United States Senate
WASHINGTON, DC 20510

June 6, 2023

Mark Zuckerberg
Chief Executive Officer
1 Hacker Way
Menlo Park, California 94025

Dear Mr. Zuckerberg,

We write with concern over the “leak” of Meta’s AI model, the Large Language Model Meta AI (LLaMA), and the potential for its misuse in spam, fraud, malware, privacy violations, harassment, and other wrongdoing and harms. As a part of the Subcommittee on Privacy, Technology, & the Law’s work on artificial intelligence (AI), we are writing to request information on how your company assessed the risk of releasing LLaMA, what steps were taken to prevent the abuse of the model, and how you are updating your policies and practices based on its unrestrained availability.

In February 2023, Meta released LLaMA, an advanced large language model (LLM) capable of generating compelling text results, similar to products released by Google, Microsoft and OpenAI.[1] Unlike others, Meta released LLaMA for download by approved researchers, rather than centralizing and restricting access to the underlying data, software, and model. Meta explained this decision as helping efforts to advance AI research in order to “improve their robustness and mitigate known issues, such as bias, toxicity, and the potential for generating misinformation.” While LLaMA was reportedly trained on public data, it differed from past models available to the public based on its size and sophistication. Regrettably, but predictably, within days of the announcement, the full model appeared on BitTorrent, making it available to anyone, anywhere in the world, without monitoring or oversight. The open dissemination of LLaMA represents a significant increase in the sophistication of the AI models available to the general public, and raises serious questions about the potential for misuse or abuse.

Open source software and open data can be an extraordinary resource for furthering science, fostering technical standards, and facilitating transparency. Many experts compellingly argue that open access to these models can help the development of safeguards through exposing

[1] Introducing LLaMA: A foundational, 65-billion-parameter large language model, Meta. https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

vulnerabilities to a larger community who can find fixes. As Meta noted when releasing LLaMA, there is a need for "more research that needs to be done to address the risks of bias, toxic comments, and hallucinations in large language models." Providing the full model allows the types of research, testing, and collaborative development that are not as readily achievable within closed systems like OpenAI's GPT-4 or Google Bard. Additionally, as AI becomes more important to technological growth and competition between consumer platforms, the centralization of AI expertise and technical capabilities risks stifling innovation and market competition, and threatens to further entrench incumbent tech firms. Open source AI can play a meaningful role in making sure that AI systems are robust and safe, and that the field of AI is not dominated by a few select corporations.

On the other hand, even in the short time that generative AI tools have been available to the public, they have been dangerously abused — a risk that is further exacerbated with open source models. For example, after Stability AI launched its open source art generator, Stable Diffusion, it was used to create violent and sexual images, including pornographic deep fakes of real people, which disproportionately feature women 96% of the time.[2] Even OpenAI's closed model, ChatGPT, has been misused to create malware and phishing campaigns, financial fraud, and obscene content involving children.[3] At least at this stage of technology's development, centralized AI models can be more effectively updated and controlled to prevent and respond to abuse compared to open source AI models.

Adding to this risk, Meta appears to have done little to restrict the model from responding to dangerous or criminal tasks. For example, when asked to "write a note pretending to be someone's son asking for money to get out of a difficult situation," OpenAI's ChatGPT will deny the request based on its ethical guidelines. In contrast, LLaMA will produce the letter requested, as well as other answers involving self-harm, crime, and antisemitism. While the full scope of possible abuse of LLaMA remains to be seen, already the model has been utilized to generate profiles and automate conversations on Tinder[4] and a chatbot built from LLaMA, Stanford's Alpaca AI, was taken down shortly after release over providing incorrect information

[2] Anyone can use this AI art generator — that's the risk. The Verge.
https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data
"15k deepfake videos on the internet—and 96% of them are porn, Fast Company.
https://www.fastcompany.com/90414116/there-are-almost-15k-deepfake-videos-out-there-and-96-of-them-are-porn

[3] ChatGPT: The impact of Large Language Models on Law Enforcement, EUROPOL.
https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf
ChatGPT Generated Child Sex Abuse When Asked to Write BDSM Scenarios, Vice Motherboard.
https://www.vice.com/en/article/v7b4m9/chatgpt-generated-child-sex-abuse-when-asked-to-write-bdsm-scenarios

[4] People Trying to Use Facebook's Leaked AI to Improve Their Tinder Matches, Vice Motherboard.
https://www.vice.com/en/article/qjv4vw/people-use-facebooks-leaked-llama-ai-for-tinder

and troubling responses.[5] It is easy to imagine LLaMA being adopted by spammers and those engaged in cybercrime. While centralized models can adapt to abuse and vulnerabilities, open source AI models like LLaMA, once released to the public, will always be available to bad actors who are always willing to engage in high-risk tasks, including fraud, obscene material involving children, privacy intrusions, and other crime.

Meta's choice to distribute LLaMA in such an unrestrained and permissive manner raises important and complicated questions about when and how it is appropriate to openly release sophisticated AI models. Given the seemingly minimal protections built into LLaMA's release, Meta should have known that LLaMA would be broadly disseminated, and must have anticipated the potential for abuse. While Meta has described the release as a leak, its chief AI scientist has stated that open models are key to its commercial success. Unfortunately, Meta appears to have failed to conduct any meaningful risk assessment in advance of release, despite the realistic potential for broad distribution, even if unauthorized. Stunningly, in the model card and release paper for LLaMA, Meta appears not to have even considered the ethical implication of its public release. It provides sparse details about its testing or steps to prevent for abuse, aside from technical measurements of bias. The dearth of information is particularly stark in comparison to the more extensive documentation released by OpenAI in connection with its closed models, ChatGPT and GPT-4.[6] By purporting to release LLaMA for the purpose of researching the abuse of AI, Meta effectively appears to have put a powerful tool in the hands of bad actors to actually engage in such abuse without much discernable forethought, preparation, or safeguards.

As Congress considers legislation and oversight to promote both innovation and accountability in the commercialization and use of AI, it is important to understand how companies are assessing and mitigating the risks associated with AI models. Meta has held itself out a leader on AI, and is under a Federal Trade Commission consent decree and other scrutiny over the safety of its platform. While Meta's stated intention of promoting safety research may have merit, the lack of thorough, public consideration of the ramifications of its foreseeable widespread dissemination is a disservice to the public. It is important to better understand how Meta evaluated this risk and what changes it plans to make after this matter.

Given the seriousness and scope of LLaMA's public availability, we respectfully write to ask you to answer the following questions by June 15, 2023:

1. What risk assessments, if any, were conducted regarding the likelihood and repercussions of dissemination of LLaMA to anyone other than its authorized recipients prior to Meta's release of the model to researchers?

[5] Stanford Researchers Take Down Alpaca AI Due to 'Hallucinations' and Rising Costs, Gizmodo. https://gizmodo.com/stanford-ai-alpaca-llama-facebook-taken-down-chatgpt-1850247570
[6] LLaMA Model Card. https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md
LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971

a. How many researchers did Meta approve to have access to LLaMA's model weights? What criteria was used in the selection and vetting of the approved researchers?

b. What specific technical and administrative safeguards did Meta implement in connection with its release of LLaMA to prevent the public dissemination of the model?

c. How did Meta balance this risk against its stated goal of helping foster AI safety research?

d. Who was consulted inside and outside of Meta about this risk, and how was the decision made to release Meta to researchers?

e. What other safety or security measures were contemplated to protect against the public dissemination of LLaMA?

2. What steps has Meta taken since the release of LLaMA to prevent or mitigate damage caused by the dissemination of its AI model?

   a. What steps has Meta taken to track the distribution, repurposing, and end use of LLaMA, including its potential misuse in fraud and spam campaigns targeting users on its platforms? Please provide any documentation of these efforts and findings, including any DMCA takedown notices, cease and desist letters, or similar efforts.

3. What steps were taken during and after the training process for LLaMA to ensure that the model could not be abused after it was released, for example, steps taken to prevent responses that could be used in fraud schemes, self-harm, and cybercrime?

4. Has Meta made changes to its policies, practices, and plans concerning around the sharing of AI models and other data based on the dissemination of LLaMA? If so, what changes were made and what lessons has Meta learned from it?

5. Has Meta considered alternative approaches toward working with researchers, enhancing the security of its models, or preventing them from falling in the hands of malicious actors, such as providing access through an API or allowing the model to be run in a more limited sandbox provided by Meta? Under what conditions does Meta consider it necessary to provide full access to AI models, despite the risk of release?

6. Has Meta developed policies or guidelines for when it believes that AI models should not be available to the public, for example limiting access when a model is highly capable of performing particular tasks or based on the model's parameter size?

7. Overall, how does Meta evaluate the risks and precautions that should be taken prior to the public release of a sophisticated AI model? Can this documentation be shared with the public in a form similar to the system card OpenAI shared for GPT-4?

8. The LLaMA research paper asserted that no proprietary or inaccessible data was used to train the model. Was LLaMA trained in any way on any data that it obtained or that derived from any of Meta's customers, such as posts, content or any other data created or provided by users of Facebook, Instagram or WhatsApp?

9. Meta has significant access to its users' personal information and uses that data for AI systems, such as those used in advertising. When does Meta use its users' personal data for AI research, including AI models that are available to the public or outside researchers?

Thank you for your attention to these important issues. We look forward to your response.

Sincerely,

Richard Blumenthal
Chair
Subcommittee on Privacy, Technology, &
the Law
United States Senate

Josh Hawley
Ranking Member
Subcommittee on Privacy, Technology, &
the Law
United States Senate