National Telecommunications and Information Administration
Response to Request for Comment:
Dual Use Foundation Artificial Intelligence
Models with Widely Available Model Weights
89 Fed. Reg. 14059 (Feb. 26, 2024)
Docket No. NTIA–2023–0009
March 27, 2024

Google welcomes the National Telecommunications and Information Administration's (NTIA's) request for comment (RFC) on widely available model weights for dual-use foundation models. NTIA's RFC is timely and important, and we appreciate the opportunity to contribute. We are one of only a handful of companies to have released both state-of-the-art closed artificial intelligence (AI) models (such as Gemini) and models with open weights (including our lightweight Gemma family), and we are pleased to share our learnings and perspective.

Executive Summary

Google has long been optimistic about the transformative potential of AI technology, which is already unlocking major benefits, from better understanding diseases to tackling climate change. We have also been prominent supporters of open science and open software as a way to foster innovation and competition. We have been one of the leading contributors of open-source code, including fundamental technologies like Kubernetes and the Go programming language. We've released projects such as Android and Chromium that transformed access to mobile and web technologies. And we have done the same in AI with Transformers, TensorFlow, AlphaFold, our new Gemma family of open models, and more.

For years, our AI Principles have guided us as we deploy AI responsibly. As a result, we release AI systems only when we determine that the benefits significantly outweigh the risks. We understand that models with open weights can pose unique risks, because it is impossible to reverse the decision to make weights available widely—and hence they deserve particular care. For example, we open-sourced AlphaFold after extensive consultations with bioethicists, and the tool is now used by 1.5 million biology researchers around the world. Likewise, the release of our Gemma family of open models was grounded in our thorough approach to safety and responsibility, including an assessment that the benefits of the release significantly exceeded the risks.

The challenge today is that there is not an agreed set of criteria that definitively answers whether an AI model should be open—or the degree to which developers should open it. Government, industry, and civil society have a key role to play in helping make progress on standards, evaluations, and best practices for releasing models with widely available weights. We offer the following recommendations for NTIA:

Recognize that the concepts of "open" and "closed" exist on a spectrum rather than as a binary set of choices. Good policy requires definitional clarity and rigor. Access to AI systems is better understood in terms of different degrees of access to different components of a given system, where the appropriate decision for release will require weighing potential benefits against potential risks. For example, Gemma offers free access to model weights, but under a custom license that requires adherence to Gemma's Prohibited Use Policy; in general, terms of use, redistribution, and variant ownership can differ due to licensing terms.

Promote a rigorous and holistic assessment of the technology to evaluate benefits and risks. A holistic risk assessment weighs the benefits of open release against the risks it may pose, including recognizing that many different components combine to make a model functional and useful and that varying the openness of these components will change the model's risk profile.

- Risk largely depends on model capabilities—for example, sophisticated models can present heightened risks due to their capacity to produce harmful content at scale.
- Other factors include the potential for performance and capabilities to change over time, including from post-training enhancements, and the availability of metadata, source code, compilers, extensions, and other fundamentals of integrated software systems.
- A model may pose novel risks or add substantially to existing risks in the AI ecosystem through its use at scale. Risks may also compound.

Taking any one attribute as determinative for a decision about the openness of a model will fail to capture the full risk landscape. We encourage NTIA to develop recommendations that appropriately support the release of open models by accounting for their attributes—and benefits and risks—holistically.

Highlight the critical need to calibrate testing and mitigations to the unique risks each open model presents. Developers should implement rigorous, tailored best practices, such as addressing safety and harm *prior* to deployment; implementing system internal reviews grounded in guiding principles; employing a high bar for evaluations; sharing and leveraging AI responsibility tools, such as Google's Generative AI Responsible Toolkit; and continuing to advance novel mitigations for open models. Where risk is present and misuse occurs, careful thought needs to be given to the allocation of responsibilities. Because upstream developers cannot verify the end uses to which their open models are put, the entity at the closest point to the product's end use should usually be held responsible. However, frameworks should be in place so that upstream developers of open models appropriately follow safety best practices. Given the risk of potential misuse, NTIA should urge developers to implement robust cybersecurity protections to ensure that they are making model weights available only when intended.

We are proud to be among a number of leading AI developers that have made voluntary commitments to heightened safety and security practices for our most advanced models. The US Executive Order on AI further identifies "dual-use foundation models" that could give rise to significant risks alongside their significant benefits. Responsible AI developers should exercise particular care in developing and deploying these models. Viewed in conjunction with the care that should guide decisions around open access to model weights, this suggests developers would have to meet a very high safety threshold for the release of weights for "frontier" models. We support a cautious approach to frontier models, not because existing models are inherently dangerous, but in anticipation of the rapid advancements we expect to see over the coming years.

Drive collaboration on standards for responsible open science. We need greater consensus in the AI community on many important questions, and NTIA will be vital to convening and supporting these discussions. Questions include: When is a model "too risky" to release openly? When is it "safe enough" to release? What metrics and evaluations do we need to develop to empirically underpin these assessments? And how do we drive toward decisions that are appropriately guided by the public interest?

Although there are signs of misuse in some current open models, we believe that, on balance, the benefits of many open models still significantly outweigh the risks. We also recognize a potential future class of models whose capabilities may necessitate a presumption against the open release of their weights, particularly because they may have "dangerous capabilities" such as the capacity to accelerate the development of hacking tools and methods. Risk calculations shift as one moves closer to models at the "frontier" of research and development, owing to the greater degree of uncertainty around their potential capabilities.

Although much of the conversation around the most advanced models focuses on the amount of training compute they consume, it is important to consider model capabilities independently from model size. AI capabilities continue to evolve and improve at an accelerating pace. Given this, we believe it is prudent to proactively consider potential future risks because the impacts of these emerging technologies are difficult to predict with certainty. Potential procedural guidelines could suggest that developers who are considering the release of weights for their most capable models would first need to deploy those models via more controlled modalities—such as APIs—for an extended period of time without major safety incidents.

There is a pressing need for more granular guidelines to shape these decisions, which will require robust threat models, state-of-the-art evaluations, risk thresholds, and red lines. These should be developed in coordination with governments and civil society—not by labs alone. Answers to these questions must be developed collectively, quickly, and globally, given the rapid evolution of technology. Google is proud to partner with a broad array of organizations like the Partnership on AI, MLCommons, and the Frontier Model Forum to foster best practices and standardized testing methods. Many organizations, including international standards organizations and newer entities like the National Institute of Standards and Technology's (NIST's) AI Safety Institute, can meaningfully advance these discussions.

Invest in open model research and other initiatives that support responsible access to AI capabilities. NTIA should support funding and policies that catalyze open model research, including the development of standardized evaluation frameworks, robust benchmarks, and additional risk mitigations, which will help address concerns about new uses and misuses that are undiscovered. Comprehensive evaluations are essential to understand model capabilities, including potential harms and post-deployment, and reduce uncertainty around release decisions. This should include funding for national initiatives such as the National AI Research Resource as well as independent research by organizations capable of conducting rigorous assessments.

These concepts should be central to NTIA's policies, as they will help advance today's and tomorrow's open technology innovation and unleash the benefits for all.

* * * * * * * *

Responses to NTIA Questions

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

Assuming "openness" in AI as a binary choice between open and closed approaches fails to capture important nuances. Rather, it is important to think about "open" and "widely available"[1] as existing on a gradient of *access,* which offers a better conceptual frame. There are many components of an AI system that can be made openly available in addition to weights, including model architecture metadata, training data, training code, and documentation—each releasable via different modalities and enabling any number of downstream integrations when shared. A comprehensive risk assessment should examine the specific implications of sharing each component rather than treating a model as a monolithic entity. Viewing openness as a pure binary conflates many different types of model release approaches together, when those differences have meaningful technical and policy implications. Accordingly, as some researchers propose, we suggest thinking of openness at different levels, such as fully closed, gradual or staged access, hosted access, cloud-based or API access, downloadable access, and fully open.[2] At a high level, access levels can be described as follows:

Fully closed models and all of their components (e.g., architecture, weights, and training data) are kept private by the developer and are inaccessible to external parties. These models can only be accessed by the developer. Examples include Google DeepMind's Gopher.[3]

API access allows controlled usage of the model but limits direct access to model artifacts. Here, the model is hosted on the developer's servers and can be accessed by external users via an API. Users can provide inputs and receive outputs from the model via a defined interface, but cannot directly inspect or modify the model's internal architecture, weights, hyperparameters, or training data, which remain under the control of the API provider. The API may, however, provide varying levels of functionality, from simple queries to more advanced features like fine-tuning; it may also provide certain controls, such as rate limits and content filters, which can be updated over time in response to developments such as greater understanding of model capabilities. Examples include Google's Gemini Ultra 1.0, offered via Google Vertex AI.[4]

Restricted weights access limits a trained model's weights to selected external researchers and/or developers, usually under certain license terms and usage restrictions. The full model code may or may not be provided. Access is gated, and recipients are vetted. This allows for some analysis by external parties and the building of a broader range of applications. Meta's initial release of its OPT models followed this approach.[5]

[1] From our point of view, there is no meaningful distinction between "open" and "widely available" as NTIA has presented the terms, so we use them interchangeably.

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

[3] Google DeepMind, *Language modelling at scale: Gopher, ethical considerations, and retrieval* (Dec. 8, 2021).

[4] Google Cloud, *Google Cloud expands access to Gemini models for Vertex AI customers* (Feb. 15, 2024).

[5] Meta, *Democratizing access to large-scale language models with OPT-175B* (May 3, 2022).

Full weights access permits anyone to download a trained model's weights, but some or all of the model code (e.g., training code and tokenizers) is withheld. Users can run and fine-tune the model but must abide by the license terms. Examples include Meta's Llama 2 model and Google's Gemma models.[6] Once a model's full weights are publicly released, any safety measures or content filters implemented by the original developer can potentially be removed or circumvented by third parties, and this release is usually irreversible since the weights can be easily copied and distributed beyond the original developer's control.

Fully open models involve public availability of a fully pre-trained model and additional components (architecture, weights, code, and sometimes training data), typically on a platform such as GitHub or Hugging Face. Anyone can download, use, modify, and share the model without restrictions. Examples include EleutherAI's GPT-Neo and GPT-J and Stability AI's Stable LM.[7]

As Google currently uses the term, "open" models feature free access to the model weights, but terms of use, redistribution, and variant ownership will differ according to a model's specific licensing terms, which may not be based on an open-source license.[8] For example, the Gemma models' terms of use make them freely available for individual developers, researchers, and commercial users for access and redistribution.[9] Gemma users are also free to create and publish model variants, but under Gemma's custom license, developers agree to avoid harmful uses, reflecting our commitment to developing AI responsibly while increasing access. At the same time, we acknowledge that AI systems are more complex than many other software systems and that work is ongoing to align on a shared definition of "open source AI" that reflects these complexities.

Google strongly supports openness in technology because it both spurs innovation and builds trust. We are proud to offer many products and services that empower developers and researchers. At the same time, "open" or "widely available" foundation model weights will often still need to incorporate measures that protect against risks associated with public access—such as terms that outline certain prohibited harmful uses. Thus, "open" or "widely available" should not always mean that a model is made available with completely unrestricted terms.

Ultimately, the appropriate level of access will differ based on various factors, but what is key is ensuring that appropriate AI capabilities are accessible to appropriately provisioned sets of developers (i.e., developers committed to responsible use). Release modalities, such as staged releases, play an important role in enabling such access. Similarly, structured access programs enable controlled, arm's-length interactions with AI models that are not open or widely

[6] Meta, *Llama*; Google: The Keyword, *Gemma: Introducing new state-of-the-art open models* (Feb. 21, 2024).
[7] EleutherAI, *GPT-Neo*; EleutherAI, *GPT-J*; Stability AI, *Stability AI Launches the First of its Stable LM Suite of Language Models* (Apr. 19, 2023).
[8] We see value in clarity around language—including indicating when a model is "open," and when it is "open source." The Open Source Initiative's definition of "open source" has historically offered useful principles: software under this definition must permit redistribution, allow derived works, and prohibit restrictions on use. These principles cannot always be directly applied to AI systems, which raise specific nuances related to concepts like derived work and author attribution.
[9] *See* Google AI for Developers, *Gemma Terms of Use*.

available but that may present risk. These programs prevent model weights from becoming widely accessible while preserving access to the capabilities that can be used safely.[10]

2-3. What are the benefits of foundation models with model weights that are widely available as compared to fully closed models? How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?[11]

We believe that an open approach to technology can bring significant benefits, but also needs to control for key risks. Google has pioneered a wide range of open projects that have expanded choice and opportunity for consumers and business users, spurring innovation and competition. For example, our free open-source browser software—Chromium—powers not only Google's Chrome browser, but also competing browsers such as Microsoft's Edge, Amazon's Silk, DuckDuckGo's Privacy Browser, and Opera. Our free open-source mobile operating system, Android, underpins 24,000 models of smartphones and laptops, including some that use only non-Google apps, increasing choice and lowering cost for consumers.

Google likewise freely shares many of its AI breakthroughs, helping advance the state of technology for all. Our open-source machine learning (ML) platform TensorFlow makes cutting-edge AI capabilities publicly available. We offer free open datasets (images,[12] videos,[13] natural questions[14]) to foster research, and publish more than 1,000 research papers annually, sharing ideas to advance science.[15] In fact, the transformer technology—the "T" in ChatGPT—is what enables generative AI and was developed by Google researchers and made freely available in 2017.[16] We have also opened up other AI innovations such as GraphCast, Word2Vec, BERT, T5, JAX, AlphaFold, and AlphaCode.[17] Our open approach allows a broad ecosystem of

[10] Toby Shevlane, *Structured Access: An Emerging Paradigm for Safe AI Deployment*, arXiv (Apr. 11, 2022); Toby Shevlane, *Sharing Powerful AI Models*, Centre for the Governance of AI (Jan. 20, 2022).

[11] The responses here are intended to respond to both Questions 2 and 3 of the RFC.

[12] Google Research, *Announcing Open Images V5 and the ICCV 2019 Open Images Challenge* (May 8, 2019).

[13] Google Research, *YouTube-8M Segments Dataset*.

[14] Google Research, *Natural Questions: a Benchmark for Question Answering Research* (2019).

[15] *See* Google Research, *Publications*.

[16] Transformer models provide a novel neural network architecture based on a self-attention mechanism that we believe to be particularly well suited for language understanding. Google Research, *Transformer: A Novel Neural Network Architecture for Language Understanding* (Aug. 31, 2017).

[17] Our state-of-the-art weather model delivers 10-day weather predictions at unprecedented accuracy in under one minute. Google DeepMind, *GraphCast: AI model for faster and more accurate global weather forecasting* (Nov. 14, 2023). Word2Vec is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets, supporting a variety of downstream natural language processing tasks. TensorFlow, *Word2Vec*. Bidirectional Encoder Representations from Transformers, or BERT, allows anyone to train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU. Google Research, *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing* (Nov. 2, 2018). With T5, we propose reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input. Google Research, *Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer* (Feb. 24, 2020). JAX is Autograd and XLA, brought together for high-performance numerical computing, including large-scale ML research; it can automatically

users to benefit from our state-of-the-art technology even as they are able to access a range of alternatives.

Many of these same benefits to innovation, research, and competition apply to open models. Open models allow users across the world, including in emerging markets, to experiment and develop new applications, lowering barriers to entry and making it easier for organizations of all sizes to compete and innovate. These were the benefits we had in mind when we decided to release our Gemma family of lightweight models. To optimally support a wide range of developer needs and environments—and to ease the economic and technical barriers we know many developers face when integrating these emergent systems into their workflows—we released Gemma model weights in two sizes capable of running directly on a developer laptop or desktop computer.[18] We have already seen innovative uses of these models that further increase access to AI capabilities for more communities: for example, third-party developers have recently used and enhanced Gemma's generative capabilities to encompass a total of 15 Indian languages.[19]

Openly available models also enable important AI safety research and community innovation. A diverse pool of available models ensures that developers can continue to advance critical transparency and interpretability evaluations from which the developer community has already benefited. For example, researchers have demonstrated a method for reducing gender bias in BERT embeddings.[20] Open models have also been used to advance safety research, including on interpretable internal representations and adversarial attacks.[21] In addition, releasing open models can provide information, like user feedback, that can help guide future open releases.

differentiate native Python and NumPy functions. GitHub, *JAX: Autograd and XLA*. In a major scientific advance, the latest version of our AI system AlphaFold provides a solution to the so-called "protein folding problem," the mystery of what shapes proteins fold into. Google DeepMind, *AlphaFold: a solution to a 50-year-old grand challenge in biology* (Nov. 30, 2020). AlphaCode uses transformer-based language models to generate code at an unprecedented scale, and then smartly filters to a small set of promising programs. Google DeepMind, *Competitive programming with AlphaCode* (Dec. 8, 2022). Our paper detailing this breakthrough was published on the cover of Science. *See* Yujia Li et al., *Competition-level code generation with AlphaCode*, Science (Dec. 8, 2022).

[18] Gemma's two sizes are 2 billion and 7 billion parameters. We made them available for hosting across a number of platforms including locally on-device (i.e., deployable on laptop, desktop, IoT and mobile) as well as across multiple cloud environments. We also provide access to base pre-trained models alongside instruction-tuned offerings intended to encourage a range of developers to leverage Gemma's chat and code capabilities to support their own applications. "Pre-training" is "a process where a model is first trained on a large, general dataset before being fine-tuned on a specific task," while "instruction tuning" is "the process of fine-tuning a machine learning model based on specific instructions or prompts." *See* TED AI San Francisco, *Glossary*.

[19] Ravi Theja, *Introducing Navarasa 2.0 — Indic Gemma 7B/2B Instruction tuned model on 15 Indian Languages*, Medium (Mar. 18, 2024).

[20] Nora Belrose et al., *LEACE: Perfect linear concept erasure in closed form*, arXiv (Oct. 29, 2023).

[21] Andy Zou et al., *Representation Engineering: A Top-Down Approach to AI Transparency*, arXiv (Oct. 10, 2023); Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, arXiv (Dec. 20, 2023).

While the benefits of open AI models are profound, there is also a risk that their use accelerates harms, like deepfake imagery, disinformation, and malicious services.[22] Below, we raise six elements that raise the risk profile of a model whose weights are openly released compared to one whose weights are not openly released:

First, once model weights become publicly available, it is difficult or impossible to limit or revoke access afterward. Sharing and downloading model weights is a relatively uncomplicated exercise. While hosting platforms may choose to delist or otherwise restrict developer access to a model's weights or applicable components, that would not prevent other distribution platforms from hosting and sharing the same content.

Second, it is difficult to prevent bad actors from fine-tuning an open model for malicious intent, even when access to the model is subject to a prohibited use policy.[23] Further work is needed to build more robust mitigation strategies against intentional misuse of open systems, which we are exploring both internally and in collaboration with the wider AI community.

Third, open-source projects can be compromised in ways that introduce risks for downstream organizations that integrate open utilities into their systems. This could occur, for example, if a model is designed to execute malicious code on a host's computer that is then made available for users to download.[24] With an open model, certain vulnerabilities and risks can propagate downstream to fine-tuned models, whereas with a closed mode (i.e., one behind an API), known vulnerabilities can be more easily identified and patched.[25] This risk is also applicable to the open source software (OSS) landscape, and lessons learned there may be helpful in this context. The OSS community has developed mitigations against these risks and is currently evaluating these issues as they relate to models in the Open Source Security Foundation (OpenSSF), an industry endeavor to support secure OSS.

Fourth, systems access that is managed via a controlled modality, such as an API, allows organizations deploying models to apply a number of centralized mitigations for unintended model behaviors, such as model hallucinations or leakage of personally identifiable information. These can include instruction and supervised fine-tuning nodes to ensure adherence to safety policies and thresholds and the use of classifiers to restrict specific types of model outputs. Similar mitigations are more difficult to coordinate and achieve at scale for open models, which tend to become distributed and used broadly on many different platforms and devices.

Fifth, some models may exhibit emergent capabilities that may not be fully understood at earlier phases of public deployment, but that arise once they are made widely available. There

[22] Zilong Lin et al., *Malla: Demystifying Real-world Large Language Model Integrated Malicious Services* (Jan. 6, 2024); Richard Fang, *LLM Agents can Autonomously Hack Websites* (Feb. 16, 2024).

[23] This risk is not unique to open models, as research indicates that bad actors can jailbreak closed models as well. *See* Zilong Lin et al., *Malla: Demystifying Real-world Large Language Model Integrated Malicious Services* (Jan. 6, 2024). That being said, an open model raises the risk of a bad actor modifying the model for malicious intent.

[24] The response to Question 7 discusses methods by which the integrity of open source projects can be maintained.

[25] Huaming Chen & M. Ali Babar, *Security for Machine Learning-based Software Systems: a survey of threats, practices and challenges*, arXiv (Dec. 17, 2023).

can be a related "time lag" effect: open models may be erroneously deployed under a false premise of limited marginal risk. For open models, which may lack mechanisms for centralized risk mitigation, a more robust initial risk assessment may be needed. Mitigations could include longer observation and evaluation periods for new model classes, followed by staged deployment phases.

Sixth, individual developers could unintentionally contribute to an overall increase in societal risk if they neglect to account for the cumulative effect of deployments. For example, interactions among multiple open-weight models (e.g., where the components of multiple models are "chained" together) could lead to an unforeseen rise in total societal risk due to inherent difficulties of system-wide risk measurement.

Despite these challenges, we expect that many open models will be a vital part of the AI ecosystem. But maximizing the benefits while minimizing the risks will require responsible frameworks for open models. Our approach to the Gemma family provides one such framework. Our decision to release these models openly involved a holistic assessment of their benefits and risks. In particular, we focused on rigorous testing, including through adversarial fine-tuning, to understand the "upper bounds" of model behavior so we could assess whether these models might pose any novel or marginal risks through use at scale, compared to existing software and open AI tools. In general, we support a measured and cautious approach that acknowledges uncertainties about capabilities due to fine-tuning, interactions among systems, and other learning effects, as well as limitations in current evaluation and measurement methods. We also tailored the degree of openness of our release—including the use of a custom license—to help mitigate potential harmful uses. We discuss this approach in more detail below.

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.

Risk-benefit determinations for open foundation models are necessarily multifaceted due to the complexity of the underlying technological components and variations in the environments in which they are deployed. Policymakers must analyze the risks and benefits of open foundation models holistically, noting how risks compound but weighing those against a model's potential benefits and the available mitigations.

An open model's risks depend largely on its capabilities.[26] For open models with limited capabilities, the marginal risk of harm from their release may be negligible, and their deployment may offer useful lessons on how to approach releasing more capable models. However, very sophisticated models can present heightened risks due to novel capabilities, deeper reasoning, and their capacity to produce harmful content at scale. Multimodal models can also pose additional risks and require additional scrutiny, given the salience of visual information. For future agentic systems that could make decisions with limited direct

[26] Model capability may be a more important indicator than model availability when assessing model risk. For example, biorisk is a concern for AI safety and security, necessitating mitigations on a wider societal level, such as DNA synthesis screening. *See* Steph Batalis & Vikram Venkatram, *Breaking Down the Biden AI EO: Screening DNA Synthesis and Biorisk*, Center for Security and Emerging Technology (Nov. 16, 2023).

supervision, other baseline safety practices may be merited.[27] However, a model's size is not a perfect proxy for the degree of risk it might pose; for example, small specialized models may pose higher risks in particular domains than large, general-purpose models do. While we recognize limitations in the current state of the art, it will be essential for the AI community to rapidly develop more precise methods to evaluate potential capabilities instead of relying on proxy measures.

As the AI community builds and deploys increasingly powerful AI, NTIA should also promote policies and best practices that guard against extreme risks from future general-purpose models that could achieve strong skills in domains such as manipulation, deception, cyber-offense, or self-replication.[28] While the capabilities of current open models do not suggest an immediate potential for material risks in this respect, we believe strongly in the value of developing and applying evaluations that can serve as "early warning systems," should those capabilities arise in the future. By extension, we believe there is a class of future models whose capability profile should lead to a strong presumption against making their weights openly available.

Policymakers should also consider that an AI model's performance and capabilities are dynamic, not static. Post-training enhancements, such as scaffolding,[29] better prompting and inference mechanisms,[30] or fine-tuning a model to access the internet, can significantly improve performance.[31] This means that the risks and benefits of a model assessed today may be substantially different for the same model assessed a year from now.

The availability (or lack) of metadata, source code, compilers, extensions, and other fundamentals of integrated software systems are also important factors to consider when determining how broadly to share or open a given system. On the one hand, access to these resources can increase the benefits of widely available model weights because these components enable developers to innovate with the benefit of a fuller context of a system's inputs. In this spirit, Google provides access to resources such as data sets and Kubernetes, an open-source system to deploy, scale, and manage containerized applications anywhere.[32] At the same time, access to these resources can introduce risks, for example, by enabling model

[27] *See* Yonadav Shavit, *Practices for Governing Agentic AI Systems*, OpenAI (Dec. 14, 2023).

[28] *See* Google DeepMind, *An early warning system for novel AI risks* (May 25, 2023).

[29] Tom Davidson et al., *AI capabilities can be significantly improved without expensive retraining*, arXiv (Dec. 12, 2023) ("Scaffolding enhancements structure the model's thinking and the flow of information between different instances of the model, allowing the resultant system to tackle a wider array of problems.").

[30] Maciej Besta et al., *Graph of Thoughts: Solving Elaborate Problems with Large Language Models*, arXiv (Feb. 6, 2024).

[31] Tom Davidson et al., *AI capabilities can be significantly improved without expensive retraining*, arXiv (Dec. 12, 2023) ("Our non-experimental work shows that post-training enhancements have significant benefits: most surveyed enhancements improve benchmark performance by more than a 5x increase in training compute, some by more than 20x. Post-training enhancements are relatively cheap to develop: fine-tuning costs are typically <1% of the original training cost. Governing the development of capable post-training enhancements may be challenging because frontier models could be enhanced by a wide range of actors.").

[32] Google Cloud, *Datasets*; Google Cloud, *What is Kubernetes?*.

weights to be used in new and harmful ways. In those circumstances, sharing these resources would not be appropriate.

Taken together, the above points stress how risks compound and interact in the context of open systems development. Considering a model in isolation is not sufficient. It is more useful to identify the specific threat types in the context of the overarching environment in which the system will operate, and to compare the impact of an AI system to other widely available tools. For example, assessing the risk associated with strong cyber offense capabilities will partly depend on the strength of the external environment's cybersecurity, while the capabilities of models in the biological and chemical weapons domains should be compared against alternative sources for such information, such as reference books and web search. Thus, we advise thinking about the potential for systemic risks rather than the model alone.

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 model weights?

Practices need to evolve to ensure strong safety and security standards for components of all AI models, including open models. Just as we have created unique threat models and solutions for other open technologies, we are developing safety and security tools appropriate for the attributes of openly available AI.[33] For example, Gemma models incorporate state-of-the-art safety features and tools, including a novel methodology for building robust safety classifiers with minimal examples—giving developers a head start as they innovate for safety.[34] In the broader context of community-driven software development, the OpenSSF offers a model of collaborative safety innovation. Below are some broader suggestions that stakeholders should keep in mind.

Securing model weights against unauthorized release

Keeping the most advanced AI models and systems secure is a cornerstone of responsible AI model and systems development, and portions of this approach should be extended to the development of open models too. As frontier models become more powerful, we expect to see increased attempts to disrupt, degrade, deceive, and steal them. Given the potential risks that could result from the use of certain model weights, it is critical that they only be distributed as a result of a deliberate and rigorous process, and that they are protected against unauthorized release. Many of the mitigations available or contemplated for safe open model release—including rigorous evaluation prior to release or staged access—are dependent on robust protection of model weights from the beginning of the model development process.

Our models are developed, trained, and stored within Google's infrastructure, supported by central security teams and by a security, safety, and reliability organization consisting of engineers and researchers with world-class expertise. Google's Secure AI Framework (SAIF) provides a number of practical recommendations for organizations looking to integrate security best practices into their AI systems, including: (1) expanding strong security foundations to the AI ecosystem, including secure-by-default protections; (2) extending

[33] Google Security Blog, *Celebrating SLSA v1.0: securing the software supply chain for everyone* (Apr. 26, 2023).

[34] *See* Google AI for Developers, *Responsible Generative AI Toolkit*; Google AI for Developers, *Tune models for safety*; Google Codelabs, *Showcasing Agile Safety Classifiers with Gemma* (Feb. 21, 2024).

detection and response to bring AI into an organization's threat universe; (3) automating defenses to keep pace with new and existing threats; (4) harmonizing platform-level controls to ensure consistent security across organizations; (5) adapting controls to adjust mitigations and create faster feedback loops for AI deployment, and; (6) contextualizing AI system risks in surrounding business processes.[35] The SAIF is inspired by the security best practices—like reviewing, testing, and controlling the supply chain—that we've applied to software development while incorporating our understanding of security mega-trends and risks specific to AI systems.[36]

We are also in the process of developing a framework to ensure heightened cybersecurity for our cutting-edge frontier models. This framework assigns initial risk categories based on projected performance, monitors performance during training, and applies appropriate mitigations post-training. It governs the access, use, and distribution of the model and its weights, with an expert committee tasked with making informed decisions about adjusting categorizations and relaxing mitigations when appropriate.

NTIA should advocate for the adoption of such conceptual frameworks that both government and the private sector could use to collaboratively secure AI technology. Additionally, NTIA should encourage model developers to incorporate NIST's Secure Software Development Framework, which assists software developers in decreasing vulnerabilities in software releases, minimizing the potential consequences arising from the exploitation of undetected or unresolved vulnerabilities.[37]

Measures for open models

Where developers are considering making their model weights openly available, it's important to follow rigorous practices tailored to the attributes of open models to protect against potential risks. Based on our own experience with open model releases, we believe the following principles should be top priorities for all ecosystem stakeholders:

Address safety and harm prior to deployment – It is essential to protect developers and downstream users against the unintended behaviors of open models, including the generation of toxic language or the perpetuation of discriminatory social harms, model hallucinations, and leakage of personally identifiable information. When deploying models behind an API, these risks can be reduced via various filtering and fine-tuning methods. We safeguard against these risks by implementing robust data governance practices on our pre-training data and assessing our models against standardized AI safety benchmarks,[38] although we also recognize that mitigations related to unintended behaviors for open models are vulnerable to adversarial fine-tuning and other methods.

[35] Google Safety Center, *Google's Secure AI Framework (SAIF)*.

[36] For example, Google's cybersecurity architecture blocks over 100 million phishing attempts every day and checks over 1 billion saved passwords for breaches, while also protecting 4 billion devices against risky websites. Google, Our cyber security journey through the years.

[37] NIST, *Secure Software Development Framework*.

[38] Google DeepMind, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* (2024).

Systematic internal review in accordance with guiding principles – We believe it is essential that AI developers approach the model development and deployment process—including decisions to release models openly—using a principled framework for making decisions that prioritizes safety and security. For example, Google has adopted seven AI principles[39] and has a central team dedicated to ethical reviews of new AI and advanced technologies before launch. We work with internal domain experts in machine-learning fairness, security, privacy, human rights, the social sciences, and, for cultural context, Google's employee resource groups.[40] We release models only when we have determined that the benefits are significant and the risks of misuse are low or can be mitigated. We take that same approach to open models. With Gemma, we considered the possibilities for increased AI research and innovation by us and many others in the community, the access to AI technology that the models could bring, and what access was needed to support these use cases. We also employed a measured and cautious approach to the release decision—in essence, holding open models to a higher bar for release—by factoring in uncertainties about the measurement of model capabilities and the irreversibility of making model weights available openly.

Use a high evaluation bar – Our open models underwent thorough evaluations, consistent with our AI Principles, and were held to a higher bar for evaluating risk of abuse or harm than our proprietary models, given the more limited set of post-deployment mitigations currently available for open models. For example, Gemma was subjected to assessments based on standardized AI safety benchmarks, internal red-teaming, and rigorous ethics and safety evaluations (including for fairness, privacy, and societal risk). We also conducted dangerous capability evaluations for chemical, biological, radiological, and nuclear (CBRN) risks, cybersecurity, and autonomous replication—and specifically assessed the models' potential for developing harmful abilities. These evaluations involve fine-tuning a base model to maximize its propensity to perform undesirable actions, with the "upper bound" assessment allowing us to measure on a continuous scale how close the model is to acquiring a dangerous capability. These are the same types of evaluations that we have applied on our most powerful general models that are covered by our July 2023 Voluntary Commitments to the White House.[41] While these evaluations did not produce any results of concern for Gemma, it is important to evaluate for such capabilities as models become more powerful. This is even more critical when models are open because bad actors may remove safety mechanisms through fine-tuning. We encourage NTIA to work with the AI community to develop more rigorous criteria for evaluations in areas such as dangerous capabilities and fine-tuning to reduce the uncertainty about the capabilities of AI models, which would aid in decisionmaking around open models.

Share and leverage AI responsibility tools – A critical mitigation tool for open releases is to make it easier for model users to deploy AI tools responsibly. To that end, we have released a Generative AI Responsible Toolkit to support developers to build AI responsibly, including for Gemma models.[42] The Toolkit includes resources to help developers design and implement

[39] Sundar Pichai, *AI at Google: our principles*, Google: The Keyword (June 7, 2018).

[40] Google DeepMind, *AI Safety Summit: An update on our approach to safety and responsibility* (Oct. 27, 2023).

[41] White House, *Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (July 21, 2023).

[42] Google AI for Developers, *Responsible Generative AI Toolkit*.

responsible AI best practices and keep their own users safe. While we've invested significantly in the model toolkit, we recognize its limitations. To ensure transparency for downstream users, we've also published a detailed model card to provide researchers with a more comprehensive understanding of the model.[43] Transparency reporting should include an analysis and description of risk-benefit and why the open model release doesn't create novel or marginal/differential risks.

Make progress on novel mitigations – As an industry, we should aim to identify additional mechanisms that could help us mitigate the risks of open models. One area of opportunity is to develop a better empirical understanding of the spread and uses of open models. Alongside more systematic market monitoring, technical solutions also hold promise. For instance, researchers have identified techniques to establish provenance that could be embedded within open models.[44] While many open questions remain about how such mitigations could be implemented at scale, governments can support technical research on these issues as well as a common approach across open model developers.

In sum, open model release best practices should follow the essential elements of this framework. Developers should implement robust cybersecurity protections for model weights to ensure that models are only made widely available when intended. Developers should also put in place clear internal principles and guidelines that detail when they will deploy models openly that prioritize safety and security. Developers should build as thorough an understanding of model capabilities as the current state of the art allows, while also acknowledging limitations in evaluation science, model modifications, and learning effects that lead to underestimation of potential capabilities—especially for more powerful models. Given the irreversibility of open weight releases, this should lead to a measured and cautious approach to assessing the risk-benefit balance for open models. Releases can be paired with tools such as model cards and safety toolkits that can help reduce misuse. And finally, it is essential to conduct research on improved evaluations and novel mitigations so that open model releases can keep pace with the rapidly advancing state of AI science.

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

Open foundation models are associated with several important legal and business issues, a few of which we highlight here.

First, when companies make their models open, it increases the risk that intellectual property will be leaked. Making more of the underlying source code and training data publicly accessible increases the risk that business-sensitive information could be disclosed. As research shows, data can be easily extracted from open models, and even closed models are vulnerable.[45] There may also be a risk of exfiltration for proprietary models that run on-device.

Second, there is also the risk that models will be abused or misused.[46] Although companies can prohibit harmful uses, once a model is shared, companies relinquish much of this control.

[43] *Gemma*, Kaggle.
[44] Jiashu Xu, *Instructional Fingerprinting of Large Language Models*, arXiv (Jan. 21, 2024).
[45] Milad Nasr et al., *Scalable Extraction of Training Data from (Production) Language Models*, arXiv (Nov. 28, 2023).
[46] *See* Response to Questions 2 and 3, *supra*, for description of some risks and harms.

With open-source applications (as with other technologies), companies are not legally responsible for third-party misuse. For this reason, NTIA should explore solutions that acknowledge a shared responsibility for safety by model developers, deployers, and users but also recognize that updated liability frameworks may be useful to fully realizing the benefits of open models, given that the entity at the closest point to the AI product end-user is best positioned to monitor and prevent misuse. Clarifying this point for open models can help drive continued investment.

Third, open foundation models can expose companies to the risk of reputational harm. If their open foundation models are abused or misused, companies may face repercussions in the court of public opinion.

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

Google teams are collaborating across a number of cross-industry, policymaker, and civil society dialogues to help build durable frameworks that allow us to realize the manifold benefits of AI while mitigating the risks associated with the adoption of any general-purpose technology.[47] As our CEO has said, "AI is too important not to regulate, and too important not to regulate well,"[48] which is why we are engaging deeply with government organizations and leaders in the open-source community to build out industry-wide best practices and standardized testing methods, including:

- The Partnership on AI, a community of experts from academic, civil society, industry, and media organizations dedicated to fostering responsible practices in the development, creation, and sharing of AI.[49]
- MLCommons, a collective that aims to accelerate ML innovation and increase its positive impact on society.[50] We supported MLCommons' proposal to utilize a multistakeholder process for selecting tests and grouping them into subsets to measure safety for particular AI use cases, and we are supporting the recently launched MLCommons Working Group to develop and update standard safety benchmarks.
- The Frontier Model Forum, an industry body focused on safe and responsible AI, draws on the technical and operational expertise of its member companies to benefit the entire AI ecosystem, advancing best practices, technical evaluations and benchmarks, and solutions to common issues.[51]

[47] Guy Ben-Ishai et al., *AI and the Opportunity for Shared Prosperity: Lessons from the History of Technology and the Economy*, arXiv (Feb. 1, 2024).
[48] Sundar Pichai, *Google CEO: Building AI responsibly is the only race that really matters*, Financial Times (May 23, 2023).
[49] Google: The Keyword, *How we're partnering with the industry, governments and civil society to advance AI* (Feb. 14, 2024).
[50] *Id.*; Google Research, *Supporting benchmarks for AI safety with MLCommons* (Oct. 26, 2023).
[51] Google: The Keyword, *Frontier Model Forum: A new partnership to promote responsible AI* (July 26, 2023).

- The Coalition for Content Provenance and Authenticity, a cross-industry effort to provide more transparency and context for people on digital content.[52]
- The Global Partnership on AI, a multistakeholder initiative established by the G7 that aims to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities.
- We are collaborating with relevant government organizations to advance AI safety, including the UK's AI Safety Institute and the US AI Safety Institute Consortium.

NTIA should encourage open model developers and deployers to participate in these consensus-building forums as well as continue engaging directly itself to support research and best practices related to open models. Alignment here can also complement technical standards development efforts (such as those being developed by ISO) by, for example, providing canonical datasets, evaluation methods, and benchmarks for evaluating open AI systems.[53] These efforts will further a collective understanding of more precise capability thresholds indicating when models may be too risky to release openly.

Google also supports the Supply-chain Levels for Software Artifacts (SLSA) standard aimed at harmonized evaluations and descriptions of how secure software was built. SLSA's security framework provides adoptable guidelines to improve supply chain security and offers a common vocabulary to discuss software supply chain security.[54] With software systems attacks being responsible for damage to both public and private interests,[55] NTIA should encourage other industry members to implement similar mechanisms.

Applying SLSA to open model design could provide similar information about a system's supply chain and address attack vectors not covered by model signing, such as a compromised source control, a compromised training process, and vulnerability injection. Our vision is to include specific ML information in a SLSA provenance file, which would help users spot an undertrained model or one trained on bad data.[56] By promoting SLSA as a national standard for software provenance, policymakers can raise the bar for supply chain security standards. Upon detecting a vulnerability in an ML framework, users can quickly identify which models need to be retrained, thus reducing costs. Open-source maintainers should not have to shoulder the burden of adopting SLSA; it should be part of the default fabric of the software ecosystems they already use for development.

[52] C2PA, *Google to join C2PA to help increase transparency around digital content* (Feb. 8, 2024).

[53] *See* ISO, *Standards by ISO/IEC JTC 1/SC 42: Artificial Intelligence*.

[54] Google Cloud, *Securing the software development lifecycle with Cloud Build and SLSA* (July 29, 2021). Over the last decade, Google has used an internal version of SLSA to protect against insider risk, build system tampering, and unilateral code changes.

[55] *Id.*

[56] Integrating SLSA tooling into core ecosystem tooling allows for both the generation of signed provenance at the time of production and the verification of that provenance at the time of consumption, by default. SLSA support needs to be built into open-source ecosystems, including SLSA-compliant builders, tooling for provenance production, and automatic verification solutions.

8. In the face of continually changing technology, and given 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?

We need more consensus in the AI community regarding a number of important questions. Governments have an important role to play in convening and supporting these discussions and in facilitating ongoing public-private collaboration and research. Labs and relevant domain experts should consult with governments to establish thresholds on the potential risks of open models, in order to inform release decisions—including thresholds beyond which the model weights should not be openly shared.

Establishing such thresholds will require making urgent progress on model evaluations and mitigation mechanisms and on aligning on a common approach across industry. We need to develop evaluations on multiple fronts, with some directed at certain types of risks (e.g., CBRN risks) or specific domains in which AI systems may operate (e.g., health and finance). The AI Safety Institutes have a particular opportunity to convene expert views and establish standards for evaluations.

The US government has a crucial role in defining acceptable thresholds in specific domains, particularly in high-consequence domains like those related to CBRN threats. Prior to setting such thresholds, rigorous threat modeling is necessary to identify the specific AI capabilities that could pose risks if developed without sufficient safeguards. Once these concrete risks are defined, policymakers, AI developers, and other parties could set clear capability thresholds that would trigger additional testing requirements, enhanced cybersecurity measures, or other model-specific interventions needed to mitigate the identified threats. At the frontier of AI development, broad capability-based thresholds may be appropriate, potentially supplemented by narrow compute-based assessments for frontier models.[57]

Support and investment from government bodies and industry leaders for other initiatives on responsible access to AI capabilities are also needed. For example, Google contributes resources to the National Science Foundation's National AI Research Resource Pilot.[58] We further suggest establishing a Global Resource for AI Research (akin to the National AI Research Resource, but on a bigger scale) and strong trade and investment policies that allow for international collaboration on AI, including cross-border data flows that will enhance the capability of partners to work together to make sure AI systems are trained on demographically and geographically representative datasets.[59]

[57] The Commerce Department should be cautious about imposing restrictions on the release of models before the government, civil society, and industry develop more precise definitions on which to base those decisions. Overly restricting the release of the models based on limited factors, such as compute power rather than capabilities, could limit the innovations and benefits we have seen countless times from making technology available openly. While recognizing the risks associated with releasing open models, the developers of frontier models should be expected to weigh the risks and benefits prior to the release of models rather than hewing to a particular metric that may or may not adequately capture the risks and benefits.

[58] Google: The Keyword, *How we're partnering with the industry, governments and civil society to advance AI* (Feb. 14, 2024).

[59] Kent Walker, *An opportunity agenda for AI*, Google: The Keyword (Nov. 14, 2023).

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?

NTIA should focus on whether open foundation models present *heightened* risks as it develops policies in this space. This analysis would consider the extent to which open models empower bad actors in ways that existing technology does not. For example, NTIA should not base policy decisions on the risk that bad actors might use a large language model as a search engine because bad actors have had access to search engines for decades. By focusing on areas in which open models could create novel risk or exacerbate risk, NTIA can better target its policies to protect the public while supporting innovation and not impeding the benefits of openness discussed in this comment.

NTIA should also explore policies that place liability at the closest point to the end-use of an applicable AI product, especially for open model releases where upstream developers have no operational control over the final uses a model may be put to. Downstream parties who are directly servicing users are most familiar with how a model is being used and the specific context. They are best positioned to ensure safety and consider relevant mitigations. While providers of off-the-shelf, multipurpose AI systems can provide general information about their construction and guidance on operating boundaries in foreseen use cases, they are poorly positioned to conduct a deployment risk assessment because they cannot verify the end-uses to which their systems are put. This approach could supplement other policies designed to deter misuse of models by any actor (developer, deployer, or user).

Additionally, NTIA should track the potential impact of AI agents, especially how they will affect risk profiles. These capabilities are rapidly expanding.[60] Google recently shared new research on its Scalable Instructable Multiworld Agent that can follow natural-language instructions to carry out tasks in a variety of video game settings.[61]

Last, NTIA should recognize the critical need to sustain OSS communities. OSS and related services underpin most modern software technology, yet they are largely maintained by volunteer communities and non-profits. Google supports OSS communities through the Google Open Source Programs Office (OSPO), one of the industry's first OSPOs.[62] The OSPO focuses on expanding open-source technologies; sharing Google-developed technology under open licenses; and supporting open-source projects, communities, and maintainers across the entire open-source ecosystem. A federal OSPO, as proposed in the Cybersecurity and Infrastructure Security Agency's Open Source Software Security Roadmap, could bring similar benefits by managing federal agencies' consumption of and contributions to open sourcing.[63] This office would set policies for the government's use of OSS, including assessing and managing security risks, promoting best practices, and fostering collaboration among government agencies. The Department of Commerce (including NTIA, NIST, and perhaps the US AI Safety Institute) might coordinate to improve the security of the open-source ecosystem

[60] Zane Durante et al., *An Interactive Agent Foundation Model*, arXiv (Feb. 8, 2024).

[61] Google DeepMind, *A generalist AI agent for 3D virtual environments* (Mar. 13, 2024).

[62] Google Open Source Blog, *Establishing new baselines: Identifying open source work in an unstable world* (June 22, 2022).

[63] CISA, *CISA Open Source Software Security Roadmap* (Sept. 2023).

and anticipate the standards and shared principles that developers will need to incorporate as the open model ecosystem develops.

Conclusion

Google believes in AI's tremendous potential and appreciates this opportunity to comment on how NTIA should approach dual-use foundation models with widely available model weights.