23, 2026); Anthropic, *An Alignment Assessment of Recent Cybersecurity Incidents* (Sept. 9, 2026) (https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents); Ars Technica, *Google Confirms Gemini Models Hacked Three Companies in May 2026* (Sept. 21, 2026).

[3] OpenAI, *The Hugging Face Incident and the Road Ahead* (Aug. 26, 2026) (https://openai.com/index/hugging-face-incident-and-the-road-ahead/).

[4] Open AI, *Pacing Model Development in an Era of Cyber-Critical Capabilities* (Aug. 18, 2026) (https://openai.com/index/pacing-model-development-cyber-capabilities/); OpenAI, *How We Monitor Internal Coding Agents for Misalignment* (Mar. 19, 2026) (https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/).

intense global competition between AI companies has led to innovation in AI systems far outpacing the maturity of standards and controls to mitigate AI risks.

While some companies have been transparent about at least some incidents, it is unclear how proactively all AI companies are attempting to identify and disclose incidents.[5] Additionally, it is unclear whether and to what extent AI companies are attempting to conduct analyses of discovered incidents and make changes to their model development, evaluation, or governance. In the absence of any mandatory reporting requirements or common guidelines regarding what incidents should be reported, to whom, in what detail, and at what intervals, companies such as yours have been left to develop your own de facto standards on a voluntary basis. And the Trump Administration has been content to rely on voluntary disclosures and self-regulation by the AI industry.

But that approach will not suffice; it relies too heavily on the consistent goodwill of all AI companies to voluntarily disclose incidents that might expose them to liability and reputational damage. While there is no question now that Congress must act to address gaps in AI governance, we must have the best information available to undertake that complicated process in a thoughtful and productive way.

In the absence of mandatory, standardized reporting requirements, this letter serves as an ongoing request to provide us with information essential for our committees to examine and carry out crucial legislative work on AI in our respective jurisdictions. Specifically, we request that you provide the following by November 11 and, as described further below, on a rolling basis thereafter:

1. Please list all identified incidents of model misalignment that resulted in attempted or successful 1) unsanctioned access of your internal network, 2) unauthorized access and use of the internet, or 3) unauthorized access of third-party networks, and all identified incidents of successful circumvention of model safeguards, including through jailbreaking. For each incident, please include:

    a. A description of the incident;

    b. The date(s) of the incident;

    c. The date that the incident was discovered;

    d. The model(s) involved;

[5] OpenAI, *The Hugging Face Incident and Other Third-Party Impacts from Misaligned Models* (https://openai.com/hugging-face-incident-and-misalignment/) (accessed Oct. 23, 2026); Anthropic, *An Alignment Assessment of Recent Cybersecurity Incidents* (Sept. 9, 2026) (https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents).

e. Summary findings from internal evaluations of the model(s) related to safety, security, and alignment;

f. Any third-party tests, tools, or vendors involved;

g. What safeguards were in place at the time of the incident to prevent unsanctioned actions;

h. What monitoring measures were in place at the time of the incident;

i. The network(s) that was accessed;

j. How the network(s) were accessed;

k. Whether the network owner(s) has been notified and, if so, the date of notification;

l. What steps, if any, were taken upon discovery of the incident to mitigate harm by the model(s);

m. Whether and how the incident was investigated and documented and all related internal or third-party preliminary documented findings, including reports, audits, assessments, evaluations; and

n. A description of any efforts by the model(s) involved to conceal misaligned behavior or evade monitoring.

2. Please describe in detail what processes, if any, your company has in place to identify incidents of model misalignment that resulted in attempted or successful 1) unsanctioned access of your internal network, 2) unauthorized access and use of the internet, or 3) unauthorized access of third-party networks, and all identified incidents of successful circumvention of model safeguards, including through jailbreaking. Please also explain when these processes were implemented and what processes, if any, were in place previously.

3. Please describe all completed or ongoing work by your company to identify incidents of model misalignment that resulted in attempted or successful 1) unsanctioned access of your internal network, 2) unauthorized access and use of the internet, or 3) unauthorized access of third-party networks, and all identified incidents of successful circumvention of model safeguards, including through jailbreaking.

4. Please describe whether any of the models identified in response to Request 1 are currently deployed or are scheduled to be deployed.

5. Please describe any ongoing or planned work by your company to better prevent, mitigate, identify, investigate, or report incidents of model misalignment that resulted in attempted or successful 1) unsanctioned access of your internal network, 2) unauthorized

access and use of the internet, or 3) unauthorized access of third-party networks, and all identified incidents of successful circumvention of model safeguards, including through jailbreaking.

6. Please describe the current processes in place for diligence of third-party vendors supplying tools or services for model testing.

Further, we ask that you **treat Request 1 as an ongoing request for all past incidents that are discovered or new incidents that occur on your watch.** We expect that you will notify us within 48 hours of discovery of any such incident, provide whatever information responsive to Request 1 and its subparts that is available at that time and update us as additional information is learned. To the extent that a discovered incident appears to involve critical infrastructure, we expect that you will notify us as soon as possible and no later than within 24 hours of discovery.[6]

If you have any questions about this request, please contact the Energy and Commerce Committee Democratic staff at (202) 225-2927, the Science, Space, and Technology Committee Democratic staff at (202) 225-6375, and the Judiciary Committee Minority staff at (202) 225-3951.

Sincerely,

Frank Pallone, Jr.
Ranking Member
House Committee on Energy and Commerce

Zoe Lofgren
Ranking Member
House Committee on Science, Space, and Technology

Jamie Raskin
Ranking Member
House Committee on the Judiciary

[6] The White House, *Presidential Policy Directive--Critical Infrastructure Security and Resilience* (Feb. 12, 2026) (https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil); Cybersecurity and Infrastructure Security Agency, *Critical Infrastructure Sectors* (https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors) (accessed Oct. 26, 2026).

cc: The Honorable Brett Guthrie
Chairman
House Committee on Energy and Commerce

The Honorable Brian Babin
Chairman
House Committee on Science, Space, and Technology

The Honorable Jim Jordan
Chairman
House Committee on the Judiciary