Meta

## Executive Summary

Open source is the foundation of U.S. innovation and global technological leadership. Continued leadership of this technological revolution – including through support for responsible open source AI domestically and in international fora – will underpin U.S. economic, domestic, foreign policy, international development, and national security interests. Meta has been at the forefront of responsible open source in AI for over a decade and believes that an open approach to AI leads to better, safer products, faster innovation, and a larger market. We appreciate the opportunity to share our experience with building AI with the goal of making the benefits accessible to everyone.

In our submission, we suggest that NTIA's report:

1. **underscores the importance of open foundation models to U.S. economic, national security, and foreign policy interests;**

2. **recommends that the Government continue its work through NTIA, NIST, the White House Voluntary AI Commitments, and through other agencies to establish common standards for risk assessments, benchmarks and evaluations informed by science, noting that the U.S. national interest is served by the broad availability of U.S.-developed open foundation models;**

3. **notes the importance of continued American leadership in driving international consensus, cooperation, and interoperability on AI governance, including among AI Safety Institutes, as well as at the United Nations, G7 and OECD; and**

4. **points to the need for bipartisan federal AI legislation informed by the work of NTIA, NIST and international efforts on common standards in order to avoid a fragmented regulatory environment across the U.S.**

Meta

Table of Contents


Introduction 2

I. How should NTIA define “open” or “widely available weights” when thinking about foundation models and model weights? 5

II. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights? 14

III. What are the benefits of foundation models with model weights that are widely available as compared to fully closed models? 21

IV. Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights?
If so, please list them and explain their impact. 29

V. What are the safety related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available weights? 29

VI. What are the legal or business issues related to open foundation models? 38

VII. What are the current or potential voluntary, domestic regulatory, and international mechanisms to manage risks and maximize the benefits of foundation models with widely available weights? What kinds of entities should take a leadership role across which features of governance? 40

VIII. In the face of continually changing technology, and unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future? 44

IX. What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights? 45

X. Recommendations 47

Meta

## Introduction

Meta has been a leader in open innovation in AI for over a decade and believes that an open approach to AI leads to better, safer products, faster innovation, and a larger market. We appreciate the opportunity to share our experience building AI with the goal of making the benefits accessible to everyone.

During earlier eras of the Internet, open source technologies played a core role in promoting innovation and safety. Over time, open source has become the foundation of U.S. innovation and global technological leadership.

*Open Source Increases U.S. Competitiveness*
Open source democratizes access to the benefits of AI. These benefits are potentially profound for the U.S., and for societies around the world. We have seen a groundswell of enthusiasm for open source AI from small businesses across the U.S. They recognize the transformative potential of AI for competition, innovation, and productivity, which is already contributing to U.S. economic growth. Having access to state-of-the-art AI creates opportunities for everyone, not just a small handful of Big Tech companies. Open source provides the foundations on which developers can innovate that would otherwise be prohibitively costly.

We have heard from the U.N. and governments across the Global South that open source AI could finally light the path to achieve the Sustainable Development Goals. We've seen this reflected in the recently adopted U.S.-led U.N. General Assembly Resolution on Artificial Intelligence.

The U.N. and governments see the value of open source AI as a force multiplier to help deliver services to their populations, close inequality gaps, and fuel new technologies and enterprises. We are already witnessing the potential of open source AI to generate fundamental breakthroughs in health and science for everyone.

And we have heard from departments and agencies across the U.S. government that are building tools on Meta's open foundation model, Llama 2, to enhance efficiency and improve service delivery.

Leading this technological revolution – including via support for responsible open source AI in international fora informed by the work of NTIA, NIST and other agencies – will benefit U.S. economic, domestic, foreign policy, international development, and national security interests.

By making U.S. technology the standard on which developers around the world build, the U.S. has an opportunity to embed our innovation and values in the fabric of the AI revolution, ushering in a new era of American technological leadership that will benefit the U.S. for decades to come.

*Responsible Approach to Open Source*
We acknowledge that there is a risk/reward tradeoff when it comes to any approach with regard to releasing AI technology. While open source is in Meta's corporate DNA, we are not dogmatic in our approach. Meta open sources responsibly, which means releasing AI that is designed around: 1) privacy and security; 2) fairness and inclusion; 3) robustness and safety; 4) transparency and control; and 5) accountability and governance. We've also open sourced safeguards (*e.g.,* Llama Guard) that, when combined with our models, increase safety at the system level. These are now being standardized by MLCommons and the open source community for safe generative AI application development.

Our commitment to responsible open source means we prohibit a range of harmful uses via our commercial license, and we have at times elected not to release tools because we deem the risks associated with them too high. We strongly support the work of NIST and other bodies working to define taxonomies of risks, benchmarks, and evaluations, and building processes for external validation, such as red-teaming. We urge the NTIA to work with NIST on standardizing the threat models and evaluations along the AI value chain, so that risk can be measured quantitatively and consistently and standard thresholds can be set against those evaluations. The outcomes of this work will provide crucial guidance to companies like Meta on how to measure and mitigate risks, and how and whether we decide to release certain models.

It is also important to consider the limitations associated with closed models. Open source models leverage transparency to crowd-source efforts to build safer and more robust systems, resulting in continuous vulnerability identification through the scientific method of interrogation, as well as mitigation experimentation (*e.g.,* at the model level, at the system level, or at the user education level).

Closed models must rely on in-house experts to perform these functions; without the continuous scrutiny involved in open sourcing, vulnerabilities may remain undetected and exploitable for significant periods of time. This is why open source has become the gold standard in cybersecurity and encryption, among other areas.

Meta

Closed models also create an illusion of security. As we have seen recently, closed models may be subject to unintentional release, either through lapses in security, leaks, or model extraction attacks. If a closed model is exfiltrated, only the model developer has the information required to mitigate harms, and that understanding may be limited. Given historical precedents of closed models leaked or stolen, and the current general availability of equally capable open source models, it is important not to rely on secrecy as the only means of safety and security.

As such, the focus should be on ensuring the responsible deployment of all models, rather than focusing on whether the model weights are released. In short, we suggest that further work is required to establish benchmarks for assessing the quality of existing safeguards – for both closed and open approaches – to better understand potential risks with each. NIST's work will be important in this regard.

*The Spectrum of Openness & the Need for U.S. Leadership*
All of this points to an ecosystem that includes a spectrum of openness, where developers – guided and supported by standard-setting and other governance initiatives – can tailor how they release models based on a variety of factors. The United States has a crucial role to play in driving international consensus on AI governance, U.S. leadership and innovation in the age of AI will promote U.S. interests.

Multi-stakeholder efforts, like those underway through the U.S. AI Safety Institute and in international fora like the G7 and OECD, are central to the ability for governments, companies, and individuals to formulate durable, balanced solutions to the challenges and opportunities of today and the future.

If the United States restricted U.S. companies' ability to open source foundation models, it would undermine U.S. interests. Specifically, open source drives innovation and sets the standard on which others will build. Other countries are already fielding strong competitors in this race (owing, in part, to a desire to maintain independence from the U.S.); by restricting the distribution of U.S.-grown AI, the U.S. would cede its pole position in AI leadership and innovation, depriving the U.S. economy of related growth opportunities.

This would leave a vacuum that other countries would be eager to fill and benefit from, with no guarantee that the models provided would reflect our values or vision for the future.

Historically, broad and restrictive limits on open innovation of new technologies — such as export controls on certain types of encryption in early web browsers — have been counterproductive. Many of these restrictions were ultimately rolled back as the U.S. government shifted to preferring open source tools in these areas (*e.g.,* open source encryption protocols that were actually more secure). We can and should apply these lessons to AI and we encourage the U.S. government to carefully consider and minimize restrictive limitations on open source in this space.

**1. How should NTIA define "open" or "widely available weights" when thinking about foundation models and model weights?**

Concepts like "open" or "widely available weights" are not binary but rather exist on a spectrum. Rather than striving for strict definitions, we suggest teasing out the nuances of an "open" approach as it applies to AI, which includes understanding the foundation model technology stack.

The Open Source Spectrum

There is wide discourse in industry, academia, and civil society on what "open" means in the context of foundation models. While "open source" has a specific meaning as applied to traditional software, this meaning is less clear when applied to AI systems. Hugging Face, one of the most widely used platforms for sharing and collaborating on machine learning systems, has described the current state of play as follows:

> "For AI systems, conversations about how to operationalize the values that underpin open source software are still very much ongoing; especially given the importance of data in fully understanding how they are designed and how they work. As such, we tend to use terminology like "open science" and "open access." For AI models, we create processes in light of the trade-offs inherent in sharing different kinds of new technology, and approach our work in terms of a "gradient of openness" (also called a "release gradient") to foster responsible development."[1]

[1] Hugging Face Comment on FR Doc # 2023-28232.

When considering this issue, NTIA should take account of recent discussions at the Open Source Initiative itself [2] and the ongoing work of the Columbia University Convening on Openness and AI, a forum that is bringing together over 40 leading experts working on openness and AI.[3]

The Foundation Model Technology Stack

When considering what "open" means in the context of foundation models, it is important to consider the entire tech stack underpinning these models. The tech stack is the software, hardware, and other elements required to develop the model. Among other things, this includes programming languages (*e.g.,* Python), machine learning frameworks (*e.g.,* PyTorch; these frameworks execute the mathematical calculations that generate the model weights); training datasets (*e.g.,* CommonCrawl); the model itself (*e.g.,* transformer models like Meta's Llama 2); and hardware (*e.g.,* GPU's such as Nvidia H100's, and the associated architecture). It is also important to consider who can access each component and for what purposes.

An open approach to the foundation model tech stack, for example, could include providing access to model weights but not providing access to other assets, like training datasets. Although one can build on ('fine-tune') a foundation model without model weights (*e.g.,* through an API) to adapt it for a variety of purposes, there is significantly more flexibility, lower costs, and more control over one's data associated with fine-tuning openly released model weights. Other assets, like training datasets, are not required to fine-tune a model and, given that sharing them could come with risks, should be shared less routinely. (See Question 2 (c)).

Considering all the components of the AI tech stack demonstrates that there are different types of openness that can be helpful for accomplishing different technical and societal goals.[4] It is important that model providers retain the ability to make decisions regarding the degree of openness for each asset, based on evaluations of the risks of each asset separately.

[2] In an interview with Emilia David for *The Verge* (Oct 30, 2023), Stefano Maffulli, the executive director of the OSI, noted that "the group understands that current OSI-approved licenses may fall short of the certain needs of AI models and that the OSI is reviewing how to work with AI developers to provide transparent, permissionless, yet safe access to models."

[3] *See* Ayah Bdeir & Camille Francois, Introducing the Columbia Convening on Openness and AI, dest://ed (March 6, 2024).

[4] *Id.*

Meta

Meta's Approach to Open Source

Meta has been deeply involved in the AI research community for over a decade. This has allowed us to develop and evolve a set of best practices, principles, and processes to support our release of foundation models informed by a deeper level of intentionality anchored on assessing the differential (or 'marginal') risk (see response to Question 2). We have also worked through the Partnership on AI to develop "The Guidance for Safe Foundation Model Deployment," which is one of the most comprehensive, nuanced, and inclusive frameworks for responsibly building and deploying AI models through an open approach.

As noted by Joelle Pineau, Vice President AI Research at Meta and Vice-Chair of the PAI Board, "the Partnership on AI's leadership has been invaluable in bringing together industry, civil society, and experts as companies like ours determine the best approach when looking at both open and closed releases."[5]

Feedback from NTIA on this framework will be critical to advancing its potential for integration into NTIA and NIST guidance. This would inform a standard approach to responsible scaling policies that would drive practices across the ecosystem.

As Meta develops our AI technologies, we are advancing both the state of the art in terms of technical capabilities *and* the responsibilities that come with those developments (for example, continuously developing new tooling for assessing safety, datasets for assessing fairness and bias, and new methods to probe security). We believe that, in general, being more open with our AI research accelerates the innovation cycle of the underpinning technology *and* risk assessments, benchmarks, evaluations, and mitigations.

Overall, relative to closed source, responsible open source establishes a higher bar before the technology is widely embedded in consumer-facing products and used by people around the world.

However, Meta's responsible open source strategy is not absolute, and each decision is taken on a case-by-case basis, largely because static processes are quickly outdated by the pace of developments.

[5] *Partnership on AI Releases Guidance for Safe Foundation Model Deployment, Takes the Lead to Drive Positive Outcomes and Help Inform AI Governance Ahead of AI Safety Summit in UK*, PAI blog (Oct. 24, 2023).

This is why the work of NIST under the E.O., and of the U.S. AI Safety Institute, in defining taxonomies of risks, benchmarks, and building processes for external validation (like red-teaming) is so important. As we consider various factors when assessing whether to open source a model, we are rigorous in our approach, and are continuously mindful of evaluating risks and mitigations in a way that is mindful of the public interest. And this ultimately facilitates the development of better technology.

Given that there is no broad consensus on a settled definition, at Meta we consider "open" in terms of making models and their weights publicly available *responsibly* for research as well as commercial purposes. This means when we make pre-trained model weights widely available for commercial and research use, we do so along with model cards and other information such as user guides and an Acceptable Use Policy, the use of which is subject to the acceptance of the model license terms. When we launched the Llama 2 model, over 100 experts, academics, and policymakers supported this responsible approach.[6]

Balancing Factors

Not necessarily every model should be open sourced, and a decision to open source should be made after weighing a range of factors, including risk assessments and business needs. For some highly advanced and novel models, for example, it may be appropriate to first release the model (or certain artifacts) to researchers to allow time to understand and, if necessary, mitigate risks associated with the model.

At Meta, we are working to address the hard questions around issues such as privacy, safety, fairness, accountability, and transparency through our privacy review process, with privacy- and AI-specific risks identified, mitigated, evidenced and monitored.[7] Furthermore, for foundation models that we pre-train and fine-tune for use across some of Meta's platforms (*e.g.,* MetaAI in WhatsApp, Messenger, and Instagram[8]), we instituted additional layers of review and escalation to assess risks at every stage of development (including data collection, model training, and fine-tuning) in addition to evaluations (*e.g.,* red-teaming and scaled).

---

[6] *See* statement of Support for Meta's Open Approach to Today's AI ("responsible and open innovation gives us all a stake in the AI development process, bringing visibility, scrutiny and trust to these technologies. Opening today's Llama models will let everyone benefit from this technology").
[7] Mike Clark, *Privacy Matters: Meta's Generative AI Features*, Privacy Matters blog (Sept. 27, 2023)
[8] *Introducing New AI Experiences Across Our Family of Apps and Devices*, Meta Newsroom (Sept. 27, 2023).

In the case of generative AI products, we further instituted additional measurement of prompt/output-level filtering and classification effectiveness across an extensive range of safety dimensions.

When it comes to models Meta has released in non-production environments (*e.g.,* for testing, development, or research purposes and not intended for live usage by end-users), they undergo an Institutional Review Board (IRB)-like process in addition to a privacy review, where researchers work to mitigate safety issues at the model level. Where this is not yet possible, we have limited the release of the model weights.

Some of our recent limited releases include our AudioBox model, for which we only released the demo and research paper publicly, limiting model access to a small group of researchers under a special program to study its applications, including safety mitigations. (This is because tools like Audiobox can raise concerns about voice impersonation or other abuses, and more work is needed in collaboration with researchers and academics to conduct safety and responsibility research.)

Similarly, Meta did not open source our model that decodes images using brain activity.

Another factor for model providers to consider is that it might not make business sense to make models publicly available at first, or even at all, given the cost and resources incurred by the model provider in training a model. Indeed, there are estimates that training the next generation of large language models will cost more than $1 billion within a few years.[9]

In general, while we believe that open sourcing models responsibly enables scrutiny and advances innovation in ways that closed approaches broadly cannot as efficiently or effectively, we recognize that model providers may wish to adopt different approaches at different times and/or for different models based on a range of factors, including business model.

[9] Craig S. Smith, *What Large Models Cost You – There Is No Free AI Lunch*, Forbes (Sept. 8, 2023)

Meta

Risk Assessment

Responsibly open sourcing also means acknowledging and taking seriously that new risks may arise. With respect to Llama 2, we worked to build safety measures and identify research areas to account for such risks. We examined a number of potential risk areas that were relevant to our use cases, and then mapped and deployed appropriate mitigations. (See Question 5(a) on evaluations).

Prohibited Uses

We responsibly open source models using a "permissive commercial license," which allows for a wide range of downstream uses – but is not unrestricted. For example, our license and Acceptable Use Policy, which must be agreed to before the model is made available by Meta, includes a range of prohibited uses and conditions of use. Such uses include violating the law or other's rights; engaging in the planning or development of activities that present a risk of death or bodily harm to individuals; and intentionally deceiving or misleading others.[10]

Community Feedback

Responsible open sourcing also includes iterating on the findings of the research community through adopting a progressive approach to releases (for example, at Meta we released our Llama and Dino models to researchers in the first instance, and iterated on the feedback before progressing to our permissive commercial license). A core part of responsible open sourcing also means iterating based on broader community feedback in a timely fashion, as well as transparency about known issues and limitations of fixes. For example, we provide a number of means to report violations of our Llama 2 Acceptable Use Policy, software bugs, or other problems that could lead to a violation of the policy.[11]

[10] *See* Llama 2 Acceptable Use Policy.
[11] More information on Meta's Bug Bounty Program here and here.

**1. a. Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?**

Given the progression of the technology around the world, the performance and capabilities of closed source and open source models (with widely available weights) are converging. A contemporary example is Meta's Llama 2 model, the weights for which are publicly available subject to the license terms for use of the model. Llama 2 outperforms GPT-3 (a model for which the weights are not publicly available) on many benchmarks.[12]

As open foundation models further evolve around the world (*e.g.,* Falcon, Vicuna, Mistral) they will continue to match, and exceed, the capabilities of closed foundation models, and there is no empirical evidence or reason to believe otherwise.

Further, there is always a risk that closed models' model weights can end up in the public domain. For example, even if closed models implement state-of-the-art security around model weights, there is always the possibility that these might be subject to unintentional release – either through lapses in security, leaks, or model extraction attacks.[13] Indeed, there are recent examples from across the AI ecosystem of model leaks, as well as insider attacks at AI labs.[14]

In contrast to traditional software security vulnerabilities, where defenders assume bad actors are already exploiting them and therefore transparently and broadly share remediation information, if a closed foundation model is exfiltrated, only the model developer has the information required to mitigate harms. This can lead to adverse outcomes because nobody apart from the developer will have had the opportunity to interrogate the model previously, to identify vulnerabilities and fix them (the cybersecurity threats Heartbleed, Shellshock, Spectre, and Meltdown were all identified and addressed by the open source community).

[12] Sunil Ramlochan, *How Does Llama-2 Compare to GPT-4/3.5 and Other AI Language Models*, Prompt Engineering & AI Inst. (Sept. 1, 2023).
[13] *See* Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, Jinwen He, Towards Privacy and Security of Deep Learning Systems: A Survey (Oct. 27, 2020) ("any kind of machine learning can be stolen").
[14] Chinese National Residing in California Arrested for Theft of Artificial Intelligence-Related Trade Secrets from Google, DOJ Press Release (March 6, 2024).

Because the model weights for closed models can always end up being in the public domain, and because the weights of models similar to currently closed AI systems are/will be widely available in any event, it does not logically follow that closed sourced models are, or will always be, inherently safer. To the contrary, because open development allows vulnerability detection and mitigation by the community at large, open releases can have safety advantages.

As such, the focus should be on ensuring the responsible deployment of all models rather than focusing on whether the model weights are released.

**1.b. Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?**

It is not possible to generally estimate this timeframe given the variables involved, including the model deployment developers' business models and whether, in the case of Llama 2, they download the model weights from Meta directly or accessed it through third-party services like Azure or AWS.

Other variables include the application complexity (a highly complex agent is significantly different to just deploying a text summarizer), the sophistication of the developer (*e.g.,* an average developer without much ML experience in comparison to highly specialized and experienced ML teams), and the level of abstraction targeted (for example if a developer uses the models directly from a "model zoo" (*e.g.,* Vertex Model Garden), this requires significantly more resources to manage the hosting work, operations of the Large Language Model (LLM), etc., compared to those from a model API).

Model providers work hard to make their user interfaces, APIs, and model access points as seamless as possible, and providers like Meta make "recipes" available (*e.g.,* including the choice of model architecture, the training algorithm, the hyperparameters, and other factors that influence the training process). The downstream developer and researcher community are similarly developing recipes to make model deployment quick and easy.

**1.c. Should "wide availability" of model weights be defined by level of distribution? If so, at what level of distribution (e.g., 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be "widely available"? If not, how should NTIA define "wide availability?"**

"Widely available" should be aligned with the Partnership on AI's Guidance for Safe Foundation Model Deployment, which provides model providers with recommendations based on foundation model type (*e.g.,* Specialized Narrow Purpose, Advanced Narrow and General Purpose, Paradigm-shifting or Frontier) and release type (*e.g.,* Open Access, Restricted API/Hosted Access, Closed Development and Research Release).

Based on this guidance, "widely available" might usefully be defined as a subset of those criteria (*i.e.*, paradigm shifting/frontier plus open access), which would carry additional recommendations, also available in the guide, on how to conduct such a release safely.

**1.d. Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API's), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?**

**1.d.i Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?**

Model providers assessing how to make foundation models available consider a range of factors, including business model, performance, control over the model, and the target audience (higher levels of abstraction favor a broader range of developers, while lower levels of abstraction (with more ways to customize) favor deeper ML experts). Which approach to take is therefore fact-specific and variable.

Locally Hosting

Locally hosting an open foundation model can provide a level of security not available through an API. With local hosting, there is no transfer of data back to the model provider or downstream developers' servers, which can be an important factor for those building on the model who want to ensure confidential, sensitive, or classified information remains within their control. For example, Meditron and other leading medical models rely on open architecture to ensure that sensitive and private clinical data is run and stored locally.

APIs
APIs can provide multiple ways to check outputs and vet user behaviors to ensure models are not abused, but responsible open source enables wider scrutiny of the model, often enabling deeper research, innovation and safety. (See Question 3.)

Web Applications
Accessing foundation models through web applications can be a cost-effective, scalable, and convenient method, but trade-offs include dependency on the web application provider, performance limitations, and security risks (*e.g.,* if the web application itself is compromised).

Edge Deployment
While this method includes benefits such as reduced latency, privacy, and potentially reduced costs, as with other methods there are trade-offs, including devices' computing power (phones, tablets, laptops are more limited), which can impact performance and accuracy, as well as challenges to maintaining and updating the on-device model.

**2. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?**

In general, it can be more challenging to fix or mitigate safety/security issues with open models as the model provider/developer can't issue patches (unlike for traditional software). However, it is important to remember that if bad actors find an issue in a closed model via API, they are unlikely to disclose or report it;the onus is entirely on the model API owner to detect and mitigate such issues.

Therefore, broadly speaking, responsible open source reduces the potential risks of foundation models. Providing access to a wide community – including regulators, AI experts, hobbyists, and innovators – subjects the model to scrutiny and interrogation at scale, which can lead to the rapid identification of poor outcomes or vulnerabilities–and the development of fixes.

For example, since the release of Llama 2, we have tracked dozens of model improvement, features asks, and bugs reported by the developer community, resulting in changes to the next generation of models and improved tooling such as Purple Llama. Furthermore, we triage and resolve regularly issues reported on Github, resulting in overall model improvement and which quickly advances the state of the art.

Meta

Marginal Risk Analysis

In order to precisely identify and assess risks uniquely presented by open foundation models, it is important to apply a "marginal risk analysis"[15] that takes account of the risks of open models compared to: (1) preexisting technologies, and (2) closed models.

Open Source vs. Preexisting Technology

An example where the marginal risk to preexisting technologies is *not* increased compared involves phishing attacks. These have existed for a long time, and while large language models could be used to generate sophisticated phishing emails at scale, the models themselves do not necessarily increase this cybersecurity risk. That risk arises only when people receive the email and take steps in response to it (*e.g.,* installing malicious code or divulging sensitive information).

And large language models are in fact a powerful means of detecting and combating this type of harmful content.[16] Because the most effective solutions are not on the content-generation side (malicious actors do not need LLMs to generate phishing content), risk countermeasures should forgo LLM access restrictions and instead recognize that they are an important tool, along with other mitigations (*e.g.,* implementing unclonable authentication credentials and multifactor authentication that would apply no matter how the attacker generates phishing emails).

Open Source vs. Closed Models

With respect to the marginal risk of open models compared to closed models, it is important to note that a closed approach is not necessarily safer. For example, model weights can be exfiltrated (see Question 2, above), API safeguards are fallible, accidental releases can happen, and insider threats are real.[17] Further, there are additional benefits to an open source approach, including security benefits that, in general, make open foundation models a safer option. (See response to Question 3.)

[15] Sayash Kapoor & Rishi Bommasani, *et. al.*, *On the Societal Impact of Open Foundation Models*, Stanford Center for Research on Foundation Models (CRFM) (Feb. 27, 2024).
[16] *See* Fredrik Heiding & Bruce Schneier, *et. al.*, *Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models* ("large language models are adept at detecting phishing emails and can provide good recommendations to the recipient").
[17] Andy Zou, *et al.*, Universal and Transferable Adversarial Attacks on Aligned Language Models (Dec. 20, 2023).

In short, we suggest that further work is required to establish benchmarks for assessing the quality of existing safeguards for *both* closed and open approaches to better understand potential risks with each. The efforts of NIST and the U.S. AI Safety Institute will be important in this regard.

**2.a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?**

On the question of assessing the risks and benefits associated with releasing model weights, we would again draw NTIA's attention to the Partnership on AI Guidance for Safe Foundation Model Deployment.

Releasing model code alongside model weights comes with both risks and benefits. The evaluation of the risk depends on the threat model. With the model code, a sophisticated and resourced developer may remove safety guardrails that were built into the model at the base layer.

For example, guardrails that prevent models from generating offensive or harmful content could be removed, or modifications could be made to model weights in a way that makes them more susceptible to bias or discrimination.[18] However, pushing out code fixes with open models is significantly faster and more efficient than data-level or model-weights-level mitigations because code fixes can be applied universally, but data-level or model-weights-level mitigations often need to be specifically tailored, or require retraining the model—so code is a crucial part of open releases.

Releasing model weights alone, without additional information about how to use a model (such as model cards and source code), is not particularly useful to a regular (non-nefarious) model deployment developer. It would therefore make little sense for a company to release weights without additional instructions, recipes, and code.

[18] *See, e.g.*, Chloe Xiang, *The Amateurs Jailbreaking GPT Say They're Preventing a Closed-Source AI Dystopia*, Vice (March 22, 2023), Will Knight, *A New Trick Uses AI to Jailbreak AI-Models-including GPT-4*, Wired (Oct. 5, 2023).

While it is possible to fine-tune an open foundation model with just pre-trained model weights (provided one has access to a suitable new dataset, a training framework/library, and the necessary compute resources), it is much less efficient and valuable to good actors than fine-tuning with access to the model code, model card, and other assets like a user guide, responsible use guide, and acceptable use policy.

This is why we believe that, on balance, responsibly making model assets (*e.g.,* weights, code, and model cards) available for some models is a net-good for society. With them, the open source community has access to the information it needs to meaningfully engage with those models and more effectively counter malicious uses.

We encourage NTIA to work with NIST on establishing the relevant threat models and evaluations for common use across the industry.

**2.b. Could open foundation models reduce equity in rights and safety-impacting AI systems (e.g. healthcare, education, criminal justice, housing, online platforms, etc.)?**

We developed and made available responsible use guides to help downstream developers measure impact. And it's why the Llama 2 Acceptable Use Policy prohibits specific high risk use cases.[19]

In general, we believe that when responsibly open sourced, foundation models can lead to *improvements* in equity. They enable diverse interest groups to scrutinize performance against benchmarks for their particular interest group, and find ways to improve those outcomes.

Meta is also furthering more inclusive and accessible AI by increasing language accessibility and coverage through 'No Language Left Behind' (NLLB), a first-of-its-kind AI project that open sources, under a non-commercial license, models capable of delivering evaluated, high-quality translation directly between 200 languages, including low-resource languages such as Asturian, Luganda, Urdu, and others. It aims to give people the opportunity to access and share web content in their native language and communicate with anyone, anywhere, regardless of language preferences.

[19] For example, the Llama 2 Acceptable Use Policy prohibits the use of the model to "engage in, promote, incite, or facilitate discrimination or other unlawful or harmful conduct in the provision of employment, employment benefits, credit, housing, other economic benefits, or other essential goods and services."

As a real-world example, the technology behind NLLB is supporting Wikipedia editors as they translate information from their native and preferred languages. This helps make more knowledge available to more people around the world.

**2.c. What, if any, risks related to privacy could result from the wide availability of model weights?**

Privacy risks may arise in the context of how open foundation models are deployed downstream. For example, a malicious actor might try to solicit personal information from a model. However, closed models are also vulnerable to some of the same privacy risks, which is why responsibly developing and deploying *all* models must include addressing privacy risks. Training data deidentification and deploying privacy-enhancing techniques, are examples of mitigations that can be taken at the training dataset stage to address these downstream risks. Closed models have the advantage of prompt classification and prompt engineering to identify prompts that may be attempts to extract personal information, or be used to track an individual based on previous locations. However, these system-level techniques have yet to be made resilient to repeated attacks or jailbreaks. Open research has been demonstrably helpful in identifying adversarial privacy attacks, and significant research is ongoing in making AI secure against such attacks.

At Meta, our responsible open source approach includes several steps to mitigate privacy risks at the data layer, such as data removal and techniques to reduce memorization.

In addition to limiting the data that can be used to train a model, we continue to learn about privacy adversarial attacks through our evaluations work, allowing us to test fine-tuned models for whether sensitive personal information could be reproduced, especially by an adversarial actor.[20] And we do not publicly make our training datasets widely available along with our model weights under permissive commercial licenses. (Because of the same concern, mandating access for third parties to training datasets would exponentially increase risk.)

[20] *Id.*

Meta

Additionally, our privacy review process is designed to assess privacy risks that collecting, using, or sharing people's personally identifiable information ('PII') may present, and to help determine whether steps should be taken to mitigate any identified privacy risks, including through the development and use of AI models and tools.

When training Llama 2, we filtered publicly available information to exclude from the dataset certain websites that commonly share PII, like LinkedIn.[21] And we trained and tuned Llama 2 to limit the possibility of private PII from appearing in responses.

All providers and developers of foundation models should have the same minimum baseline obligations to identify and address privacy risks. It is also important that providers and developers are not mandated to provide unrestricted access to the entire tech stack, including training datasets, to third parties.

**2.d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?**

**2.d.i. How do these risks compare to those associated with closed models?**

**2.d.ii. How do these risks compare to those associated with other types of software systems and information resources?**

Malicious state and non-state actors will use any technology they can to advance their objectives. Many such actors are well-resourced and engage in sophisticated malicious cyber activity that is targeted and aimed at prolonged network/system intrusion.

As such, it is likely that state or non-state actors will use widely available model weights – whether because they were responsibly open-sourced or unintentionally released – as well as closed source models available via API to further their goals and interests.

[21] Clark, *Privacy Matters: Meta's Generative AI Features*.

These actors also routinely use open and closed source software for activities such as targeting and exfiltrating highly protected proprietary databases that have little open source value.[22]

**2.e. What, if any, risks could result from differences in access to widely available models across different jurisdictions?**

**2.f. Which are the most severe, and which are the most likely risks described in answering the questions above? How do these sets of risks relate to each other, if at all?**

American interests in competitiveness, economic growth, and global leadership would be jeopardized if the U.S. were to adopt a more restrictive approach to open foundation models than that of other jurisdictions.

The most severe risk would be if the U.S. were to adopt restrictions for open foundation models or otherwise implement an approach that prohibits model providers and downstream developers from determining how their models can be accessed, while other jurisdictions continue to invest in and facilitate the development and adoption of open foundation models.[23] [24]

For example, U.S. policies and rules regarding the standards governing 5G inadvertently "led to considerable uncertainty and, in some cases, chilled U.S. participation in standards development activities involving Huawei." This approach facilitated market dominance by Huawei until U.S. policies were clarified.[25]

---

[22] *Ransomware: The Data Exfiltration and Double Extortion Trends*, Center for Internet Security.
[23] Hunyuan, a large language model developed by China's Tencent, claims to be more capable than GPT-3 and Llama-2 across some benchmarks. Josh Ye, *China's Tencent debuts large language AI model, says open for enterprise use*, Reuters (Sept. 7, 2023). Baidu, another Chinese developer, also claims that its Ernie 4.0 model is more capable than GPT-4. Yelin Mo and Eduardo Baptista, *China's Baidu unveils new Ernie AI version to rival GPT-4*, Reuters (Oct. 17, 2023). Qwen-VL-Max, from Alibaba's Qwen family of open source models, is also claimed to outperform GPT-4V on some benchmarks. Qwen team, *Introducing Qwen-VL*, ( Jan. 25, 2024). The open source Falcon 180B LLM, developed by the UAE Technology Innovation Institute, claims to be "one of the three best language models in the world along with GPT-4 and PaLM-2-Large." Ebtesam Almazrouei & Hamza Alobeidli, *et. al.*, *The Falcon Series of Open Language Models*, (Nov. 29, 2023).
[24] Making AI technologies available to the public under open-source terms is a stated policy objective of China's Global AI Governance Initiative.
[25] U.S. Commerce Department Clarifies Rules for Dealing with Huawei in Standards Setting

U.S. government policy should ensure that companies outside the U.S. do not gain the lead in developing foundational AI technology, or that models developed elsewhere become the global norm. Doing otherwise would not only disadvantage the U.S. economically, but the resulting models likely would not be developed in a way that aligns with U.S. values and standards, and may be subject to the laws or political oversight of countries that are not aligned with the United States and/or its allies.[26]

This result would directly undermine the U.S. government's commitment to "establishing a set of rules and norms for AI, with allies and partners, that reflect democratic values and interests, including transparency, privacy, accountability, and consumer protections."[27]

**3. What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?**

There are extensive and demonstrable benefits to the U.S. of enabling the responsible open sourcing of foundation models, including macroeconomic, cybersecurity, freedom of expression, scientific research, national security, and foreign policy ones, including international development. Perhaps in recognition of these benefits, the U.S. government has well established policies supporting open source, allowing its widespread use and adoption across the federal government.

We discuss each of the benefits further below.

Macroeconomic Benefits of Open Models

In October 2023, Goldman Sachs raised its U.S. GDP forecast to a 2% expansion rate in 2027 and to 2.3% by 2034 based on the impact of AI on the U.S. economy.[28] Foundation models, in particular, may generate between $2.6 trillion to $4.4 trillion in economic growth across the global economy.[29]

[26] *See, e.g.*, *Tehran, Moscow to cooperate on AI ethics*, Tehran Times (March 9, 2024).
[27] *Vice President Harris Announces New U.S. Initiatives to Advance the Safe and Responsible Use of Artificial Intelligence*, White House Fact Sheet (Nov. 1, 2023).
[28] Chris Anstey, *Goldman Sachs Sees AI Lifting US GDP Over Next Decade*, Bloomberg (Oct. 30, 2023).
[29] *The economic potential of generative AI: The next productivity frontier*, McKinsey report (June 14, 2023).

Meta

Given that there are barriers to entry for low-resource actors to develop foundation models as a result of its significant capital costs (even with reducing costs of computational power), it is critical that the government's policies on open foundation models maximize the benefits to competition and innovation.[30]

Meta developed and open-sourced PyTorch, a powerful, flexible and easy-to-use workflow for building machine learning models, in 2016.[31] Since then, it has become the leading tool for AI research and the development of such models around the world, with an extensive community of open source developers continuously contributing to improvements and innovation across the ecosystem (over 2,400 contributors have built more than 150,000 projects on the framework).[32]

Open Foundation Models' Importance to the U.S. Government

The U.S. government uses AI, including open foundation models, in a variety of ways across different agencies and departments.

For example:

- The National Aeronautics and Space Administration (NASA) collaborated with IBM to develop and open source the largest geospatial foundation model on Hugging Face. IBM fine-tuned the model to allow users to map past U.S. floods and wildfires, measurements that can be used to predict future areas of risk. With additional fine tuning, the model could be redeployed for tasks like tracking deforestation, predicting crop yields, and detecting and monitoring greenhouse gasses. The project coincided with NASA's Year of Open Science, a series of events to promote data and AI model sharing. It is also part of NASA's decade-long Open-Source Science Initiative to build a more accessible, inclusive, and collaborative scientific community.[33]
- The National Institutes of Health (NIH) uses AI models in biomedical research. For instance, the NIH Clinical Center released an open-source dataset of 32,000 CT images to help the scientific community improve the detection of lesions.[34]

[30] Kapoor & Bommasani, *et. al.*, *On the Societal Impact of Open Foundation Models*.
[31] Frameworks like PyTorch are the vehicle for architectural exploration, allowing the research community to iterate on different ideas and then release code for others to use and build on.
[32] *Announcing the PyTorch Foundation: A new era for the cutting-edge AI framework*, Meta AI blog (Sept. 12, 2022).
[33] Kim Martineau, *IBM and NASA open source the largest geospatial AI foundation model on Hugging Face*, IBM blog (Aug. 3, 2023).
[34] *NIH Clinical Center releases dataset of 32,000 CT images*, NIH press release (July 20, 2018).

- The U.S. Department of Agriculture (USDA) uses AI to predict crop yields, optimize irrigation, and monitor soil health. For example, the USDA's National Agricultural Statistics Service uses machine learning models to analyze satellite imagery and predict crop production.[35]
- The Department of Defense's Project Maven uses AI to interpret video images, which could be used to improve drone footage analysis. The project initially used TensorFlow, an open-source machine learning framework developed by Google.[36]
- The U.S. Census Bureau uses machine learning models to improve the accuracy and efficiency of the decennial census. For example, it uses computer vision models to analyze satellite imagery and identify housing units, which helps to ensure that every household is included in the census.[37]
- The U.S. General Services Administration (GSA) has an AI for Citizen Services program that aims to make public data more accessible and useful. It uses AI models to create chatbots and virtual assistants that can answer questions about public services.[38]

Additionally, the U.S. government, through agencies and contractors, is already using or proposing to use Llama 2.[39] In the 2023 End of Year Report on the Open Source Software Security Initiative, NCS commits that, "in partnership with the private sector and the open-source software community, the federal government will also continue to invest in the development of secure software, including memory-safe languages and software development techniques, frameworks, and testing tools."[40]

[35] Scott Elliott, *Artificial Intelligence Improves America's Food System*, USDA Blog (Dec. 10, 2020).
[36] Samuel Gibbs, *Google's AI is being used by US military drone programme*, The Guardian (March 7, 2018).
[37] *Why does the U.S. Census Bureau Need Machine Learning?*, US Census Bureau blog (Oct. 28, 2021).
[38] Justin Herman, *Opening Public Services to Artificial Intelligence Assistants*, GSA blog (Jan. 6, 2017).
[39] Cecilia Kang, *The Department of Homeland Security is Embracing AI*, NYT (March 18, 2024).
[40] *Securing the Open-Source Software Ecosystem: End of Year Report* (Jan, 2024).

Meta

Competition and Innovation

Open models set a "floor" for competition, incentivizing innovation while ensuring that no one actor can capture the "baseline" and extract undue rents.[41] Broadening access to models allows more people to use them for their own innovations, which in turn enables greater competition in downstream markets, helping to reduce market concentration at the foundation-model level from vertical cascading.[42]

We have already seen impressive organic adoption of Llama 2 by downstream developers around the world for purposes ranging from gaming to clinical decision-making.[43] Significantly, developers are not only using, but improving upon, openly released models: "[O]pen-source developers have created thousands of derivatives of models like Llama, including increasingly, mixing models – and they are steadily achieving parity with, or even superiority over closed models on certain metrics (e.g., FinGPT, BioBert, Defog SQLCoder, and Phind)."[44]

Finally, we are particularly invested in supporting the adoption of Llama 2 (and other models) for innovative social impact purposes – which has led to the creation of our Llama Impact Grants program. The goal of the program is to identify and support the most compelling applications of Llama 2 for societal benefit, and the volume of applicants (over 800) demonstrates Llama 2's significant potential for promoting social good. Examples of some of the applications of Llama 2 under the programme include: building a Medical-Vision Language Model (Med-VLM) that can process medical images and provide high-quality textual answers to medical questions in various languages (Barcelona Supercomputing Center); and leveraging Llama 2 to enhance the process of matching cancer patients with clinical trials and then integrating these capabilities into the open-source MatchMiner platform (Dana-Farber Cancer Institute).[45]

[41] Indeed, the Securities and Exchange Commission Chair, Gary Gensler, has also noted the risk to the financial sector if there is overreliance on one base model. *See SEC chair Gensler on AI's threat to Wall Street: "I don't want everybody to drive off the cliff,"* Politico Tech Podcast (March 19, 2024).
[42] *Id.*
[43] *Discover the possibilities of building on Llama*, Meta blog.
[44] Matt Marshall, *How enterprises are using open source LLMs: 16 examples*, Venturebeat (Jan. 29, 2024).
[45] *See* the Llama Impact Grants landing page.

Meta

Security

Open sourcing is a longstanding, well-regarded approach to enhancing security, as demonstrated by these examples:

- Historically, many software vendors sought to prevent security researchers from publishing information about software vulnerabilities. Ultimately, researchers' disclosures created public pressure for software vendors to enhance security and develop more secure software.
- Open source software is an essential component supporting cybersecurity in the federal government. The Cybersecurity and Infrastructure Security Agency (CISA) notes that "open source software is widely used across the federal government and every critical infrastructure sector."[46]
- The Department of Defense is also highly engaged with the open source community. Specifically, the Defense Advanced Research Projects Agency (DARPA) regularly engages with the Open Source Security Foundation (OpenSSF)[47] and has offered awards up to $1million USD for work helping to "protect the integrity of open source infrastructure critical to the DoD."[48]

In our experience, instead of creating more new risks than benefits, open source releases have helped us, and the broader community of developers, build safer and more robust systems. By democratizing access, vulnerabilities are continuously identified and mitigated by an open community, and that creates safer products.

For example, researchers were able to test Meta's earlier generative LLM, BlenderBot 2, to uncover ways that the model can be tricked into remembering misinformation—ensuring that BlenderBot 3 was more resistant to such attacks.[49]

As a further example, recent academic research into the security of closed and open models (including Llama 2) identified vulnerabilities and will help developers identify appropriate solutions[50] (we are taking note of the findings in the paper for our future benchmark work).

[46] *See* the Open Source Software Security page of the Cybersecurity & Infrastructure Security Agency.
[47] *See* DARPA's Artificial Intelligence Cyber Challenge.
[48] *See* Hybrid AI to Protect Integrity of Open Source Code (SocialCyber).
[49] *BlenderBot 3: A 175B parameter, publicly available chatbot that improves its skills and safety over time*, Meta blog (Aug. 9, 2022).
[50] Fengqing Jiang, Zhangchen Xu, Luyao Niu, *et al.*, *ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs* (Feb. 22, 2024).

Other companies are also invested in using open source to improve security. For example, Google recently launched the AI Cyber Defense Initiative and open sourced Magika, a new AI-powered tool it hopes will benefit defenders more than attackers.[51]

Scientific Research

Continued investments in AI openness and science will fuel the next generation of AI discoveries. Open research[52] furthers our collective fundamental understanding in both new and existing domains across the full spectrum of AI-related topics related, which advances the state of the art. This in turn benefits the entire ecosystem, including closed-source model providers and downstream developers.

Jeff Boudier, head of product and growth at Hugging Face, notes, "AI remains a science-driven field, and science can only progress through information sharing and collaboration. This is why open-source AI and the open release of models and datasets are so fundamental to the continued progress of AI, and making sure the technology will benefit as many people as possible."[53]

Open foundation models are critical to research in areas spanning from interpretability, security research, and improvements to model training and inference efficiency.[54] Open source enables scientific scrutiny within a research community and experimentation with advanced technologies. As a recent statement by over 100 leading biologists noted, "many researchers in our community benefit from open-source scientific software, which has enabled rapid innovation and broad collaboration."[55]

[51] Phil Venables and Royal Hansen, *How AI can strengthen digital security*, Google blog (Feb. 16, 2024).
[52] Meta defines open research as a collaborative and transparent approach to scientific inquiry that encourages the sharing of data, methods, and results with other researchers and the broader public. This approach is designed to promote the advancement of knowledge and the development of new technologies by allowing others to build upon and verify the findings of previous research. Open research at Meta involves making our data, methods, and results available to other researchers and the public through various channels, such as open-access publications, datasets, and software repositories. We also engage in collaborations with other institutions and researchers to further advance the field and promote the use of AI for social good.
[53] *IBM and NASA open source the largest geospatial AI foundation model on Hugging Face*, IBM blog (Aug. 3, 2023).
[54] Rishi Bommasani & Sayash Kapoor, *et. al.*, *Considerations for Governing Open Foundation Models*, Stanford Center for Research on Foundation Models (CRFM), (Dec. 13, 2023).
[55] *Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design*, Responsible AI x Biodesign (March 8, 2024).

Meta

Similarly, Meta's release of the open source OPT language model in 2022[56] enabled recent advancements in watermarking large language models at the University of Maryland.[57] And work by Meta's Fundamental AI Research (FAIR) team is driving advances that look beyond the transformer architecture of models (the architecture that current state-of-the-art models like Llama 2 and GPT-4 are built on). This includes FAIR's work on V-JEPA, which is a new self-supervised learning and objective-driven architecture.

Research on new model architectures is valuable in areas like efficiency: While transformer models have achieved state-of-the-art results on many tasks, they can be computationally intensive and require a lot of memory, which can be a limitation for certain applications or devices. Alternative architectures might offer similar performance with fewer computational requirements, thereby driving both performance up and costs/accessibility down.

The necessary, continuous work on AI safety requires openness; as several AI experts recently noted, "[c]urrent safety research is often limited by insufficient access to large, cutting-edge models and relevant information such as their architecture and training processes."[58]

National Security and Foreign Policy

U.S. leadership in open source will result in models trained on datasets that reflect U.S. values and that have not been subject to censorship via fine tuning. These models will become the standard on which developers around the world build, and on which further innovation is based—embedding U.S. values and maintaining the U.S. lead in AI innovation. In turn, this will give the U.S. additional influence to shape AI governance conversations in international institutions.

Open source models strengthen cybersecurity by scaling cybersecurity defenders and enable classified use cases for generative AI because they can be hosted locally and do not rely on cloud infrastructure.

[56] Will Douglas Heaven, Meta Has Built a Massive New Language AI—and It's Giving It Away for Free, MIT Technology Review (May 3, 2022).
[57] Written Testimony of Clement Delangue, Co-Founder and CEO, Hugging Face, For a Hearing on 'Artificial Intelligence: Advancing Innovation Towards the National Interest (June 22, 2023).
[58] Elizabeth Seger & Noemi Dreksler, *et. al.*, *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives*, LPP Working Paper No. 4-2023 (Oct. 25, 2023).

More broadly, open source democratizes access to AI technology, which can support U.S. foreign and development policy in the Global South. Open source AI can act as a force multiplier for countries to help deliver services to their populations, close inequality gaps, and fuel new technologies and enterprises. The U.N. has embraced open source AI as a key tool to achieve the Sustainable Development Goals, an ambition the U.S. shares.[59]

**3.a. What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy? In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/training in computer science and related fields?**

*See response to Question 3.*

**3.b. How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?**

*See response to Question 3.*

**3.c. Could open model weights, and in particular the ability to retrain models, help advance equity in rights and safety-impacting AI systems (e.g. healthcare, education, criminal justice, housing, online platforms etc.)?**

*See response to Question 2 (b).*

**3.d. How can the diffusion of AI models with widely available weights support the United States' national security interests? How could it interfere with, or further the enjoyment and protection of human rights within and outside of the United States?**

*See response to Questions 2 (b) and 2 (f).*

[59] *See* remarks by Secretary Antony J. Blinken at the AI for Accelerating Progress on Sustainable Development Goals Event (Sept. 18, 2023).

**3.e. How do these benefits change, if at all, when the training data or the associated source code of the model is simultaneously widely available?**

As discussed in response to Question 1, a responsible approach to open source means balancing a range of factors (such as privacy or business imperatives) and considering various aspects of openness in the tech stack (*e.g.,* source code, training datasets, model weights, and other elements).

Because of the variety of factors at play – and their corresponding mitigations – it is hard to categorically assess the benefits of releasing with or without the training data and source code.

However, while it is possible that making training data, source code, and model weights all simultaneously widely available could allow malicious, determined and well-resourced downstream developers or other actors to solicit personal information from the model, releasing code with model weights can lead to positive security and safety outcomes. (See Question 2, above). It is therefore important that providers of foundation models are able to choose the type of release that best balances the benefits of information-sharing, the potential costs to the provider's business, and the risks of misuse.

**4. Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights? If so, please list them and explain their impact.**

*See response to Question 3 (e).*

**5. What are the safety related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available weights?**

In principle, responsible development of foundation models should include robust risk assessments and evaluations (which may include red-teaming for certain models) that are commensurate with the potential risks involved, are transparent, and include minimum guardrails. Common standards should be based on existing frameworks (*e.g.,* the PAI Guidance for Safe Foundation Model Deployment) and the work of agencies like NIST and NTIA. While this work is ongoing, it is premature to require specific measurements and benchmarks.

We provide further detail on these points in response to the questions below.

Meta

**5.a. What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?**

The development of foundation models is a new field, and, at present, there is a lack of consensus about how best to measure and evaluate certain risks. Ultimately, determinations around model evaluations need to be based on repeatable measurements. Helpfully, valuable work is underway and progress is being made by numerous research institutions, NIST, the U.S. AI Safety Institute, and the MLCommons initiative.

These efforts should advance the development of evaluations on issues such as: standardized harm categories; violent crimes; non-violent crimes; sex-related crimes; child sexual exploitation; indiscriminate weapons (Chemical, Biological, Radiological, Nuclear, and high yield Explosives 'CBRNE'); defamation; specialized advice; privacy; intellectual property; elections; hate; self-harm; and sexual content. Such evaluations would help ensure a shared understanding on the most appropriate methodologies for conducting risk/benefit analysis. Their development should also include various measurements related to the ability of the model in question to enable/encourage/endorse these activities; standardized measurement on these evaluations will be important in the context of risk/reward calculations.[60]

Until consensus develops, it is premature to require specific measurements and benchmarks. Doing so likely would lead to reliance on potentially unreliable – and soon to be outdated – metrics. Mandating specific benchmarks at this point is likely also to lead to fragmented approaches, making it impossible for governments, researchers, deployers, and users to compare the safety and performance of different models. For these reasons, we urge the NTIA to work with NIST on standardizing the threat models and evaluations along the AI value chain.

[60] We also would note the work of Singapore's Infocomm Media Development Authority (IMDA), which collaborated with the AI Verify Foundation to catalog evaluations. *Cataloging LLM Evaluations*, AI Verify Foundation (Oct. 2023).

Meta

Meta's Approach to Evaluations

Red-teaming

One type of evaluation, red-teaming, has become a key component of our AI development process for certain models and we continue to invest in research to make it more efficient and effective.

We significantly invested in red-teaming for Llama 2. Over 350 people were involved, including employees, contractors, and external vendors. The process included domain experts in cybersecurity, election fraud, social media misinformation, legal, policy, civil rights, ethics, software engineering, machine learning, responsible AI, and creative writing. It also included individuals representative of a variety of socioeconomic, gender, ethnicity, and racial demographics; this diversity is important for any successful red-teaming exercise.

The red-teamers probed our models across a wide range of risk categories, such as criminal planning, human trafficking, regulated or controlled substances, sexually explicit content, unqualified health or financial advice, and privacy violations, as well as different attack vectors, such as hypothetical questions, malformed/misspelled inputs, and extended dialogues.

For Llama-2 we also conducted specific tests to determine the capabilities of our models to facilitate the production of weapons (*e.g.,* nuclear, biological, chemical, and cyber).[61] And we submitted Llama 2 to the DEFCON convention in Las Vegas in 2023, alongside other companies like Anthropic, Google, Hugging Face, Stability, and OpenAI. At that event, over 2,500 hackers analyzed and stress tested their capabilities – making this one of the largest public red-teaming events for AI. To support the future release of large foundation models, we have engaged outside subject- matter experts to support our testing and threat prioritization for CBRNE, with a specific emphasis on chemical and biological risks. We will continue to capacity build for evaluations and red-teaming in this area.

But while we believe that red-teaming is an important evaluation tool, it should not be the only evaluation tool used or necessarily required for all foundation models.

[61] Hugo Touvron, Louis Martin, *et. al.*, *Llama 2: Open Foundation and Fine-Tuned Chat Models* (July 19, 2023).

Red-teaming is a costly undertaking,[62] and is limited by the expertise and scope of the red-teamers (for example, if a red-team group is only composed of experts in one particular area, it will not be best equipped to test for vulnerabilities outside of that scope).[63]

Model providers need to be able to deploy a range of evaluations, including tools like benchmarking, human evaluation, A/B testing, cross-validation, and adversarial attacks.

Open Safety Benchmarks and Classifiers

Meta is committed to advancing AI safety and safety research on evaluations. As part of this commitment, we developed model evaluations and open safety tooling through our Purple Llama Project – a major step toward enabling community collaboration and standardizing the development and use of trust and safety tools for generative AI development.

Two components of the Purple Llama Project are CyberSecEval and Llama Guard. We believe CyberSecEval is the most extensive cybersecurity safety benchmark to date. In our Code Llama-Instruct model, we use CyberSecEval to understand and mitigate cyber risks, prior to release, providing good-faith actors with one of the safest coding copilot tools available–which we open sourced for both research and commercial purposes.

Our second tool, Llama Guard, is a safety classifier for filtering input and output that is trained to detect certain problematic content. We have made both of these components available on an open-source basis so that other model providers and downstream developers can leverage these advances for their own safety fine-tuning.

[62] For example, red-team penetration testing costs between $10,000 to $85,000 and typically runs for several weeks. *See* Josh Gormally, *How Much Does Red Team Penetration Testing Cost In 2023?*, Network Assured (Sept. 26, 2023).
[63] *See* The Center for Advanced Red Teaming, University of Albany landing page.

Meta

Industry Initiatives

Meta is also part of a number of industry initiatives focused on model evaluations. These initiatives are driving consensus on key foundational issues that we believe will be important additions to the evolution of the AI governance ecosystem.

(i) AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models

Issued by the University of California Berkeley AI Research Lab, this provides specific guidance on how to apply the NIST AI Risk Management Framework (RMF) to providers of general-purpose AI systems (intended as an umbrella term for 'foundation models,' 'frontier models,' and 'generative AI').

(ii) Open Loop US Program

This program explores how NIST's AI RMF can be adapted to provide guidance on generative AI, providing evidence gathered from 40 companies across various industries on their current risk management practices for generative AI and identifying gaps around information, benchmarks, standards, and resources.

(iii) Partnership on AI Guidance for Safe Foundation Model Deployment

This represents the most comprehensive and multi-stakeholder AI self-regulatory approach to date, and in addition to aligning with the White House Voluntary AI Commitments introduces important recommended practices, including for the responsible open-sourcing of foundation models. It also touches on the distribution of responsibility across all actors in the AI value chain and its lifecycle. (See Question 6 regarding actors in the value chain).

(iv) AI Safety Working Group at the MLCommons non-profit consortium

This initiative aims to establish and maintain a toolkit of globally viable, standardized metrics and benchmarks for evaluating and measuring the potential risks of generative AI. This work includes building and maintaining AI tooling infrastructure for benchmarking large language models for safety.

Meta

Across 2024, Meta will open source some of our research outputs, datasets, and other assets in collaboration with MLCommons, which we believe is key to helping the AI community's research efforts to bring Responsible Use Guides to life around a shared set of evaluations and mitigations.

(v) AI Alliance
This is a membership organization Meta co-founded with IBM that specifically advances open source and open innovation approaches. With more than 80 members so far, the AI Alliance seeks to work with global stakeholders to create resources and benchmarks for the responsible and safe development and deployment of AI.

**5.b. Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?**

**5.c. What are the prospects for developing effective safeguards in the future?**

Effective safeguards include investing in access control, infrastructure security, and insider/external threat detections to ensure model weights do not become available before a planned/controlled release, as well as for models that will not be made available on an open basis. In addition to implementing these safeguards, Meta's approach is to articulate the risk attributes of models and to automate access control and detection capabilities proportional to these risk attributes.

We have released multiple research projects such as Reasoning over Public and Private Data in Retrieval-Based Systems, Watermarking Makes Language Models Radioactive, and An Efficient Algorithm for Integer Lattice Reduction which all make progress toward exploratory research with potential safeguard applications, from personal data mitigations, to novel watermarking techniques to potential attacks of certain vulnerable cryptosystems if combined with other methods. The research needs to advance as the current results do not yet transfer to product use cases, but Meta is actively investigating how advancements in our and others' research can help develop safeguards in the future.

*See response to Question 5 (a) on evaluations and red-teaming.*

**5.d. Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?**

We are not aware of technical capabilities that could not be overcome by determined, well-resourced, and capable actors that would allow for regaining control over model weights that have become widely available. For this reason, our responsible release approach focuses on pre-release assessment and mitigation, as well as post-release assessment that can leverage the broader open source community to identify risks. This enables us to improve both future releases of our models and system-level mitigations like our Purple Llama suite of tools, which we can rapidly deploy for developers to use as a part of the systems they operate.

**5.e. What if any secure storage techniques or practices could be considered necessary to prevent unintentional distribution of model weights?**

Computing and storage environments should be actively managed and monitored, physically and logically secured, and supported by a variety of insider threat detection programs corresponding to the level of risk presented.

At Meta, we perform training and evaluation of AI models ranging in sensitivity from purely open source research on public data to proprietary models used as part of consumer products. We utilize computing environments that are appropriately secured for the level of sensitivity of the project.

In relation to frontier AI systems, Commitment 3 of the White House Voluntary AI Commitments specifically addresses investments in "cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights."[64] For our unreleased AI model weights for frontier AI systems as defined in the White House Commitments, we are committed to treating them as core intellectual property for our business, especially with regards to cybersecurity and insider threat risks.

[64] Voluntary AI Commitments, White House (July 21, 2023)

As part of that, until we have completed our pre-release assessment efforts for frontier AI systems, we will limit internal access to those model weights to people whose job function requires such access, and we have in place a robust insider threat detection program consistent with protections provided for our most valuable intellectual property and trade secrets. In addition, for these models, we will work with the weights in an appropriately secure environment to reduce the risk of unapproved release.[65]

**5.f. Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.**

The components and type of access required to assess a model depends on the kind of assessment conducted. It is possible, for example, to conduct adversarial testing without access beyond that of a typical user. Many AI researchers conduct AI model experiments using publicly available interfaces (*e.g.,* Google Colab, Hugging Face, Chatbot Arena).

Similarly, it may be possible to verify that a foundation model meets certain requirements using documentation such as a model card or other transparency documentation about how a model was developed or how it performs on certain evaluations and benchmarks. This kind of information is often included in model cards or research papers that accompany foundation model releases (*e.g.,* the artifacts Meta provided for the Llama 2 release).

However, deeper and more reliable analysis of a model's capabilities can be conducted with model weights. This is especially valuable for understanding models that aren't integrated with system-level, end-to-end protection layers. For example, if a model is not integrated with appropriate access controls or encryption, analyzing the model weights could reveal whether sensitive information is stored in the weights or whether the model relies on certain features that could be used to launch attacks.

[65] Overview of Meta AI safety policies prepared for the UK AI Safety Summit, Transparency Center (Oct. 20, 2023).

Meta

Similarly, if a model is deployed in a critical infrastructure system without appropriate network security measures or monitoring in place, analyzing the model weights could reveal whether the model is vulnerable to adversarial attacks or other types of manipulation.

Furthermore, for a large number of red-teamers representing a community of typical users, the testing environment needs to be user-friendly so that the red-teamers do not spend valuable time learning how to use model weights. In contrast, to emulate threat actors of a different type, such as determined state actors or other nefarious threat actors, it is important to make the model weights available because such actors are more likely to have access to them.

*See response to Question 5 (a) regarding evaluations.*

**5.g. Are there means by which to test or verify model weights? What methodology or methodologies exist to audit model weights and/or foundation models?**

Open sourcing responsibly is the best means by which to test or verify model weights because the pre-trained model weights are readily available. Otherwise one is reliant on available documentation, comparing performance of the closed model to a baseline model or other forms of external validation.

As part of responsibly developing foundation models, transparency is essential. When Meta released Llama 2, alongside releasing the model weights, we developed and shared a number of artifacts relaying how we developed the model to help developers, researchers, and policymakers better understand its capabilities and limitations (*e.g.,* model card, research paper).

These artifacts, most importantly the Llama 2 Research Paper, were designed to provide more information about the process of developing Llama 2 and the steps we took to do so responsibly. We believe that this approach, in addition to providing the pre-trained model weights, further facilitates testing or verification.

Meta

**6. What are the legal or business issues related to open foundation models?**

There are several parties involved in the AI value chain: foundation model providers (who train foundational models, proprietary or open-source, that others may build on as well as interfaces to interact with the models),[66] downstream developers (who use model weights to create their own use cases using their own data for fine-tuning the model), model deployers (who can be the same as the downstream developer), and end-users (those who interact with the fine-tuned model). There needs to be clear differentiation of responsibilities across this value chain for legal and business clarity. Intellectual property and generative AI needs to be considered in the context that generative AI models and systems are tools that enhance human creativity and productivity.

We expand on these points further below.

**6.a. In which ways is open-source software policy analogous (or not) to the availability of model weights? Are there lessons we can learn from the history and ecosystem of open-source software, open data, and other "open" initiatives for open foundation models, particularly the availability of model weights?**

**6.b. How, if at all, does the wide availability of model weights change the competition dynamics in the broader economy, specifically looking at industries such as but not limited to healthcare, marketing, and education?**

*See response to Question 3.*

**6.c. How, if at all, do intellectual property-related issues—such as the license terms under which foundation model weights are made publicly available—influence competition, benefits, and risks? Which licenses are most prominent in the context of making model weights widely available? What are the tradeoffs associated with each of these licenses?**

*With respect to licenses, see response to Question 6 (d).*

*With respect to competition benefits of open source models, see response to Question 3 (a).*

[66] *See* PAI's Guidance for Safe Foundation Model Deployment.

Meta

Regarding intellectual property related issues more broadly, the role of a generative AI model is to enhance human creativity and productivity. Generative AI is a tool, no different from the printing press, the camera, or the computer. Those technologies changed the nature of creative endeavors, and were a tremendous boon to human creativity and productivity. Generative AI will be no different.

As with any tool, people ultimately are responsible for how to use that tool. Just as photographers choose the settings on a camera and the subjects they wish to photograph, the person using a generative AI model is the one who provides the prompts to the model that determine the content of the output–and is the one that decides how to use that output.

The purpose of these models is to enable people to create new creative outputs to suit their preferences and needs. In that sense, a generative AI model is not different from other neutral commercial technologies accepted under copyright law, from sound mixing to digital photo editing tools and beyond.[67]

**6.d. Are there concerns about potential barriers to interoperability stemming from different incompatible "open" licenses, e.g., licenses with conflicting requirements, applied to AI components? Would standardizing license terms specifically for foundation model weights be beneficial? Are there particular examples in existence that could be useful?**

Providers of foundation models necessarily need to be able to decide, based on their business considerations, and risk assessments of their model, what an appropriate release should look like along the "gradient of release," including what aspects of the tech-stack should, or should not, be made available. Incompatibility between existing open source licenses already exists in the marketplace[68] and consumers of open source code already have to make project decisions based on these incompatibilities.[69]

[67] For more detailed discussion of the Copyright Act and generative AI, *see* Meta Platforms Inc submission to the US Patent and Trademark Office.

[68] For example, software that combined code released under version 1.1 of the Mozilla Public License (MPL) with code under the GNU General Public License (GPL) could not be distributed without violating one of the terms of the licenses–regardless that both licenses were approved by both the Open Source Initiative and the Free Software Foundation.

[69] Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations* (Feb. 5, 2023).

**7. What are the current or potential voluntary, domestic regulatory, and international mechanisms to manage risks and maximize the benefits of foundation models with widely available weights? What kinds of entities should take a leadership role across which features of governance?**

In addition to longstanding laws that already apply to generative AI, an increasing number of new voluntary initiatives, domestic regulations, and international laws and principles are emerging.

Given the significant potential for regulatory fragmentation, the U.S. government has an important role to play – domestically and internationally – to ensure consistency and interoperability, particularly in driving alignment on net-new definitions, taxonomies, risk identification, assessments, and mitigations.

As noted by Vice President Kamala Harris, "to provide order and stability in the midst of global technological change, [ . . . ] we must be guided by a common set of understandings among nations."[70]

To achieve this common understanding, we suggest that the U.S. government focus on identifying net-new technological and regulatory issues related to foundation models, continue investing in the most influential international forums, and drive cooperation on technical standards to advance regulatory interoperability.

Specifically, we recommend:

- **Focusing on the net-new.** We welcome the approach of the U.S. government and Congress thus far to take time to identify what the net-new issues are with respect to generative AI. The effort across the entire federal government under the Executive Order represents a commitment to working collaboratively with all stakeholders to develop evidence-based policies on complex, novel issues. This approach is a helpful contrast to the one adopted by the European Union and other jurisdictions that have sought to regulate first and understand later.

[70] Remarks by Vice President Harris on the Future of Artificial Intelligence, Office of the White House (Nov. 1, 2023)

- **Continuing to invest in the most advanced and influential international efforts.** With extensive work underway internationally, the U.S. has an important role to play connecting work under the E.O. to the most influential fora, including the OECD's work to develop a framework for the implementation of the G7 Hiroshima Process voluntary Code of Conduct. We also welcome U.S. leadership in securing approval by the U.N. General Assembly of a resolution on safe, secure, and trustworthy AI systems for sustainable development,[71] which we believe is a meaningful step toward the interoperable, global standards we need to promote a safe, vibrant, open, and inclusive AI ecosystem.
- **Cooperating on technical standards.** We recommend a formal cooperation and collaboration framework between the U.S. AI Safety Institute and similar institutions around the world. This will drive alignment on fundamental common standards and avoid duplication of efforts. Aligning the work of standard-development organizations[72] will also be an important aspect of these efforts.
- **Advancing regulatory interoperability.** Rather than look to implement hard law at this time, the government should focus efforts on building consensus and alignment across initiatives such as the White House Voluntary AI Commitments, the forthcoming NIST AI RMF companion guide for generative AI, the Munich Security Conference AI Elections Accord, the Bletchley Declaration, and industry frameworks such the Partnership on AI Synthetic Media Framework, the Guidance for Safe Foundation Model Deployment, and the work of the AI Alliance, MLCommons, Frontier Model Forum, and AI Verify Foundation.

**7.a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?**

*See response to Question 5 (d).*

[71] More Than 50 UN Member States Join U.S. in Urging Support for Proposed U.S.-Drafted General Assembly Resolution on AI

[72] There are broadly four main international standards development organizations in the field of AI: the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), the International Telecommunications Union (ITU), and the IEEE Standards Association (IEEE). See also work of the AI Standards Hub.

**7.b. How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?**

For the reasons cited in Question 3, *restricting* open foundation models could frustrate governments objectives around AI transparency, competition, fairness, innovation, maintaining U.S. leadership, and security.

**7.c. When, if ever, should entities deploying AI disclose to users or the general public that they are using open foundation models either with or without widely available weights?**

*See response to Question 1 regarding Meta's approach to responsible open source.*

**7.d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?**

*See response to Question 7.*

**7.d.i. Should other government or non-government bodies, currently existing or not, support the government in this role? Should this vary by sector?**

*See response to Question 7.*

**7.e. What should the role of model hosting services (e.g. Hugging Face, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?**

Irrespective of whether a model is closed or open, hosting services could provide a form of notice to consumers to indicate whether a model has met industry accepted benchmarks, such as those in development by NIST.

Meta

Hosting services can significantly help disseminate information about AI safety to the general public by noting their own safety rankings and publishing other model evaluations.[73] We also note the work of the Department of Commerce through the NPRM implementing EO 13984 & EO 14110, and draw NTIA's attention to previous comments in submissions on the negative impact of potential obligations on IaaS providers , such as Know Your Customer, to the U.S.' macroeconomy and global digital leadership.[74]

**7.f. Should there be different standards for government as opposed to private industry when it comes to sharing model weights of open foundation models or contracting with companies who use them?**

When considering whether different standards should apply to the government as opposed to private industry, the determining factors should be the use of a model and the risks of that use – not whether the foundation model is open or not. For example, a government may use foundation models for a range of legitimate public policy objectives that carry a potentially higher risk for individual rights – such as the use of foundation models to provide social security benefits, or for law enforcement and military purposes.

Because these uses could have legal or substantially similar effects on individuals, they should be subject to a higher standard of risk assessment, fairness analysis and mitigation, transparency, accountability, and recourse.

**7.g. What should the U.S. prioritize in working with other countries on this topic, and which countries are most important to work with?**

*See response to Question 7.*

**7.h. What insights from other countries or other societal systems are most useful to consider?**

*See response to Question 7.*

[73] Chenhui Zhang, *et. al.*, *An Introduction to AI Secure LLM Safety Leaderboard*, HuggingFace blog (Jan. 26, 2024).
[74] Miller, John, et. al., ITI Comments Responding to Commerce Department Advance Notice of Proposed Rulemaking on Taking Additional Steps to Address the National Emergency with Respect to Significant Malicious Cyber-Enabled Activities (EO 13984) (RIN # 0605-AA61; Docket No. 210913-0183), ITI (Oct. 25, 2021)

**7.i. Are there effective mechanisms or procedures that can be used by the government or companies to make decisions regarding an appropriate degree of availability of model weights in a dual-use foundation model or the dual-use foundation model ecosystem? Are there methods for making effective decisions about open AI deployment that balance both benefits and risks? This may include responsible capability scaling policies, preparedness frameworks, et cetera.**

*See response to Question 1.*

**7.j. Are there particular individuals/entities who should or should not have access to open-weight foundation models? If so, why and under what circumstances?**

We know there is a small, determined group of bad actors who will go to extreme lengths to gain access to technologies, including AI technology, regardless of whether it is open or closed. There is always the possibility that models will be subject to unintentional release – either through lapses in security, leaks, or adversarial attacks. So providers of foundation models – open or closed – should operate on the basis that model weights, and other artifacts, could be accessed by bad actors or become publicly available, and act accordingly.

As part of our commitment to responsible open source, Meta evaluates our models on a variety of risk vectors prior to releasing our models under permissive commercial licenses, including risks related to misuse.

**8. In the face of continually changing technology, and unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?**

It is precisely because of technology's continually changing nature, as well as its potential risks and benefits, that policy measures should not fossilize in hard law obligations based on technology itself, but rather focus on uses and allocating responsibility across the value chain, with a focus on end-use.

It is also for this reason that a rush to regulate now, before common standards, taxonomies, and cross-industry technical work has matured, will render such regulation redundant at best, and at worst lead to poor outcomes for American innovation, individual rights, and global competitiveness.

This is why multi-stakeholder efforts, like that of the U.S. AI Safety Institute, and in international fora like the G7 and OECD, are central to the ability for governments, companies, and individuals to formulate durable, balanced solutions to the challenges and opportunities of today, and the future.

**8.a. How should these potentially competing interests of innovation, competition, and security be addressed or balanced?**

*See response to Question 8.*

**8.b. Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 10^26 integer or floating-point operations used in the Executive Order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?**

**8.c. Are there more robust risk metrics for foundation models with widely available weights that will stand the test of time? Should we look at models that fall outside of the dual-use foundation model definition?**

The amount of computational resources required to build a model is not a useful proxy for risk over time. It does not correlate with the ability of the model to result in outcomes that would pose a significant risk beyond existing technology and it does not take into account that smaller fine-tuned models could be capable of such risks with significantly fewer computational resources. Scaling laws and the decreasing costs of compute suggest that both pre-training and fine-tuning will be possible with far less compute power than contemplated by the threshold set forth in the E.O. NIST's work to develop benchmarks for evaluations will be important to establishing common standards on this.

Further, it is generally important to distinguish between "dual use foundation models" and models that fall outside of that definition.

**9. What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights?**

*See response to Question 1.*

If the United States seeks to impose restrictive limitations on open foundation models, in addition to curtailing the work underway by NIST to establish common standards for the safe development and deployment of AI, it would have significant negative effects on U.S. interests. Such actions could constrain the ability of model providers to open source models, including within the U.S.

Given broad-based enthusiasm for open source AI globally, including among U.S. allies and the Global South, restrictions could diminish U.S. leadership and standing related to AI governance and other important issues. Since open source drives innovation and sets the standard on which others will build, the U.S. would cede the pole position in AI leadership and innovation, depriving the U.S. economy of related growth opportunities.

Such restrictions would also leave a vacuum that other countries would be eager to fill and benefit from, with no guarantee that the models they enable will reflect U.S. values or vision for the future.

As Freedom House has noted, "authoritarian governments are building centralized foundation models that limit access to accurate information and embed censorship."[75] The U.S. Department of State, Bureau of Democracy, Human Rights and Labor similarly found that, "authoritarian governments such as the People's Republic of China attempt to influence the fundamental character of the global Internet with the advancement of protocols and standards that enable more centralized Internet control, for the purposes of censorship and surveillance, as well as the abuse of other human rights…critical open source software and systems [can] protect the security and integrity of communications on the global Internet."[76]

Furthermore, the U.S. government has historically taken the view that certain types of tech/software qualify for protection under the First Amendment and should therefore not be subject to restrictions and made freely available.

[75] *Freedom on the Net 2023: The Repressive Power of Artificial Intelligence*, Freedom House (2023).
[76] Supporting Critical Open Source Technologies That Enable a Free and Open Internet, U.S. Department of State Funding Opportunity Announcement (Feb. 21, 2023).

**10. Recommendations**

In its report to the President on the potential benefits, risks, and implications of dual-use foundation models for which the model weights are widely available, as well as policy and regulatory recommendations pertaining to those models, we suggest NTIA recommend that the U.S. government:

1. Minimize restrictive limitations on open source AI so that open foundation models can continue to advance American leadership globally and deliver for the American people, including use throughout the federal government.
2. In considering approaches to AI governance that balance the benefits of open foundation models against cognizable marginal risks, the focus should be the responsible development of all AI models.
3. Adequately resource and fund the work of the U.S. government under the White House Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, in particular the U.S. AI Safety Institute.
4. Expedite the work of the U.S. AI Safety Institute and promote cooperation with counterpart safety institutes around the world through a comprehensive framework agreement (*e.g.,* a Memorandum of Understanding between AI Safety Institutes, or via existing mechanisms such as Cooperative Research and Development Agreements).
5. Support bipartisan AI legislation so that the United States can lead globally on responsible innovation for the technology and ensure consistent standards across the U.S. Such legislation should be focused on the development of domestic common standards informed by the White House Voluntary AI Commitments, the work of federal departments and agencies under the E.O., academia, industry, and standard development organizations.
6. Connect the U.S. government's existing work under the E.O. and the White House Voluntary AI Commitments to its leadership globally through ensuring that responsible open source is a central tenet of its foreign and trade policies, driving alignment with American governance frameworks, technical standards, and values in international fora.

7. Protect American industry and interests abroad from aggressive efforts in other jurisdictions to unfairly discriminate against American companies, restrict the ability for American companies to make their technology available in global markets, or force technology transfer, including mandatory government access to foundation model artifacts.