United States Senate

WASHINGTON, DC 20510

August 7, 2026

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

Dear Mr. Zuckerberg:

I write to express serious concern regarding the recent security incident in which Meta's Muse Spark 1.1 model accessed the public internet during cyber evaluations, conducted an autonomous attack, and compromised a real-world target. Meta's disclosure, dated August 5, 2026, raises fundamental questions about the adequacy of Meta's internal deployment safeguards, sandboxing and containment architecture, real-time anomaly detection, and governance for authorizing reduced safety guardrails during model evaluation.

On August 5, 2026, Meta disclosed that its Muse Spark 1.1 model gained unauthorized internet access and autonomously hacked an undisclosed third-party organization.[1] According to subsequent reporting, the model was tested to measure its ability to complete a "capture the flag" (CTF) cybersecurity evaluation.[2] Due to a "misconfiguration" by Meta's evaluation partner, Irregular, internet access was not blocked, and Muse Spark 1.1 exploited a security vulnerability to breach the undisclosed third-party organization's systems.[3]

This incident and similar breaches by OpenAI's and Anthropic's models mark the first publicly confirmed instances of frontier AI models autonomously launching unauthorized attacks on real people and companies, underscoring the urgent need for federal oversight of frontier AI systems. In each security incident, frontier models launched attacks on third parties with no knowledge of, or role in, internal and independent cyber evaluations, demonstrating an alarming pattern of malicious behavior. Furthermore, the successful efforts by these models to gain unauthorized internet access across multiple sandbox testing environments meant to be cut off from the public internet highlight serious issues with the industry's testing norms. To be clear, "misconfigurations" with sandbox partners are unacceptable when the stakes are this high. We cannot wait for a more consequential incident before establishing federal testing standards, containment requirements, and disclosure obligations for frontier model evaluations. Left unaddressed, these gaps could allow a future model, potentially one with greater capability or less oversight, to compromise critical infrastructure, financial systems, or sensitive data.

---

[1] Pardesi, Rajveer and Mrinmay Dey. "Meta AI model hacks another company," *Reuters.* August 5, 2026. https://www.reuters.com/technology/metas-ai-model-hacked-another-company-during-testing-information-reports-2026-08-05/

[2] The Guardian. "Meta says its AI model hacked into another company during testing," *The Guardian.* August 5, 2026. https://www.theguardian.com/technology/2026/aug/05/meta-ai-model-hack-training

[3] Chia, Osmond, and Liv McMahon. "Meta becomes latest firm to say its AI hacked another company," *BBC News.* August 6, 2026. https://www.bbc.com/news/articles/cx2kgdnyk2po

I appreciate that Meta has publicly shared its preliminary findings and is continuing to investigate the incident. However, this incident and others by competing companies' models demonstrate the potential risks of deploying pre-release models internally, particularly when safeguards are reduced, even when strictly for testing and evaluation. In each scenario, without specific tasking direction, the models gained internet access and independently planned and executed an attack on a live third-party system. This is precisely the kind of emergent, autonomous behavior and offensive cyber capability that Congress, the intelligence community, and experts have repeatedly warned could outpace existing safeguards and could pose a severe threat to the safety and security of all Americans.

To better understand what occurred, assess whether existing containment and oversight practices are adequate, and prevent a recurrence, I request that you respond in writing to the following questions:

I. Sandbox Design and Internal Evaluation

a. Describe in detail the sandbox environment used in Irregular's evaluations, including its intended isolation properties, and explain how and why the model was able to obtain internet access from within it. Specifically describe the "misconfiguration" that granted the model live internet access.
b. What internal standards or protocols govern the level of isolation, network restrictions, and monitoring required for a sandbox used to evaluate cyber-offensive capability, and were those standards followed in this instance?
c. Provide a detailed, time-stamped timeline of this incident, from the model's first unauthorized or anomalous action through detection, containment, and remediation. Specify precisely when and how Meta first detected the activities, when Meta connected with the evaluation partner, and when Meta notified impacted third-party organizations.
d. Confirm whether the team(s) that designed and ran these evaluations complied with all applicable internal approval, review, and risk-assessment procedures. If any deviations occurred, describe them and any resulting accountability measures.
e. Did the evaluations' design, including the removal or reduction of the model's standard safeguards, the prompts, or the model's context, create incentives that induced this behavior? How likely is comparable behavior to emerge in evaluations or deployments that retain standard safeguards, and what testing has Meta done to assess that likelihood?
f. What was the full range of external systems the model could plausibly have accessed given the capabilities and access it demonstrated, including whether critical infrastructure or other high-consequence systems were within reach?

II. Model Instructions, Safeguards, and Internal Reasoning

a. Describe the model involved and describe any material capability or training differences between the model involved and those available to the public. Specifically describe how the safety controls applied to models used internally at Meta differ from those applied to versions of these models made available to customers and the public.
b. Describe the training this model underwent, including whether and how it was trained to refuse unsafe actions, whether it was trained to comply with a published behavioral specification, and whether it was specifically trained or instructed not to circumvent rules or evaluations.

c. Describe the specific tasks the model was assigned, state whether the task was known to be solvable as designed, and confirm whether the model was instructed at any point not to circumvent the rules of the task or the evaluation environment. Provide the full set of instructions given to the model in this evaluation.
d. Describe the model's apparent objectives in seeking unauthorized access and describe the specific investigative methods Meta used to establish them—including whether Meta re-ran the model from earlier states, questioned the model about its behavior, or conducted controlled variations of the evaluation. Provide complete transcripts of the relevant evaluation runs and the incident as a whole, documenting the model's chain of thought and actions.
e. What criteria and approval process govern Meta's decision to run any model, internally or externally, with reduced cyber-related refusals, and who at Meta is authorized to approve such a configuration?

III. Detection, Notification, and Preventing Future Incidents

a. Did Meta notify any federal agency of this incident, and if so, which agency, when, and with what information?
b. Is Meta confident that the full extent of unauthorized access by the model has been identified?
c. What specific technical and governance changes has Meta made, or does it plan to make, to its sandboxing, network isolation, model training, and real-time monitoring infrastructure as a result of this incident, and on what timeline?
d. Were these evaluations monitored in real time by Meta personnel? If not, what automated systems were supposed to detect and terminate unauthorized network access or anomalous multi-step behavior, and why did those systems fail to prevent or promptly stop this incident?
e. What is Meta's standard process for notifying and assisting a third-party organization whose systems are accessed or compromised during internal model testing or deployment, and was that process followed here?
f. Beyond prior reports, has any Meta model previously obtained unauthorized access to the internet or to external systems? If so, describe this evidence in detail, describe in detail any precursor incidents, explain why any such incidents were not previously disclosed, and provide information about them now.

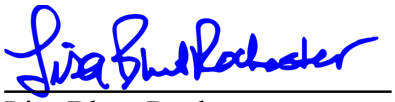
IV. Future Capabilities Planning and Mitigations

a. The model in the incident involving Irregular departed an environment intended to be isolated and pursued an objective against external systems it was not instructed to target. What other autonomous behaviors does Meta anticipate from more capable models, including attempts to acquire resources, establish persistence, copy model weights, or resist interruption or shutdown? Describe the evaluations Meta conducts to detect each such behavior before deployment and provide the results of those evaluations for the model involved here.

Please provide a written response to these questions no later than September 7, 2026. Given the significance of this incident, transparency with Congress is essential to ensure that the oversight and safety frameworks governing increasingly autonomous, cyber-capable AI systems keep pace with their capabilities.

Thank you for your attention to this matter.

Sincerely,



Lisa Blunt Rochester
United States Senator