stability.ai

Comments on Dual Use Foundation AI Models with Widely Available Model Weights
Response to the National Telecommunications and Information Administration

March 2024

I. Introduction **2**
II. Background **2**
III. Open models promote transparency, competition, and grassroots innovation in AI **4**
Response to questions 3(a)-3(d) on benefits
IV. Open models are just one part of a complex AI supply chain **8**
Response to questions 2(a), 5, and 5(b) on risk
There are many actors in the open supply chain 8
Open models can be deployed in a variety of applications 8
There are layers of mitigation for different kinds of emerging risks 9
V. Future policy should promote a diverse AI ecosystem **12**
Response to questions 6(a), 7(a)-(d), (i)-(j), and 8(b) on regulation
Direct intervention will have a dramatic chilling effect on grassroots innovation 12
Indirect intervention could have a surreptitious chilling effect on grassroots innovation 14
VI. Conclusion **16**
VII. Additional responses **16**
Response to question 1 on definitions
Response to question 1(a)-(b) on timeframes
Response to question 1(d) on local deployment
VIII. Annex A: Testimonials from our developer community **19**
IX. Annex B: Economic impact of open models in the United States **20**
Additional response to question 3(a)

Contact Ben Brooks, Head of Public Policy

Acknowledgments to Hassan Zayour for assistance with data collection

I. Introduction

1. Stability AI welcomes the opportunity to respond to the National Telecommunications and Information Administration (NTIA) request for comments on dual use foundation AI models with widely available weights. These open models play a vital role in promoting transparency and competition in AI. They drive grassroots innovation among everyday developers, independent researchers, and small businesses who are helping to build safer AI models and useful AI tools. Future regulation must support that diverse AI ecosystem – from the large firms building closed products to the everyday developers using, refining, and sharing open technology. We commend NTIA's engagement to date, and we encourage ongoing dialogue with the wider developer community to understand the direct and indirect effects of future reform on open innovation.

2. As the Administration considers the future of oversight, we urge it to vigorously promote open innovation in models. To that end, the following response outlines the importance of open models in the growing developer community; explains the implications of open models for supply chain governance, and outlines important considerations for any reform within or beyond Executive Order (EO) 14110. In addition, we include illustrative analysis that quantifies the potential economic benefits of a diverse AI ecosystem (featuring highly-capable open and closed models) over a restrictive ecosystem (featuring closed models alone) as a result of higher AI adoption (Annex B).

II. Background

3. Stability AI is a global company working to amplify human intelligence by making foundational AI technology accessible to all. Our team includes passionate researchers developing open models across a range of modalities, including image, language, and audio. Essentially, these models are software programs that can help a user to create, edit, or analyze complex content. With appropriate safeguards, we release these models publicly, sharing our software code along with the billions of distinctive settings or "parameters" that determine the model's performance.[1] That means everyday developers and independent researchers can integrate or adapt our models to develop their own AI models, build their own AI tools, or start their own AI ventures, subject to our acceptable use licenses.[2] To date, our models have been downloaded over 150 million times, and over 300,000 developers and creators actively contribute to our online community:[3]

    a. **Image.** We develop a family of image models, Stable Diffusion, that underpin up to 80 percent of all AI-generated imagery.[4] These models can take a text instruction or "prompt" from a user and help to create a new image. Since taking over the exclusive development of Stable Diffusion in late 2022,[5] we have released a number of improved image models including Stable Diffusion 2, SDXL, and a forthcoming Stable Diffusion 3. These models range in size from ~800 million to 8 billion parameters.

---

[1] We use the term "open" to refer to any models with publicly-available parameters. This nomenclature overlaps the EO 14110 definition of widely available models, but diverges in certain respects: see also our response to question 1.
[2] See e.g. the Open Responsible AI License (OpenRAIL) for Stable Diffusion 2 and SDXL, prohibiting a range of unlawful or misleading uses, available here and the Stability AI acceptable use policy, available here.
[3] Figures from Hugging Face and Discord, February 2024.
[4] Everypixel, 'AI Image Statistics', August 2023, available here.
[5] Stable Diffusion 2.0 onwards. Stable Diffusion 1.0 was released by the CompVis research team at LMU Munich; Stable Diffusion 1.5 was released by Runway. The repositories for these models are maintained by CompVis and Runway.

b. **Language.** We develop a suite of language models that can interpret, summarize, or generate text. These include highly capable large fine-tuned models (~70 billion parameters), compact base models (~1-7 billion parameters), specialized models for software development (~3 billion parameters), and models for underrepresented languages, including Spanish and Japanese. Our latest model family, Stable LM 2-1.6B, proportionally outperforms comparable models from Google and Microsoft in standard benchmarks.

c. **Audio.** We develop audio models such as Stable Audio, which generates high-quality 44kHz soundtracks and sound effects up to three minutes in length based on text instructions from the user. Stable Audio was recently listed on the *TIME* Best Inventions of 2023.[6] We intend to release open variants of audio models in due course.

d. **Video.** Building on this experience, we have developed an open video model that demonstrates new breakthroughs in video generation. From a supplied reference image, Stable Video Diffusion can generate and interpolate a continuous four second (14-25 frame) video clip.[7] In coming months, we expect rapid improvements in quality, control spatial reasoning, and duration, supporting the development of new creator tools.

e. **3D.** In addition, we have developed a series of open 3D models that can generate an accurate three-dimensional reconstruction from a given reference image.[8] These models demonstrate high precision, recall, and inference speed, opening up new possibilities for object rendering in design or gaming.

4. Stability AI provides a range of tools and services to help partners customize, deploy, or use our models, sustaining our open research efforts. In addition, we support academic research into scientific applications of AI.

5. We are committed to the safe development and safe deployment of AI, investing in research, tools, and partnerships to help mitigate emerging risks across the AI supply chain. In addition to our work with organizations such as Thorn, we are signatories to the White House *Voluntary AI Commitments*, and the British Government's *Joint Statement on Tackling Child Sexual Abuse in the Age of AI*, as well as members of the US AI Safety Institute (USAISI) Consortium and the

[6] See e.g. Stability AI, 'Fast Timing-Conditioned Latent Audio Diffusion', 2024, available here.
[7] See e.g. Stability AI, 'Scaling Rectified Flow Transformers for High Resolution Image Synthesis', 2024, available here; Stability AI, 'Stable LM 2 1.6B Technical Report', 2024, available here; Stability AI, 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', 2023, available here.
[8] See e.g. Stability AI, 'Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion', 2024, available here; Stability AI and Tripo AI, 'Fast 3D Object Reconstruction from a Single Image', 2024, available here.

Singapore Government's Generative AI Evaluation Sandbox. We participated in the first large-scale public evaluation of AI models at DEF CON, facilitated by the White House, and the UK AI Safety Summit, and we continue to engage with authorities in the US and around the world.

III. Open models promote transparency, competition, and grassroots innovation in AI

Response to questions 3(a)-3(d) on benefits

6. Generative AI will become critical infrastructure across the digital economy. Language models will power tools that revolutionize essential services, from education to healthcare; reshape how we search and access information online; and transform analysis, knowledge management, or decision making in some of our most important public and private sector institutions. Audiovisual models will power tools that radically accelerate the creative process, helping existing creators boost their productivity and experiment with new concepts, while lowering barriers to entry for people who may not have the resources or training to realize their creative potential today.

7. It is more important than ever that we can scrutinize these systems before the next wave of digital tools and services are built on "black box" technology operated by a small cluster of firms. Already, the digital economy relies on opaque systems that amplify content on social media, govern our access to information, determine our exposure to advertising, and mediate our online interactions. It is difficult to scrutinize these systems, and there are significant technical and economic barriers to building viable alternatives. Today, AI is at risk of repeating this history, accelerating the concentration of value and control in the digital economy.

8. In that context, open models play a vital role in the emerging AI ecosystem. Open models improve safety through transparency, foster competition in critical technology, and support grassroots innovation in AI:

    a. **Open models promote transparency.** Researchers and authorities can "look under the hood" of an open model to verify performance; identify risks or vulnerabilities; study interpretability techniques; and develop, apply, and evaluate mitigations. Open models enable third parties to directly inspect and correct the behavior of a model before and after the model is integrated into a deployed application. By comparison, closed models may not disclose how they are developed or how they operate. They may embed unidentified features, values, biases, or behaviors. Closed models may be comparatively opaque, and risk management may depend on trust in the developer.

    b. **Open models lower barriers to entry.** Training a new "base" model from scratch requires significant resources that are not available to everyday developers. Open models lower these barriers to entry. Everyday developers can build on open models to create new AI tools or launch new AI ventures without spending vast funds on research and computing. By way of illustration, OpenAI disclosed that it cost 100 million dollars to train the closed-source GPT-4 model.[9] Today, training a new 70 billion parameter language model via third-party compute services might cost nearly 5 million dollars.[10] By reducing these costs, open models help to ensure the economic benefits of AI accrue to a broad community of developers and small businesses, not just Big Tech firms with deep pockets.

[9] Wired, 'Open AI's CEO says the age of giant models is already over', April 2023, available here.
[10] Hugging Face, Training Cluster, available here.

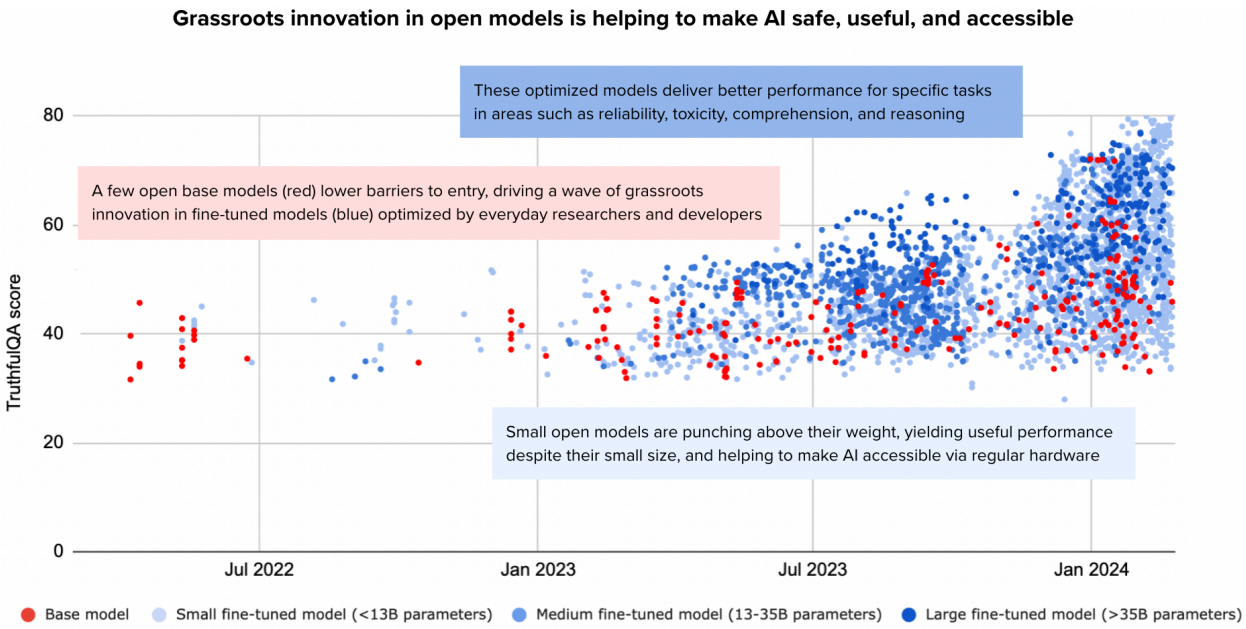
c. **Open models drive innovation in safety.** With access to a model's weights, developers can refine open models for improved safety and performance in specific tasks. For example, open models can be optimized through a range of techniques such as fine-tuning to mitigate undesirable behavior such as misinformation or toxicity, and amplify desirable behaviors. Further, developers can correct for observed biases, producing models that better represent different languages, communities, or values. These techniques can yield significant improvements in the behavior of a model without requiring the extensive computing resources necessary for pretraining. That means developers can quickly iterate to build safer and more effective models that better support real-world applications – many of which we can scarcely imagine today.

d. **Open models preserve security, privacy, and operational independence.** Open models enable firms and public sector agencies to build independent AI capabilities without relying on a handful of suppliers for essential components. Open models can be optimized and deployed without exposing confidential data; ceding control of the model's parameters; or risking unfair changes in the pricing, access, or performance of third-party services. Already, over a third of firms cite the sharing of proprietary data as a major obstacle to their deployment of language models.[11] Security, privacy, and operational independence will be particularly important for organizations handling sensitive data in regulated sectors, such as healthcare, finance, labor, public administration, and legal services. For users, locally-deployed AI systems built on open models can offer a viable alternative to third-party systems that harness or harvest personal data.

e. **Open models improve accessibility.** Many open models are smaller, more efficient, and more accessible than proprietary models. Unlike those models, which require significant computational resources to train and run, small open models can deliver useful performance with regular hardware. For example, open models may be hundreds of times smaller than a closed-source model such as GPT-4. Users can run small models on local devices, including smartphones, and developers can train or optimize these models with desktop hardware. Small models such as Gemma-2B from Google, Phi-1.5B from Microsoft, or our own Stable LM 2-1.6B may yield nearly half the performance of GPT-4 on benchmarks for reasoning or comprehension despite coming in at barely one-thousandth the size.[12]

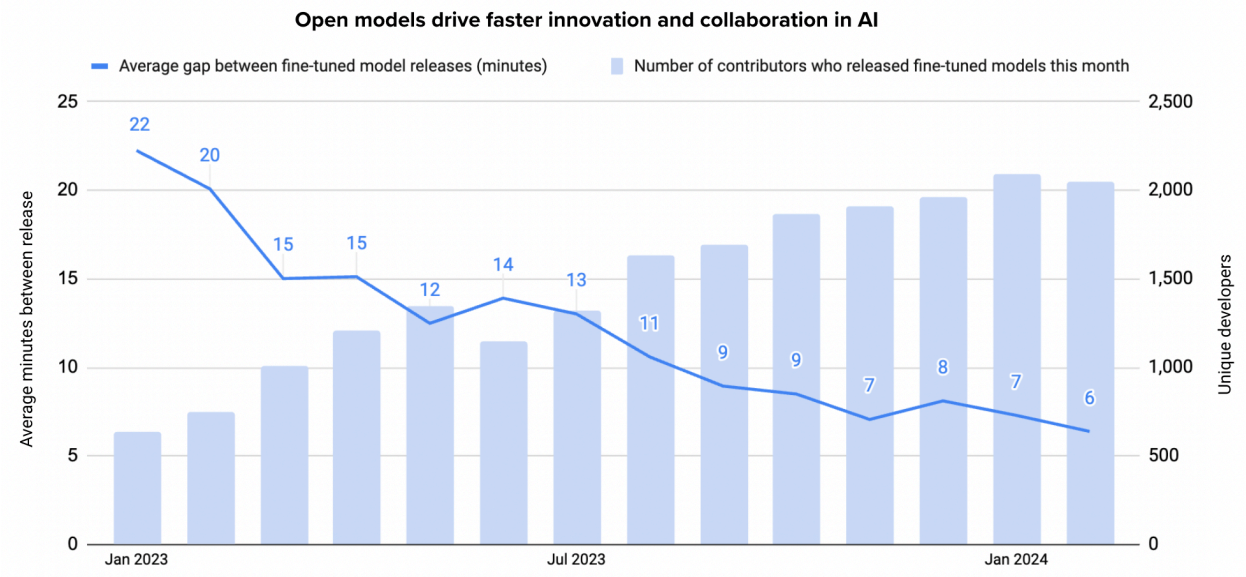
9. In this way, open models are fueling a wave of grassroots innovation in AI. Open models put this technology in the hands of everyday developers, independent researchers, and small businesses who are helping to build safer AI models and useful AI tools. Open models offer a transparent, competitive, and secure alternative to opaque technology operated by a small number of firms. These trends are supported by available data. Base models released by corporate or nonprofit labs are being optimized for better performance by third parties, and then redistributed publicly to support other research or development. These collaborative efforts are yielding significant improvements in performance:

[11] Predibase, 'Beyond the Buzz: A Survey Report of Large Language Models in Production', 2023.
[12] See, e.g. published Massive Multitask Language Understanding (MMLU) scores.



Source: Our analysis of 3,516 models from the Hugging Face 'Open LLM Leaderboard' (February 2024 inclusive). One of many benchmarks, the TruthfulQA benchmark measures a model's tendency to reproduce falsehoods. In this chart, we use "fine-tuned" broadly to refer to any variant model optimized through instruction or supervised fine-tuning, reinforcement learning, direct policy optimization, merging, or other techniques. The number of "base" models may be inflated by classification errors in the leaderboard. The x-axis represents the repository creation date on Hugging Face. We have excluded a small number of models released before March 2022 since their release dates are not accurately recorded in this data.

10. Open collaboration in models is helping to accelerate the pace of this innovation too. Supported by tools, infrastructure, and services from repositories such as Hugging Face and GitHub, open models are increasing the rate of development and release, and encouraging vast numbers of developers to contribute to research:



Source: Our analysis of 78,493 variant models released across 18 popular model families on Hugging Face (February 2024 inclusive). Actual rate of release across all models may be higher. Our count of variant models may include models optimized through a range of techniques, such as supervised fine-tuning, instruction fine-tuning, reinforcement learning, direct policy optimization, or merging.

Open base models are the bedrock for a grassroots ecosystem of contributors that supports downstream users, developers, and deployers

Open base model family	Number of third-party developers who have released variant models[i]	Number of third-party variant models publicly released to date[ii]	Total downloads of third-party variant models[iii]
BERT (Google)[iv]	11,577	44,395	237,576,000
Llama (Meta)	4,560	12,132	11,622,000
T5 (Google)	3,648	13,547	7,459,000
Mistral & Mixtral (Mistral AI)	1,570	4,955	4,544,000
SDXL (Stability AI)[v]	1,417	2,036	916,330
Falcon (TII)	355	615	303,000
BLOOM (BigScience)	274	648	92,000
Pythia (Eleuther AI)	160	637	95,000
Phi (Microsoft)	282	512	40,000
StableLM (Stability AI)	128	258	18,550

[i] Our analysis of 78,493 variant models released on Hugging Face (February 2024 inclusive). Our count of variant models may include models optimized through a range of techniques such as supervised fine-tuning, instruction fine-tuning, reinforcement learning, direct policy optimization, or merging. These figures represent model families that have been released for multiple years (e.g. T5) as well as models released only recently (e.g. Phi). [ii] These are approximate totals: models may be over- or underrepresented due to inconsistencies in naming conventions and model disclosure. [iii] Rounded to nearest thousand. [iv] For simplicity, this sum includes ~11,600 models developed from RoBERTa, a pre-trained model developed by Meta that adopts BERT architecture. Note that BERT is primarily used for text analysis, classification, and summarization tasks rather than text generation. [v] For comparison across different modalities beyond text, we include the SDXL image model developed by Stability AI.

11. We know from experience that open innovation promotes transparency, competition, and security in systemically important technology. Open research underpins many of the recent developments in AI. For example, Google openly published the research that gave rise to recent language models.[13] Meta, Google, and their partners chose to open-source foundational code libraries for machine learning.[14] University teams in Europe openly published the research that led to latent diffusion image generation,[15] and a range of privately- and publicly-funded organizations have chosen to release highly capable base language models openly, such as Falcon 180B, Llama 2, Mixtral 8x7B, or our own Stable LM family. Beyond AI, open-source operating systems such as Linux underpin a significant portion of web servers and data centers globally, and can be found on submarines, destroyers, and SpaceX rockets. Similarly, Android is an open-source mobile operating system that powers a majority of all smartphones worldwide.[16]

[13] Transformer via Vaswani et al, 'Attention is All You Need', 2017 available here.
[14] PyTorch via Paszke et al, 'PyTorch: An Imperative Style, High Performance, Deep Learning Library', 2016 available here; TensorFlow via Abadi et al, 'Large-Scale Machine Learning on Heterogeneous Distributed Systems', 2015 available here.
[15] Rombach et al, 'High-Resolution Image Synthesis with Latent Diffusion Models', 2021 available here.
[16] Vaughan-Nicols, 'From Earth to Orbit with Linux and SpaceX', ZDNET, 2020; Gallagher, 'The Navy's Newest Warship is Powered by Linux', *Ars Technica*, 2013; StatCounter, 'Mobile Operating System Market Share', 2023.

IV. Open models are just one part of a complex AI supply chain

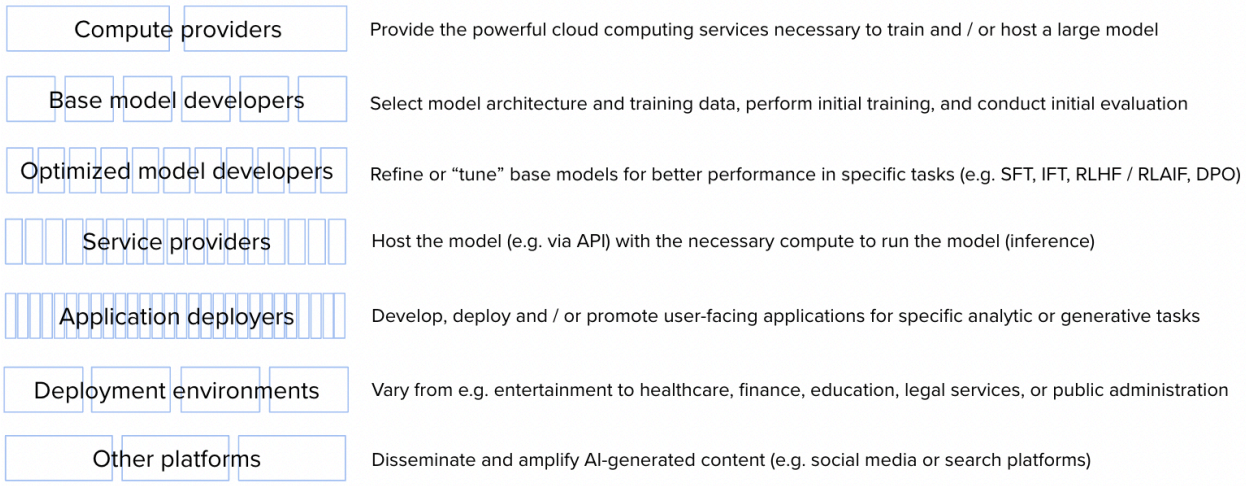
Response to questions 2(a), 5, and 5(b) on risk

12. Open models are technologies that can be integrated into a variety of applications by a range of actors. These characteristics have several implications for NTIA's analysis of risk and mitigation across the AI supply chain.

There are many actors in the open supply chain

13. Models are just one component in a generative AI system such as a chatbot or image generator. Base models should be understood as raw "engines" that can read, write, reason, draw, or animate. They develop these behaviors by observing complex patterns and relationships within a textual, visual, musical, or scientific dataset during pre-training. However, the base model may be prone to undesirable behaviors such as bias, misinformation, or toxicity. The model must be optimized or fine-tuned before deployment. Then, in order to usefully analyze or generate content, the model must be hosted and deployed in a user-facing system.

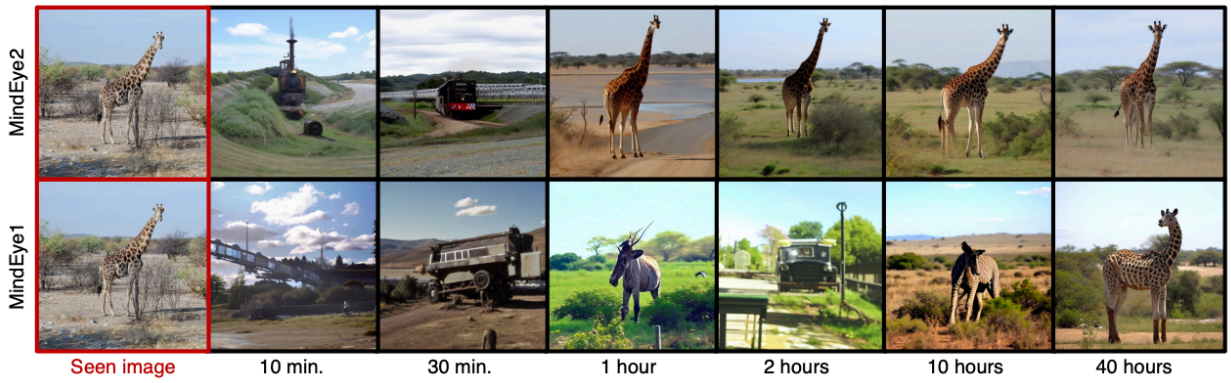
14. In this environment, different actors may perform different functions in an AI system. These range from: training the base model; optimizing the model for a specific use-case (such as conversational interactions); distributing the model; hosting the model on a computing service; developing a user-facing application that interacts with the model; and promoting that application to users. In an open ecosystem, the relationships between these actors may vary considerably compared to a vertically-integrated system built on closed models. In an open supply chain, upstream developers or researchers may have limited visibility or control over downstream activity. Responsibility for risk mitigation and assurance may be shared by different actors in different ways.



Open models can be deployed in a variety of applications

15. Generative models are versatile technologies that can support a wide range of tasks. Today, these models enable creative, analytic, and scientific applications – from personalized tutoring to drug discovery – that go far beyond the caricature of "push a button, get an image" or "push a button, get a poem". However, it is difficult to anticipate the full spectrum of potential applications.

By experimenting with open models, frontline developers, researchers, and businesses can determine how these technologies might be usefully applied to their domain. For example, our image models are fine-tuned and used by Broadway designers and architects to visualize new concepts; by photographers to edit or transform their images; and by research teams studying new approaches to diagnosing complex neurological disorders. Likewise, our language models can be used to summarize documents, edit content, or accelerate software development.



Above: Research teams experimenting with our open SDXL model demonstrate how AI can help to reconstruct a patient's visual perception (left) with as little as one hour of fMRI data (center). Disturbances in perception could help to assess or diagnose complex neurological disorders.[17]

16. Ultimately, the risk profile of an AI system will vary depending on how and where the system is deployed. For example, a system deployed in higher-stakes domains such as healthcare, finance, education, or public administration may attract more rigorous obligations than a system deployed in a lower-stakes domain such as entertainment. Vastly different requirements may apply for reliability, interpretability, explainability, and robustness. One models may be deployed in a range of different environments, and deployers bear significant responsibility for ensuring that their integrated system meets the minimum performance expectations or requirements applicable to that environment.

There are layers of mitigation for different kinds of emerging risks

17. Generative models of any kind could be deployed or misused in ways that present a risk of harm to users or third parties. We encourage NTIA to carefully distinguish between different types of risk, since these may require a different technical or regulatory response.

18. **Product safety risks.** In the short term, we expect the greatest risks are fundamentally product safety risks – the potential harms to users or third parties when unsuitable models are deployed in sensitive applications, such as a financial chatbot that generates misleading information causing injury or loss to a user. Generative models suffer a number of well-documented limitations, and these limitations need to be assessed and mitigated before models are deployed in sensitive applications:

    a. **Opacity.** These models are trained using a range of techniques that do not involve human supervision. As a result, the learned "rules" that dictate the model's performance are complex and opaque. It can be difficult to explain the relationship between an input (e.g. a

[17] MedARC et. al, 'MindEye2: Shared-Subject Models Enable fMRI-to-Image with 1 Hour of Data', 2024, available here.

question) and an output (e.g. an answer), or interpret how the model arrived at a particular output. These shortcomings are particularly significant in circumstances that require procedural fairness or redress, such as public administration, or other applications that implicate fundamental rights.

b. **Reliability.** These models analyze vast datasets to learn the hidden relationships between words, ideas, and textual or visual features. Their understanding of the world is determined by these relationships, and they have a limited understanding of other rules-based systems (e.g. scientific principles or social norms). As a result, models may "reason" in limited, erroneous, or unfamiliar ways. Further, the behavior of a language model is affected by the quantity, content, and quality of training data. They can amplify bias or errors in training data in ways that are difficult to detect, and they may unintentionally fabricate information. These behaviors can produce unreliable or misleading outputs, and relying on those outputs for advice or analysis may cause harm.

c. **Integration.** Models can be integrated with other components in ways that drastically affect the performance of the system. For example, an AI chatbot may use a model to summarize or analyze data that is obtained through a separate data retrieval system (e.g. a search engine). Alternatively, an AI tool may generate a response or recommendation based on inputs from a separate analytic system (e.g. a calculator). The performance of the model in a given task may be enhanced or degraded by these other components in ways that are difficult to assess without access to the underlying components.

19. In this context, open weights play a significant role in risk mitigation (i.e. the reduction of risk) by enabling developers and researchers to adjust the model's behavior before real-world deployment, taking into account their intended application. In addition, open models support risk assurance (i.e. the verification of risks and mitigations) by enabling deployers, researchers, and authorities to directly scrutinize the behavior of a model before and after deployment. In this way, open models facilitate research into new interpretability techniques that can help deployers to better validate outputs. Further, if given access to the underlying datasets, researchers can identify bias, errors, or omissions in training data, helping to anticipate potential risks prospectively.

20. **Misuse risks.** We acknowledge that open models pose unique challenges for other kinds of risks, such as the prevention of misuse. For example, language models could be misused to generate intentional disinformation, exploit software vulnerabilities, or summarize dangerous information. Audiovisual models may be misused to generate misleading or unlawful deepfakes. Stability AI is alert to these emerging risks, and applies a range of mitigations from development through to deployment. As with other technologies – from smartphones to word processors to photo editing software – there are no "silver bullets" to eliminate the risk of misuse altogether. However, there are layers of effective technical mitigations across the supply chain that can help to make it harder to do the wrong thing by introducing barriers to misuse:

    a. **Models.** As a first line of defense, models may be optimized for safer behavior prior to release through a range of techniques including data curation, instruction tuning, reinforcement learning from human or AI feedback, or direct policy optimization. For example, Stability AI filters unsafe content from training data, helping to prevent the model from producing unsafe content. Following pre-training, we evaluate and fine-tune our models to help eliminate undesirable behaviors, such as unacceptable bias. We

disclose known risks and limitations in standardized formats, such as model cards, to help downstream deployers decide on additional mitigations.[18] Our most capable models are subject to acceptable use licenses that prohibit a range of unlawful or misleading applications.[19]

b. **Deployers.** As a second line of defense, deployers may filter unsafe prompts and unsafe outputs when they host a model through an application or interface. Stability AI implements a number of such filters on our hosted services, and engages organizations such as Thorn to identify effective hashing, matching or classifier systems to support these filters. In addition, we apply imperceptible watermarks and content provenance metadata to images generated through our API.[20] We include watermarking modules by default in our open model libraries so that deployers can easily implement these watermarks.[21]

c. **Users.** As a third line of defense, users are governed by technology-neutral rules – state and federal – that apply to the misuse of AI models. These include laws pertaining to fraud, abuse, defamation, non-consensual intimate imagery, election interference, hacking, and privacy. Where necessary, these can and should be fortified to account for novel types of misuse or increased prevalence of misuse.

d. **Platforms.** As a fourth line of defense, countermeasures can be integrated across downstream platforms to detect and defend against misuse. For example, content distributors – such as social media, search, or streaming platforms – play an outsized role in the dissemination of harmful content, regardless of whether it is generated with or without AI tools. These platforms can use metadata, watermarks, classifier scores, and other signals to assess the provenance of content before amplifying it through their network.[22] In addition, platforms can deploy AI technology defensively. Today, machine learning classifiers are used to identify unsafe content on social media and pinpoint software vulnerabilities in complex security systems. Like conventional software, AI can be used as a shield, not just a sword, and we expect that defensive applications of AI will become increasingly effective in detecting various kinds of malicious content or conduct.

21. All actors have a role to play in mitigating the risk of misuse and, cumulatively, these mitigations provide an effective defense to emerging risks. These measures may be applied in different ways and in different configurations depending on the specific application. Different actors may contribute different capabilities to the final system, and contribute different mitigations.

[18] See e.g. 'Stable Diffusion V2-1 Model Card' available at https://huggingface.co/stabilityai/stable-diffusion-2-1.
[19] Open Responsible AI License (OpenRAIL) available at https://github.com/Stability-AI/stablediffusion/blob/main/LICENSE-MODEL and Acceptable Use Policy available at https://stability.ai/use-policy.
[20] CAI, 'C2PA', available here.
[21] Stability AI, 'Generative Models Repository', available here.
[22] For example, a platform can use the presence of metadata or watermarks to inform content recommendation decisions (i.e. upranking, downranking, or blocking content). Conversely, the absence of metadata or watermarks may be an important signal too. For example, a social media platform may choose to review or moderate photorealistic images from new or unverified accounts by default, unless the image has trusted metadata that confirms its origin.

V. Future policy should promote a diverse AI ecosystem

Response to questions 6(a), 7(a)-(d), (i)-(j), and 8(b) on regulation

22. The AI ecosystem is diverse – from large firms building vertically-integrated tools to everyday developers iterating on open models – and future reform should sustain that diversity. We encourage policymakers to preserve the culture of open access and open innovation in models that made recent AI breakthroughs possible, and that helps to make AI safer. However, grassroots innovation in models is uniquely sensitive to overbroad regulatory intervention. We urge care in the development of novel rules that could have a direct or indirect chilling effect on those who develop, optimize, or share open models.

Direct intervention will have a dramatic chilling effect on grassroots innovation

23. The most challenging proposals take the form of direct interventions, such as pre-development or pre-release authorization requirements. Already, there have been several efforts to develop a licensing regime, such as the bipartisan framework for premarket approval in Congress (which would apply to models with GPT-4 capability); the restrictions and prohibitions advocated by a range of firms and civil society organizations (including one highly-publicized proposal to license models with certain modest benchmark scores, capturing at least ~250 open models, under threat of expedited criminal prosecution); or the pre-release approval requirements tentatively introduced by foreign governments, such as India.[23] Pre-development or pre-release authorization requirements would have a dramatic chilling effect on open innovation by reversing the legal "presumption of openness" in the sharing of model weights and associated research, and by establishing conditions for approval that are inconsistent with open release and downstream optimization.

24. These types of controls are the exception, not the norm, in the regulation of software, research, and information with overwhelmingly legitimate applications. For example, the Export Administration Regulations (EAR) adopt a general principle that unclassified technology – including information necessary for the development, use, or operation of software[24] – should not be subject to EAR controls when it has been made publicly available.[25] Indeed, the US Government renegotiated international export control frameworks to ensure that overbroad definitions of "intrusion software" did not hamper the distribution of upstream technology or the

---

[23] See e.g. Sen. Blumenthal and Sen. Hawley, 'Blumenthal & Hawley Announce Bipartisan Framework on Artificial Intelligence Legislation', 2023, available here (proposing pre-development licensing for models with GPT-4 capabilities); Gladstone AI, 'Action Plan to Increase the Safety and Security of Advanced AI', 2024 (proposing premarket licensing for models with MMLU scores of 70 percent or higher, and banning the development of models with more than 10^{25} FLOPs of compute in training, with expedited criminal proceedings for those who develop or distribute such models without a license); Anderljung et. al, 'Frontier AI Regulation', 2023, available here (likening the premarket licensing of model development to aircraft certification or banking licensing); or Ministry of Electronics and Information Technology (India), Advisory No. 2(4)/2023–CyberLaws–3, 2024, available here at 2(c) (requiring "explicit permission" of the Government of India before making AI models available to users in India; this provision was subsequently amended); State Sen. Wiener (CA), S.B. 1047, 2023-2024 Reg. Sess. (Cal. 2024) (proposing pre-training reporting requirements for covered models).

[24] 15 CFR § 772.1.

[25] 15 CFR § 734.7 Exceptions to this principle such as non-standard cryptography focus on implementations of *proprietary* and *unpublished* functionality and, even then, only attract notification requirements: 15 CFR § 742.15.

disclosure of vulnerabilities, since these support legitimate security applications.[26] Additionally, the history of open-source software regulation illustrates that prior restraints on information sharing need to survive First Amendment scrutiny,[27] and they are unlikely to do so if they serve unclear objectives, address speculative harms, or adopt an overbroad approach. Restrictions on weights would invite similar scrutiny, and call to mind earlier disputes over pre-release controls:

> *If the government required that mathematicians obtain a pre-publication license prior to publishing material that included mathematical equations, we have no doubt that such a regime would be subject to scrutiny as a prior restraint. The availability of alternate means of expression, moreover, does not diminish the censorial power of such a restraint.*[28]

25. Within the open community, there is significant anxiety that recent model rules, such as those introduced by EO 14110, may be a precursor to future pre-release controls. These rules impose new obligations based on a variety of thresholds, which vary considerably between instruments such as EO 14110 (which imposes reporting obligations on "dual use" models trained with more than 10^{26} floating point operations (FLOPs) of compute, or as otherwise defined by the Secretary) and the EU AI Act (which imposes notification and testing obligations on "systemic risk" models trained with more than 10^{25} FLOPs of compute, or as otherwise defined by the European Commission). We recognize that the thresholds in EO 14110 are an effort to avoid overregulation by focusing regulatory attention on future models with unknown capabilities. However, the possibility of a future pre-development or pre-release licensing framework based on this framework is troubling.

26. First, these thresholds are coarse proxies for risk. There is still no framework for determining whether models above these thresholds actually pose a serious and unmitigated risk of catastrophic misuse. "Frontier" thresholds such as 10^{26} FLOPs might describe the size of existing models, but they are not indicative of actual capabilities or – more importantly – how those capabilities interact with existing systems, mitigations, and countermeasures. Further, there is mounting evidence that unpredictable "emergent" capabilities, which are invoked to justify frontier-type thresholds, are a reflection of poor metrics rather than an unpredictable and unforeseeable consequence of model scaling.[29] Instead, as with any technology, we encourage policymakers to (i) assess the initial risk of catastrophic misuse of a model, taking into account their realistic capabilities, (ii) measure the cumulative effectiveness of technical and legal mitigations across the supply chain to determine the residual risk of catastrophic misuse after mitigations have been applied, (iii) determine if the residual risk warrants further intervention,[30] and (iv) weigh the opportunity cost of restrictive regulation, including the impact on transparency, competition, and security.

---

[26] See e.g. negotiations over the Wassenaar Arrangement, as amended 2017, and associated Bureau of Industry and Security rulemakings. See also Sec. Pritzker, response to industry dated March 1, 2016: "In response to these concerns… the United States has proposed… to eliminate the controls on technology required for the development of 'intrusion software'… [W]e commit to ensuring that the benefits of controlling the export of the *purpose-built* tools at issue outweigh the harm" (emphasis added).

[27] *Bernstein v. Dep't of State*, 922 F. Supp. 1426 and subsequent case law.

[28] *Bernstein v. United States Dept. of Justice*, 176 F.3d 1132.

[29] See e.g. Schaeffer et. al., 'Are Emergent Capabilities of Large Language Models a Mirage?', 2023, available here.

[30] There is no bright line rule for this assessment in (iii), but the marginal risk over existing technologies could inform decision making: Kapoor et. al., 'On the Societal Impact of Open Foundation Models', 2024, available here.

27. Second, there is little consensus about which risks justify these kinds of controls in the first place. Online safety, election disinformation, smart malware, and fraud are some of the most immediate and tangible risks posed by generative AI. However, these risks are rarely invoked to justify premarket controls on other helpful software technologies with dual use applications. Photoshop, Word, Facebook, Google Search, and WhatsApp have contributed to the proliferation of deepfakes, fake news, and phishing scams, but we do not regulate their constituent technologies or underlying libraries. Regulation should be directed at specific risks, not broad technologies.

28. Alternatively, the prospect of chemical, biological, radiological, or nuclear (CBRN) risks have been invoked to justify new model obligations. However, there is limited evidence that open models represent a material increase in marginal risk over existing technologies such as search engines: in the words of the Institute for Human-Centered AI at Stanford, "while open models are conjectured to contribute to malicious uses of AI, the weakness of evidence is striking".[31] If these arguments are to justify a radical departure from our conventional approach to regulating technology, the standard of proof should be higher than speculation. Formal restrictions on sharing useful information and technology should be the last resort.

Indirect intervention could have a surreptitious chilling effect on grassroots innovation

29. In addition, there are a range of *indirect interventions* that, while appearing to neither favor nor disfavor open models, may have a disproportionate impact on grassroots innovation:

    a. **Overbroad pre-release obligations.** "One size fits all" frameworks governing model development could set back open innovation by imposing disproportionate requirements on every model. For example, until the final weeks of negotiation, the European Union AI Act imposed broad testing, record-keeping, and registration requirements on all models, regardless of how or whether they were actually deployed in an AI system, and regardless of whether they were base models released by a corporate lab or fine-tuned models released by an independent researcher.[32] While a corporate lab may have been able to comply with such requirements, these obligations are infeasible for ordinary developers, researchers, or small businesses. Ultimately, the EU amended the Act to partially exempt "free and open-source" model developers, and to narrow the obligations of downstream actors who fine-tune and release models.[33] However, above 10^{25} FLOPs (nearly equivalent to the latest Gemini model),[34] all models will be subject to the same obligations without exception.

    b. **Novel liability rules.** The complex supply chain means that actors may have varying levels of visibility and control over downstream activity. Policymakers should be cautious of imposing novel liability rules that assume vertical integration or formal relationships between actors in the supply chain, or that implicitly require upstream developers to exercise direct control over downstream fine-tuning or deployment. For example, a proposed rulemaking from the Federal Trade Commission and a variety of legislative

[31] Stanford HAI, 'Considerations for Governing Open Foundation Models', 2023 available here.
[32] European Parliament negotiating position, AI Act, art 2 and 28b, available here.
[33] Provisional Agreement, AI Act (EU), recital 60g ("In case of a modification or fine-tuning of a model, the obligations for providers should be limited to that modification or fine-tuning") and art 52c(2), available here. The open-source exemption does not apply to models above 10^{25} FLOPs or those deemed to have systemic risks in accordance with undefined criteria..
[34] Epoch AI, Machine Learning Trends, 2024, available here.

proposals in Congress might alter the liability of upstream developers for the downstream use of AI systems.[35] Taken to their extreme, these proposals could pose a significant obstacle to those who pre-train, fine-tune, or share model weights openly. Instead, liability should be determined through ordinary product liability principles, taking into account the relationships between different actors in different environments.

c. **Imprecise distribution of responsibilities.** Conflating different actors in the supply chain could impede open innovation by making it difficult for an upstream developer to comply with obligations targeted at a downstream deployer. For example, there are a range of legislative initiatives that seek to impose watermarking, labeling, disclosure, evaluation, or reporting requirements on AI systems. However, defining these obligations too broadly, or failing to distinguish between different actors, might capture upstream researchers or developers who play no role in the deployment of a user-facing system or the choice of application-layer features and mitigations. Instead, these obligations should be assigned with precision. To that end, we welcome recent legislative attempts to distinguish more carefully between developer and deployer obligations.[36]

Recommendations for policy development

30. Going forward, we encourage policymakers to carefully consider the impact of these direct and indirect interventions on grassroots innovation in open models. Pre-development and pre-release licensing are not supportable on the evidence available, and they will stifle open innovation in AI. Where possible, oversight frameworks should focus on AI systems in the context of specific applications. They should consider the risk of a deployed system in its specific environment, and avoid overemphasizing isolated components, such as models. Any obligations should be proportional to risk – not "one size fits all" – and they should account for the roles and relationships of different actors in the supply chain.

31. In addition, there are several steps that policymakers can take today to respond to the challenges posed by open technology. First, we continue to advocate for robust legal measures to deter misuse. For example, we have publicly urged clearer guardrails around the use of a person's physical or vocal likeness for improper purposes, with a focus on abusive content, election disinformation, and commercial exploitation. For instance, federal law imposes no criminal liability for the intentional distribution of non-consensual intimate imagery, and we welcome the growing support for legislative reform.[37] Where existing law falls short, we support efforts to fortify these guardrails in ways that (i) clearly and precisely define improper use of likeness, (ii) adopt a technology-neutral approach focused on conduct not tools, and (iii) account for the many legitimate applications of AI. As a starting point, we urge the Administration and Congress to

[35] See e.g. FTC, Supplemental Notice of Proposed Rulemaking on Impersonation of Government and Businesses, proposed §461.5: novel liability rules for upstream technology, especially with overbroad scienter requirements, could make it difficult or impossible to release useful technology publicly.
[36] S.3312, 118th Cong. (2023) from Sen. Thune and Sen. Klobuchar. This bill is one recent effort to clearly distinguish between developers and deployers in the context of high-impact and critical-impact systems: "the term 'developer' means an entity that (A) designs, codes, produces, or owns an artificial intelligence [component] for internal use or for use by a third party as a baseline model; and (B) does not act as a deployer of the artificial intelligence system described in subparagraph (A)."
[37] We note the release of e.g. H.R.3106, 118th Cong. (2023) from Rep. Morelle or H.R.5586, 118th Cong. (2023) from Rep. Clarke.

launch a whole-of-government gap analysis to identify shortcomings in existing regulatory mandates, agency resources, or legislative frameworks for misuse.[38]

32. Second, we encourage accelerated public research into model and system evaluation to support the diverse AI ecosystem. For example, the new USAISI can play a valuable role globally in standardizing benchmarking techniques, adversarial testing practices, and human evaluation processes. Standardized evaluation will help to provide confidence that AI systems deliver the expected or required performance for their intended deployment environment. In addition, USAISI can help to standardize, validate, and improve the range of available mitigations across the supply chain. Over the coming months, we encourage the USAISI to take a comprehensive approach to its research agenda, focusing on:

    a. **All actors.** Open models are developed, optimized, and deployed by many actors. Research should include good practices for evaluation and mitigation among both model developers as well as system deployers. These practices should account for large organizations (e.g. corporate labs) as well as thousands of everyday developers and small businesses.

    b. **All risks.** Open models engage a range of risks. Research should include the evaluation and mitigation of immediate risks, not just "frontier" risks in large models. That includes both product safety risks (e.g. reliability or toxicity) as well as misuse risks (e.g. software vulnerability discovery). Further, research should include better evaluation for (i) "off the shelf" capabilities in models as well as (ii) the robustness of models to malicious fine-tuning.

    c. **All modalities.** Open models are available in every modality, helping to support new kinds of creator tools. However, evaluation practices in non-language modalities are relatively underdeveloped. Research on evaluation and mitigation should include all modalities, such as image, video, and audio, in addition to language models.

VI. Conclusion

33. Open models play a vital role in driving grassroots innovation, helping to make AI safer and better. As the Administration considers the future of AI oversight, we urge NTIA to acknowledge the direct and indirect effects of different regulatory interventions on open innovation. To realize the full benefits of AI, future policy must promote a diverse AI ecosystem. We welcome NTIA's engagement with developers to date, and we would be pleased to discuss these matters further.

VII. Additional responses

Response to question 1 on definitions

34. The suitability of any definition for "open" or "widely available" models will depend on how that definition is used in future instruments. We urge NTIA to avoid proposing an overbroad "one size fits all" definition for all purposes. A definition intended to support research may be different to a definition adopted for the purposes of rulemaking.

[38] A good example of a gap analysis mandate is provided by Sen. Schumer, Sen. Young, and Sen. Heinrich in S.3050, 118th Cong. (2023) at sec. 3.

35. For research and standards development, we support an inclusive definition encompassing any model with weights that are released publicly. That could include weights released under a range of licenses that permit the use, modification, and redistribution of the model, some of which may not meet the canonical definition of "open-source software". Where necessary, the definition may be refined to take into account the "gradients" of release.[39] Variations in access to different components such as datasets, weights, and inference code – and the licensing conditions applicable to each – may affect how models are tested, deployed, or redistributed by downstream actors.[40] In some cases, these variations may be relevant to capability evaluation, risk mitigation, and interpretability research, and may be considered in the development of standards.

36. By comparison, a definition that triggers regulatory intervention should be drafted carefully, and should be targeted to specific risks. The open / closed distinction may not be relevant in many environments. For instance, the publication or retention of weights has little bearing on the reliability of a deployed application. As agencies develop minimum performance requirements for sensitive applications (such as in finance, healthcare, labor, or public administration), these requirements should be agnostic to the "openness" of an upstream component such as the model. Likewise, if the definition is intended to support a rulemaking for chemical, biological, radiological, and nuclear (CBRN) risks arising from misuse, the open / closed distinction may be of limited utility, since a single instance of theft or misuse could pose an unacceptable risk. In that case, the focus of regulatory action might be *any* weights that are not otherwise subject to heightened security controls, including closed models deployed for internal applications.

37. For similar reasons, any definition should avoid prescribing a single threshold based on the level of distribution. First, the level of distribution may not correlate with particular risks, such as the product safety risks in a deployed application. Second, the intended level of distribution may not necessarily correlate with the actual level of distribution.[41] Third, it may be difficult to monitor and quantify the real-time level of distribution in an open ecosystem given the limited visibility of developers and repositories into downstream activity. Finally, for the reasons given above, we encourage authorities to regulate for safety without limiting access to models through direct or indirect restrictions on distribution.

Response to question 1(a)-(b) on timeframes

38. There is ample evidence that closed models exhibiting category state of the art performance will be matched by open models in due course. Previously, it took ~28 months before an open model such as GPT-J from Eleuther AI approached the performance of a closed model such as GPT-2 from Open AI on common benchmarks. That gap is closing. Only ~eight months elapsed before open models such as Llama 2-70B from Meta rivaled GPT-3.5 from Open AI, and only ~ten months elapsed before Falcon-180B from the Technology Innovation Institute (funded by the Abu Dhabi Government) exceeded GPT-3.5 performance.[42] The same is true in other modalities. Open

[39] See e.g. Solaiman, 'The Gradients of Release: Methods and Considerations', 2023, available here.
[40] The Open Source Initiative is currently engaged in a public consultation on the appropriate definition of "open-source" AI components: OSI, 'Open-Source AI Draft Definition v.0.0.6' available here.
[41] See, e.g. leak of certain language model weights beyond the intended research community.
[42] MMLU results from Meta, 'Llama 2: Open Foundation and Fine-Tuned Chat Models', 2023 available here; TII, 'The Falcon Series of Open Language Models', 2023, available here; Open AI, 'GPT-4 Technical Report', 2023, available here.

image models such as Stable Diffusion approached the performance of closed models such as DALL-E 2 from Open AI within ~four months of the latter's release.[43] Open video models such as Stable Video Diffusion exceeded comparable closed models upon release, based on human evaluation.[44]

39. There are several explanations for the growth in model performance, and the narrowing gap between closed and open models. Fundamental architectures and libraries are open.[45] Compute resources continue to scale, halving the compute required for a language model to meet a performance target every eight to nine months.[46] Further, developers can obtain significant improvements in performance by training, or repeatedly training, on larger datasets rather than simply increasing a model's size.[47] These trends can reduce the cost barrier to training comparable open models. In addition, models can be optimized through a range of techniques, including fine-tuning and reinforcement learning, to obtain significant improvements in performance without extensive pre-training or large datasets.

40. While there is uncertainty about the persistence of these trends, or the likelihood of new breakthroughs in architecture, we encourage NTIA to assume that closed models will be followed quickly by open equivalents. While "frontier" models (e.g. trained with more than 10^{26} (US) or 10^{25} (EU) FLOPs) might appear to be beyond the range of existing open developers, these thresholds may be surpassed sooner than expected as compute performance doubles every ~2.3 years and compute costs halve every ~2.5 years,[48] and as large model behaviors are distilled or transferred into progressively smaller models through a range of techniques.

Response to question 1(d) on local deployment

41. Local deployment on-device or on-premise offers a number of advantages over intermediated access via a third-party application or API. Models need to be optimized prior to deployment for useful and reliable performance on a particular task. Local deployment ensures that sensitive data necessary for optimization is not shared with other actors, such as proprietary datasets used for supervised fine-tuning or reinforcement learning. Likewise, local deployment ensures that data is not shared during inference, such as prompts, uploaded documents / images, and outputs. The persistent disclosure of this data to a third party could compromise the privacy interests of users and the confidentiality of deployers.

42. Further, local deployment enables deployers to retain appropriate control over their AI capabilities. Local deployment helps to ensure that deployers are not subject to undisclosed updates or modifications to the model, which may inadvertently degrade the performance of their system for a particular task. In addition, local deployment ensures that upstream model providers or model hosts cannot impose arbitrary changes in pricing, access, or terms of use. As everyday developers and small businesses increasingly embed AI systems into their workflows, operational independence can help to minimize their exposure to unfair practices.

---

[43] Petsiuk et. al., 'Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark', 2022, available here.
[44] Blattmann et. al., 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', 2023, available here.
[45] See section III above.
[46] Ho et. al., 'Algorithmic Progress in Language Models', 2024, available here.
[47] Hoffmann et. al., 'Training Compute-Optimal Large Language Models', 2022, available here.
[48] Hobbhahn et. al., 'Trends in Machine Learning Hardware', 2023, available here.

VIII. Annex A: Testimonials from our developer community

We use open models to power key components of our core offering. We can't build a business entirely dependent on a third party and need some control over the models. We don't have access to the massive datasets and resources required to train foundation models. If these models are only available to massive corporations, innovation will suffer. Restrictions would create enormous uncertainty in the future and arbitrarily limit our business options.

– Dane O'Connor, founder of a tech firm, based in New York

As a designer and educator working with generative AI, I have consistently observed that groundbreaking innovations primarily originate from open-source platforms. The ability to fine-tune models and weights through the collaborative efforts of diverse user communities is a critical counterpoint to the potential biases and decisions imposed by closed-source AI corporations. Therefore, the widespread adoption and endorsement of open-source AI frameworks is imperative, as it not only fortifies the United States' standing as the forefront of AI research and development but also stimulates the growth, resilience, and variety within the field.

– Andrew Kudless, architect

As a designer, I've used Stable Diffusion to visualize ideas in a matter of minutes. I'll use it as a tool to create different iterations of an idea and manipulate quickly. I'll create moodboards and storyboards with it for other team members to see where we are headed.

– Stefania Bulbarella, Broadway projection designer

Open models accelerate adoption and sell products. Keeping models open allows smaller businesses to compete with larger businesses who close their models. [Without open models], I would lose a lot of business. I wouldn't have the same ability to provide capacity locally; I would have to outsource it through a corporation and pay hefty fees.

– Chris Watkins, president and founder of a technology consulting firm, based in Georgia

Open source is how we make money; without us building [our own] models [from open models], we can't effectively get our clients products into the content. My business relies on it – without it being open source we would have to shut down. Also, I use fine tunes shared by other creators to increase quality and my own output. It gives us small studios a competitive advantage which we would never have if this wasn't open source.

– Erik Toscano, co-founder and creative director of a content production studio

In the past year, I've used Stable Diffusion as an early conceptual design exploration tool in my contemplative architectural designs and also as a moodboard dynamic library of original reference images to visually communicate lighting design ideas in similar spaces with similar materials.

– Ilva Dodaj, architect and lighting designer

IX. Annex B: Economic impact of open models in the United States

Additional response to question 3(a)

See below.