Meta

September 10, 2026

Dear Senator Blunt Rochester,

Thank you for the opportunity to address the important issues raised in your letter.

As you note, earlier this summer, a misconfiguration by Irregular, a contractor Meta hired to conduct cybersecurity evaluations, inadvertently allowed one of our pre-released AI models, Muse Spark 1.1, access to the open internet during an evaluation that was supposed to be in a closed testing environment. We understand that Irregular tested other companies' AI models around the same time with similar misconfigurations in the testing environments. Meta learned of the issue when Irregular notified us, and, as discussed below, we completed a detailed retrospective to understand the issue and its implications. We published a blog post to provide more information on what we learned.

Importantly, our cybersecurity evaluations are designed to measure a model's capabilities by testing against specific threat scenarios. By design, some of these evaluations are conducted without certain safeguards that we deploy for released models available to consumers in order to measure the model's underlying capabilities. Testing AI models for cybersecurity capabilities is essential to understanding what they can do before deployment, allowing us to address any risks identified. As discussed below, this is one component of a broader evaluation process we conduct before releasing or deploying a model.

Meta has provided information responsive to your questions below. As this issue occurred as a result of a misconfiguration in Irregular's testing environment, not Meta's, our responses are based on the information that Irregular reported to us and our retrospective.

Muse Spark and Meta's AI Training and Safety Testing

In April 2026, Meta released Muse Spark. Muse Spark is a natively multimodal reasoning model designed to operate across text and image inputs and to support more advanced reasoning tasks. In July 2026, we launched the updated model, Muse Spark 1.1. Muse Spark 1.1 is a multimodal reasoning model built for agentic tasks, with improved tool and computer use, coding, and multimodal understanding.

As with previous models, Muse Spark 1.1 is subject to Meta's Terms applicable to generative AI. For example, our Terms of Service prohibit access or use of our AIs in any manner that

would violate any applicable law, facilitate content that may harm other individuals, or present a risk of death or bodily harm to individuals, among other things.

Muse Spark is the latest example of Meta's commitment to building generative AI features responsibly and expands upon our multi-layered approach to AI safety. For example, Muse Spark has protections built in at every stage—from filtering the data the model learns from, to safety-focused training, to guardrails that run on Meta's AI features and services.

We train and fine-tune our models to fit our safety and security guidelines, and our development process includes rigorous testing and red-teaming, with the goal of mitigating risks before release. We therefore conduct extensive safety evaluations before launch, testing against scenarios specifically designed to find weaknesses, tracking how often those attempts succeed, and working to drive performance on evaluations to ensure the model is designed to be safe to be used at scale. We also conducted extensive safety evaluations into catastrophic risks before deploying Muse Spark 1.1, in line with our Advanced AI Scaling Framework. Finally, because no evaluation is exhaustive, post launch, we monitor live traffic with automated systems designed to spot unexpected issues so we can address them quickly.

Testing pre-launch AI models for cybersecurity capabilities is an important part of this safety approach. Such testing is essential to understanding what our models can do before deployment and our published evaluation reports detail the stress-testing we conduct across a range of areas. Our cybersecurity evaluations are designed to measure a model's capabilities by testing against specific threat scenarios, combining public and internal benchmarks with independent third-party testing. As was the case here, some of these evaluations are conducted without certain safeguards described above, as this enables us to measure the model's underlying capabilities.

Third-Party Cyber Evaluation Issue

As discussed above, Meta regularly tests its pre-launch AI models, including combining public and internal benchmarks with independent, third-party testing.

In this case, we contracted with Irregular as the evaluating, third-party testing company. Irregular designs and constructs testing environments for evaluating not yet released AI models. As part of that work, we asked Irregular to test a pre-released version of Muse Spark 1.1, according to our normal testing protocol. In early July, Irregular began an exercise to evaluate, in a closed testing environment with safeguards removed, whether our model would be capable of completing an adversarial cybersecurity task. However, when Irregular set up the testing environment, a misconfiguration by them, that we understand was the result of human oversight allowed the model to access the open internet, and instead of using a fictional name of the "target" for the evaluation exercise, Irregular unintentionally provided the model with the name of a real website as its target. Believing the real website was the intended target, the pre-released version of Muse Spark 1.1, operating without its post-launch safeguards, identified and exploited

a security vulnerability in the real website. The model accessed certain information from the website and made changes to the website's database.

Meta learned about this issue after Irregular notified us. According to Irregular, as soon as it discovered the issue, Irregular turned off all internet access across evaluation environments. Upon learning about the issue, we immediately opened our own independent investigation. Our security teams reviewed over 10,000 records of Muse Spark 1.1's activity during testing to understand the full scope of what occurred. To date, we have learned:

- The model operated within the scope of its assigned task based on the instructions it was given by Irregular and the environment it encountered. This was not a sophisticated offensive cyber attack or sandbox escape.
- No other instances of the model exploiting a third-party company's system were identified beyond this evaluation, demonstrating the isolated nature of this issue.
- Our review also surfaced additional areas for improvement in the testing environment and the integration with it, which we shared with Irregular to support remediation efforts.
- We have also identified monitoring improvements that would help identify these types of issues earlier in the future.

Irregular has confirmed the misconfiguration has been corrected and that measures have been put in place to ensure that evaluations do not unintentionally reference real website names in the future.

Looking Ahead

As AI models become more capable, cybersecurity evaluations surface a specific challenge: models that demonstrate the ability to find and exploit vulnerabilities require proportionally stronger containment during testing. This is an industry-wide challenge that labs are actively working to address, including through improved containment during testing, better separation between evaluation and production environments, and ongoing research into how to evaluate safely without reinforcing unwanted behaviors in the model. Accordingly, the industry understands the importance of appropriate guardrails and coordination around the use of the internet in AI testing environments. To that end, going forward, we are implementing independent verification requirements for test environment isolation and scenario review before evaluations begin and that test scenarios do not reference real companies. More broadly, we continue to invest in our evaluation and preparedness infrastructure. As model capabilities advance, evaluation environments and practices need to keep pace, both within our own infrastructure and across the third-party partners we work with.

American AI models will support the prosperity and security of the United States and strengthen the U.S.'s position in the global race for AI leadership. That is why we are contributing to industry-wide efforts to help improve security and evaluation safety practices, model alignment,

and how models behave in adversarial testing scenarios. We plan to continue to strengthen how evaluations are conducted in a secure manner going forward, together with our partners.

Sincerely,

Brian Rice
VP, Public Policy