HUGGING FACE

Hugging Face, Inc.
20 Jay Street, Suite 620,
Brooklyn, NY 11201

Hugging Face Response to the National Telecommunications and Information Administration Request for Comments on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Hugging Face applauds the ongoing work of the National Telecommunications and Information Administration (NTIA) in examining dual use foundation models (FMs). The following comments are informed by our experiences as an open platform for AI systems, working to make AI accessible and broadly available to researchers for responsible development.

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI.

Questions

In order to best address the risks and benefits of widely available model weights, we explore both the dimension of wide availability of a model and access to model weights in our responses to avoid conflating risks inherent in the technology with risks associated with release methods. We refer to models with widely available model weights as “open-weight models”.

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

Both “open” and “widely available” as terms are distinct, and require appropriate definitions when discussing risks and benefits.

“Open” should consider components that shape model impact. The risks and benefits of foundation models are shaped by decision along their development and deployment chain (e.g. when considering risks of biases and discriminatory outcomes[1]). The contribution of openness

[1] Let's talk about biases in machine learning! Ethics and Society Newsletter #2

to supporting beneficial development of the technology,[2] as well as the governance processes it requires to help mitigate risks of the technology, will depend on access and transparency into not just the model weights but also its development datasets[3], development conditions,[4] and deployment infrastructure.

Open weights for a foundation model are notable on two main accounts. First, given how resource-intensive the data curation and training of the models has become, sharing this particular artifact in the AI value chain is necessary to enable downstream research that only requires model access from many stakeholders outside of well-resourced companies.[5] Second, the release of a model's weights has historically served as a catalyst for development projects that prioritize overall transparency[6] and proactive consideration of regulatory processes[7] that enable more comprehensive research on risks and benefits along the development chain.

To account for all those dynamics, we strongly recommend considering at least three main components in FM openness: its **datasets**, the **trained model**, and the model's **training and deployment mechanisms**.

"Widely available" should consider breadth (not just scale) of access. In general, providing model access to a more diverse set of stakeholders can have a net positive impact on the safety of the technology; access should be available to stakeholders who are best positioned to identify risks and limitations[8] – and particularly to third-party organizations that do not have a direct relationship (or conflicts of interest) with the developer and direct beneficiaries.[9] Breadth of access can be achieved independently of scale through mechanisms such as gating, where access is granted by an (independent) organization for specific well-motivated research and development projects.[10] Simple availability is also distinct from accessibility; model weights being available may not be accessible to people without the necessary compute infrastructure or skill set for model hosting. For uses of a model that do not require specific customization, a managed API supported by extensive computational resources greatly increases the scale of availability independently of whether model weights are directly accessible.

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

Past examples include staged releases of model weights where closed weights are released over time, as seen with OpenAI's GPT-2[11], Meta's LLaMA[12], and Stability AI's Stable Diffusion[13]. All recent prominent closed models have been followed by an open comparable alternative.

[2] AI Policy @🤗: Open ML Considerations in the EU AI Act
[3] Policy Questions Blog 1: AI Data Transparency Remarks for NAIAC Panel 📚
[4] The BigCode Project Governance Card
[5] The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?
[6] Introducing The Foundation Model Transparency Index
[7] Do Foundation Model Providers Comply with the Draft EU AI Act?
[8] A Safe Harbor for Independent AI Evaluation
[9] Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance
[10] Access form for the BigScience training corpus
[11] Release Strategies and the Social Impacts of Language Models
[12] Introducing LLaMA: A foundational, 65-billion-parameter large language model
[13] Stable Diffusion Public Release — Stability AI

Examples include OpenAI's GPT-3 deployed via API in 2021 and model weights for BigScience's BLOOM and Meta's OPT both released in 2022. Recently, DataBricks was reported to have trained a competitive large language model for 10M$, significantly lower than the initial development cost of closed products;[14] indicating that relying on cost or trade secrets is unlikely to limit availability of top-performing models to their current set of developers.

Being "first to market" presents a strong economic competitive advantage for closed systems, and reproduction efforts of the model itself are motivated by a range of values, including scientific understanding, free and open source software (FOSS) ideals of accessibility, and customization, including fitness for purpose and harm mitigation strategies.[15]

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

Timelines vary and estimates will change by utility of the model and costs. Factors include computational cost, training data accessibility, and potential anti-competitive practices from original developers. Research focus and more compute and data-efficient algorithms are advantages for open models, whereas product focus, data concentration, high resources, and legally untested or contested practices are advantages for close development. Overall, socially responsible research in open reproduction of closed models will necessarily take more time for release[16] as legal and ethical concerns are addressed proactively rather than retroactively after issues have been observed at sufficient scale.

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

Model weights can be shared individually between parties, on platforms with or without documentation and with or without access management, and via p2p/torrent. Usability depends on resources and infrastructure; subsidized compute from cloud compute providers can increase access drastically, including for uses that go against the spirit of ToU but are hard to detect. APIs can broaden model availability, even to malicious actors, especially already relying on developer infrastructure. APIs also can be attacked to provide model information[17]. Even controlled risks can outweigh open models in magnitude through deployment scale.[18]

[14] Inside the Creation of DBRX, the World's Most Powerful Open Source AI Model | WIRED
[15] Policy Readout - Columbia Convening on Openness and AI
[16] BigScience: A Case Study in the Social Construction of a Multilingual Large Language Model
[17] Stealing Part of a Production Language Model
[18] Google's Bard AI chatbot is vulnerable to use by hackers. So is ChatGPT.

d. Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access? i. Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?

There is no standard for weighing risks and benefits, nor are there standards or clear methods for definitive access decisions. Distributed methods for sharing via torrent are arising such as tweeting magnet links as seen with xAI's Grok[19]. These are broadly accessible but do not allow for updates in information and rely on trust and robustness of the torrent sharer.

Web applications with user interfaces lower the barrier for users with little to no AI experience to engage with the model. Functionality depends on application and interface, such as the ability to toggle temperature. This is often surface-level access, and cannot be built upon.

API deployment can offer varying functionality, allowing only querying or allowing fine-tuning. APIs ease use for prescribed use cases as they only require internet connection and possibly a user account. APIs are often commercialized. Deployment benefits include monitoring, last-layer interventions, rate-limiting. Limitations include research and auditing inaccessibility, privacy concerns, and infrastructural reliability issues. API terms of use can be monitored and detected, and users can be blocked or revoked access. However content moderation proves difficult; blocks can easily be bypassed; false positives can block good faith researchers.[20] Last-layer interventions include watermarks after output generation. Limited API access makes research less reliable or presents additional barriers when limiting access to logits, forcing version changes, or obfuscating input and output processing that confound research findings. API deployment also introduces additional privacy challenges when user inputs are collected. APIs also entirely rely on a model hoster and their infrastructure, which may have significant demand and varying reliability.

Platform sharing of model weights enables broad access to models for local use and modification, controlled by access management (e.g. gating) and visibility (privacy, sensitive content tags) features. Content guidelines and ToU on model sharing platforms are defined at the model rather than at the use level. Releases on platforms can range from fully permissive to research-only (e.g. Cohere Command-R[21]).

Local hosting provides full control of the model and inputs to the model, satiates any privacy and sensitive information concerns, and opens many possible research directions. However, this is less monitorable for the deployer, should the user break terms of use. *Edge deployment* is a type of local hosting that has environment, portability, and privacy benefits.

[19] Grok model release via torrent on the X platform
[20] A Safe Harbor for AI Evaluation and Red Teaming
[21] Command R

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

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

In most cases, the risks associated with open-weight models are broadly similar to any other part of a software system (with or without AI components), and are similarly **context-dependent.**[22][23] Risks should also be weighed against non-AI systems.[24] Prominent considerations include new capabilities and systems designed specifically for harm.

New AI use cases arise fast, as seen with the easy creation of photo-realistic images enabled by diffusion models, and more recently video and audio generations. Accelerated development timelines can outpace societal and technical interventions. Both closed development and large-scale release with rapid distribution contribute to risks arising and timeline considerations. Open development often provides additional time and opportunities to question choices, allowing many perspectives to shape systems earlier design decisions[25]. Conversely, closed product development followed by rapid distribution can lead to significant disruptions in important sectors for sometimes uncertain benefits.[26][27] Rapid distribution depends on a combination of the availability of model weights, cost and technical difficulty of running a model, availability of a product or API version of an AI system, and broad advertisement and awareness of the AI system. Ensuring that scientific understanding of foundation models and development of new applications of the technology is more gradual helps identify issues and understand technological impacts earlier.

AI systems that are designed specifically for harmful uses, such as harassment or NCII, present distinct challenges from AI systems with dual use risks. Such models should most aptly be compared to malware in the context of traditional software.[28] On model sharing platforms, such models are subject to moderation following content policies.[29] Providing explicit guidance on categories of misuse and learning from existing mechanisms of platform governance to promote cybersecurity will be essential to managing those risks. In some cases, broadly useful models can be fine-tuned or adapted to specifically re-purpose them for such harmful uses. Safety by design practices can address those risks, for example limiting the prevalence of data that would make the models easier to misuse in general pre-training[30] or researching new methods for

[22] Open Source Software Security | CISA
[23] Artificial Intelligence | CISA
[24] On the Societal Impact of Open Foundation Models
[25] Deconstructing Design Decisions: Why Courts Must Interrogate Machine Learning and Other Technologies
[26] College professors are in 'full-on crisis mode' as they catch one 'ChatGPT plagiarist' after another
[27] The impact of ChatGPT on higher education
[28] Malware, Phishing, and Ransomware | Cybersecurity and Infrastructure Security Agency CISA
[29] Content Policy – Hugging Face
[30] StarCoder 2 and The Stack v2: The Next Generation; developers opted to remove known malware from pre-training data to make it harder to fine-tune for misuse

securing models at the weights level[31]. Additionally, fine-tuning models still requires collating datasets of harmful uses; limiting access to those represents another important layer of risk mitigation.

Taking action after a model release when new risks are identified presents **distinct challenges and opportunities for closed and open weights models**. API access models can be deleted either voluntarily by the deployer or by court request when found to be in breach of existing regulation,[32] but the cost of doing so when they are tightly integrated in a widely used commercial product may disincentivize organizations from taking actions, and internal scrutiny may not always be sufficient.[33] Conversely, good platform governance can more flexibly provide warnings and documentation, switch adoption to safer models, add technical access management such as gating, and remove models that violate platform policy; drastically reducing availability as a risk mitigation strategy.[34]

For models shared through peer-to-peer systems such as torrents, updates or removals will often rely on alternative communication channels and interventions in deployed contexts.

Releasing training and especially evaluation and testing datasets is usually a net positive from a risk perspective. Information about how a model was trained, where its data may present risks, what mitigation strategies were adopted, and how it was tested, can help to identify model vulnerabilities or limitations and conduct risk-benefit analyses. Challenges include intellectual property and privacy of data subjects. We note several approaches available to effectively manage these tensions[35].

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

Maximally open systems, including training data, weights, and evaluation protocol, can aid in identifying flaws and biases. Insufficient documentation can reduce effectiveness.[36] To maximize benefits, we recommend developers maintain up-to-date information about systems and publicly disclose biases and vulnerabilities as they are identified. The Hugging Face Hub platform supports a collaborative approach to model documentation.[37] We encourage adopters to maintain an inventory of which open systems they are using to help understand relevant risks to study and document. An example could be an inventory of systems used in the federal government.[38]

[31] Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models
[32] FTC settlement includes removal of trained facial recognition model
[33] Microsoft engineer sounds alarm on company's AI image generator in letter to FTC
[34] Content Policy – Hugging Face
[35] 📚 Training Data Transparency in AI: Tools, Trends, and Policy Recommendations 🗳️
[36] Black-Box Access is Insufficient for Rigorous AI Audits
[37] Model Cards - Hugging Face
[38] Hugging Face Response to OMB RFC on federal use of AI

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

Research has shown trained models can regurgitate private information from their pre-training. This is a sign of improperly trained models, and the behavior is difficult to predict. Weight availability can marginally exacerbate attacks, but is rarely the most efficient method. Model memorization is highly dependent on the model architecture, size, and training procedure.[39][40][41] Mitigations include proper data governance and management practices during training.[42] Open development has exemplified these practices, including via privacy-conscious data sourcing (e.g. BigScience[43]), and developing pseudonymization and personal data redaction methods for the training data domain (e.g. the StarCoder[44]). Privacy risks tied to the use of foundation models chiefly arise from API deployment settings that store user data,[45] however, which open-weights models can help mitigate (see our answer to 3b below).

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

Open-weight models are unnecessary for state actor threats; should state actors prefer to use FMs, they likely would prefer to build their own capabilities.[46] Narrow AI systems are more tailorable to a given threat. **Models with realistic outputs are expensive to use, often more so than human alternatives or narrower systems.** For example, the costs of human and AI generated disinformation do not always favor large language model (LLM) use.[47]

All existing models exacerbate risks given the following conditions: models introduce a new attack vector[48]; models reduce the overall cost of attack through generation or distribution[49]; malicious actors are unable to build a tailored model. When closed-weight models allow higher scale of use via free or compute-subsidized access, relative risks should be reevaluated.

Open-weight model risks can outweigh those of APIs when the **following additional conditions** are met: models need fine-tuning in ways unavailable via API; malicious content can easily be automatically identified as malicious (e.g. generated outputs for malware creation or legitimate software[50]); safeguards are robust to basic evasion (e.g. "jailbreaking"[51] or traditional content moderation evasion techniques[52]).

[39] Quantifying Memorization Across Neural Language Models
[40] Extracting Training Data from Diffusion Models
[41] Understanding and Mitigating Copying in Diffusion Models
[42] https://github.com/allenai/fm-cheatsheet/blob/main/app/resources/paper.pdf
[43] Documenting Geographically and Contextually Diverse Data Sources: The BigScience Catalogue of Language Data and Resources
[44] StarCoder: may the source be with you!
[45] AIDB: ChatGPT Banned by Italian Authority Due to OpenAI's Lack of Legal Basis for Data Collection and Age Verification
[46] Yi-34B, Llama 2, and common practices in LLM training: a fact check of the New York Times | EleutherAI Blog
[47] How Much Money Could Large Language Models Save Propagandists? | Center for Security and Emerging Technology
[48] Google's Bard AI chatbot is vulnerable to use by hackers. So is ChatGPT.
[49] The criminal use of ChatGPT – a cautionary tale about large language models | Europol
[50] Incident 443: ChatGPT Abused to Develop Malicious Softwares
[51] [2307.15043] Universal and Transferable Adversarial Attacks on Aligned Language Models
[52] Microsoft CEO responds to AI-generated Taylor Swift fake nude images

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

Cross-jurisdictional legal uncertainty can be detrimental to research collaborations that span sectors, organizations, and geographies. This includes liability per contributor, IP and copyright law, and foundational laws around digital privacy and regulation of misuses such as CSAM.

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

In terms of novelty and scalability of the application, disparate impact of openness, and severity of the harms, non-consensual intimate imagery (NCII) in static images and videos remains the most serious and pressing risk factor of new FMs. Openness, including access to training datasets and training code, aids mitigation by enabling external scrutiny through investigating the highest contributor factors (especially at the dataset level[53]) and enabling safety by design.

Marginal risk for open-weight models is less certain for pressing risks such as nonconsensual voice cloning; hosted models can be equally harmful[54] with little recourse. The most concerning models fall well below proposed thresholds for dual use, and can easily be independently trained from scratch by medium-resourced actors. **Effective policy action should target use cases and provide guidelines for all models across availability.**

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

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

Open-weight models contribute to competition, innovation, and broad understanding of AI systems to support effective and reliable development. In terms of **value creation of open technology**, the historical and economic impact of Open Source Software (OSS) provides important context for the expected impact of open-weight models. Two main phenomena bear strong relevance: First, the estimated demand-side value of OSS is several orders of magnitude larger ($8.8 trillion) than its supply-side value (>$4 billion).[55] Investment in open technology pays off over 1000x in value in other areas of the economy. Second, adoption enables users to reallocate resources to internal development that supports market diversification and sustainable self-driven growth[56], which is critical for start-ups and SMEs. OSS trades off capital for human expertise. **Recent work has shown that these dynamics are extending to AI;**

[53] Multimodal datasets: misogyny, pornography, and malignant stereotypes
[54] AI Startup ElevenLabs Bans Account Blamed for Biden Audio Deepfake - Bloomberg
[55] The Value of Open Source Software
[56] Open Source Software and Global Entrepreneurship

cost efficiency and customizability increasingly outweigh managed solutions, given sufficient company expertise.[57]

Open-weight models are particularly relevant to **mitigating market concentration at the data level**. High-quality data, both in the form of licensed content,[58][59] interaction data, and data uploaded by customers of managed systems, is shaping up to be the next most likely point of monopoly behaviors as training compute costs decrease (see our answer to Q1.a). Open weights models increase the diversity of product offerings from providers and even enable adopters to manage their own on-premise solutions. **Companies using open-weights models or systems built on them are in a position to keep their own data value,**[60] **preserve privacy, and build consortia on their own terms as needed** to support better technology development for their applications without having to give up their intellectual property.[61] This is needed to sustain high-quality data creation and fair valuation of the data contribution to AI systems. Additionally, open-weight models have been customized and adapted to run on a greater variety of infrastructure, including individual GPUs and even CPUs, reducing points of **market concentration with cloud providers** and reducing costs for procurement.[62][63]

AI foundation models combine innovative ways of building AI systems–and **most innovation supporting recent AI product releases has come from open and academic research.** For example, the video generation system SORA builds on many public research contributions.[64] Software behind development is largely open source (Pytorch, Tensorflow, HF libraries) and development often requires access to the weights. More convenient[65] and more secure[66] weight storage formats are developed on OSS settings, among many framework improvements. This includes fine-tuning major LLMs and research breakthroughs such as model merging.[67] Open-weight models' role in enhancing existing scientific research is evident in works on topics ranging from knowledge distillation[68] to watermarking[69]. Robust innovation on both performance and safety questions requires scientific rigor and scrutiny, which is enabled by openness and external reproducibility[70]. **Supporting that research requires sharing models** to validate findings and lower the barrier to entry for participation given the growing resource gap between researchers in different institutions.[71] Recent open-weights releases of foundation models from AI system providers such as Cohere's Command+R,[72] Google's Gemma,[73] and OpenAI's Whisper 3[74], among others, represent a welcome contribution to this research.

[57] 16 Changes to the Way Enterprises Are Building and Buying Generative AI | Andreessen Horowitz
[58] Exclusive: Reddit in AI content licensing deal with Google | Reuters
[59] OpenAI In Talks With Dozens of Publishers to License Content - Bloomberg
[60] Will StarCoder 2 Win Over Enterprises?
[61] SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
[62] GitHub - ggerganov/llama.cpp: LLM inference in C/C++
[63] llamafile: bringing LLMs to the people, and to your own computer - Mozilla Innovations
[64] Sora Reference Papers - a fffiloni Collection
[65] GGUF model saving format
[66] GitHub - huggingface/safetensors: Simple, safe way to store and distribute tensors
[67] Evolving New Foundation Models: Unleashing the Power of Automating Model Development
[68] Knowledge Distillation of Large Language Models
[69] A Watermark for Large Language Models
[70] EleutherAI: Going Beyond "Open Science" to "Science in the Open"
[71] The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?
[72] Command R
[73] Gemma: Google introduces new state-of-the-art open models
[74] Introducing Whisper, openai/whisper-large-v3 · Hugging Face

Open Weights Models and Evaluation

The availability of open-weights foundation models is of particular importance to the development of the **robust and reliable evaluation ecosystem** required to properly govern this category of technology. Access to open-weights models serves several distinct purposes for evaluation. First, some categories of evaluation **require direct access to model weights**. Evaluation of a model's environmental impact and energy consumption, for example, involves running models on controlled hardware.[75] Evaluating model memorization behaviors[76] (related to privacy and intellectual property concerns), covert biases,[77] and data contamination[78] all require varying levels of access from output logits to model activations **without additional input or output processing** that are not consistently available through API deployment.

Second, access to model weights makes research and **development of new evaluations significantly cheaper**. Developing a new evaluation technique typically requires extensive iteration on variations of a system. Being able to run a model locally in a controlled environment makes this process significantly more efficient and cheaper, and has been instrumental for example in our own work on developing bias evaluations in image generation settings before applying them to commercial systems.[79]

Third, while model developers should be incentivized to thoroughly evaluate their models, their **incentives and priorities may be mis-aligned with those of external stakeholders**, especially when models are part of commercial product development. Developer-run evaluations may underplay limitations and harms of a model – often without explicitly setting out to do so; to address this, recent work has called for safe harbor regimes to enable external research[80], and outlined limitations of "black-box" access[81] and auditing without sufficient agency.[82] Similarly, self-reported performance numbers on task can often be misleading or over-represent a model's suitability for use. Common issues that contribute to this phenomenon include misleading evaluation framing[83], choice of metrics (leading to different perceptions of "emergen capabilities"[84]), rampant issues of inappropriate benchmark decontamination,[85][86] and reporting results on different model versions than the ones deployed in products.[87] Finally, evaluations of foundation model applications in specific domains should be led by experts in these domains rather than by model developers to best reflect the risks and benefits of AI adoption.[88]

[75] Power Hungry Processing: Watts Driving the Cost of AI Deployment?
[76] Quantifying Memorization Across Neural Language Models
[77] Dialect prejudice predicts AI decisions about people's character, employability, and criminality
[78] Membership Inference Attacks against Language Models via Neighbourhood Comparison - ACL Anthology
[79] Stable Bias: Analyzing Societal Representations in Diffusion Models
[80] A Safe Harbor for AI Evaluation and Red Teaming
[81] Black-Box Access is Insufficient for Rigorous AI Audits
[82] AI auditing: The Broken Bus on the Road to AI Accountability
[83] GPT-4 and professional benchmarks: the wrong answer to the wrong question
[84] Are Emergent Abilities of Large Language Models a Mirage?
[85] Investigating Data Contamination for Pre-training Language Models
[86] Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs
[87] An In-depth Look at Gemini's Language Abilities
[88] The Shaky Foundations of Foundation Models in Healthcare

Fourth, **results obtained through third-party use of a model's managed API are unreliable** without extensive documentation of the model versions, processing steps, or risk mitigation strategies. Model providers can change the underlying version of a model at any time,[89] and do not always disclose sufficient details about post-processing steps that might have an impact on the results, such as automatic edition of prompts.[90] This is particularly relevant since, regardless of any intentional misrepresentation, commonly used evaluations have been shown to be sensitive to specific implementation details that often require extensive evaluation to characterize.[91][92]

Access to open-weights models that are substantially similar to those used in commercial products is thus **necessary to support a robust evaluation ecosystem**, both to guide further development of the technology and to support the development of appropriate standards.

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

Open-weight models can help in identifying attacks before they can be leveraged at scale; for example, recent work on "stealing" the last layer was developed with LLaMA 2[93]. Research has also found some adversarial attacks apply to open and closed models,[94] flagging the need to understand limitations of safety techniques. As discussed in 3a, reproducibility increases trust in AI systems by ensuring scientific rigor. Replicable evaluations with full transparency of replicable pieces can resolve misunderstandings or misleading comparisons.[95] Deployers canalso update models as needed, without being locked into hosted model contracts and agreements - which is particularly relevant when model vulnerabilities are identified or when a plausibly more secure model becomes available.

Prominent **privacy concerns arise from API deployment** and user data storage[96][97] and training on user data that contains private information.[98] This tendency is particularly concerning as access to private information has been identified as a significantly stronger risk factor than model capacity in some misinformation settings.[99] **Open-weight models can mitigate this risk** by enabling a more controlled deployment environment and obviating the need to send data to third party organizations. Open practices can enable broader red-teaming and evaluation practices focused on privacy that can help secure all models.[100]

[89] GPT-4 API general availability and deprecation of older models in the Completions API
[90] DALL·E 3
[91] What's going on with the Open LLM Leaderboard?
[92] Open LLM Leaderboard: DROP deep dive
[93] Stealing Part of a Production Language Model
[94] Universal and Transferable Adversarial Attacks on Aligned Language Models
[95] What's going on with the Open LLM Leaderboard?
[96] March 20 ChatGPT outage: Here's what happened
[97] AI Incident database: Incident 513: ChatGPT Banned by Italian Authority Due to OpenAI's Lack of Legal Basis for Data Collection and Age Verification
[98] "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents
[99] On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial
[100] Privacy risks in deployment: Introducing the Chatbot Guardrails Arena

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

Yes; in addition to examples in 3a, LLaMA's adaptation into Alpaca[101] significantly eased accessibility while maintaining high performance. Some API restrictions prevent equity-advancing research.[102] Open practices, such as open data work, can help address historical biases and prevent "hate-scaling".[103]

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

Access to the training data of an open-weights foundation model provides substantial additional benefits from the perspective of research, rights and governance, and safety.[104] Direct access top the training dataset can support research on model privacy and memorization,[105] risks tied to hate content[106] or NCII generation[107], intentional and unintentional impacts of data filtering approaches[108], among others. Interactive or managed access to a dataset can similarly help answer questions about data provenance and privacy[109] and support data governance centered on the rights of data subjects.[110] Support for projects that develop foundation models in fully open settings by making both trained weights and access or extensive documentation of their dataset available can enable more socially responsible and inclusive technology development.[111]

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.

In addition to examining artifacts listed in 1, often-overlooked artifacts include process and documentation. Project governance shapes technical artifacts, including goals, trade-offs, funding, and mechanisms for internal feedback. In fully open development, processes include goal setting, value alignment, and enabling various stakeholders to question upstream choices. The BigScience project provides this level of openness for a multilingual LLM,[112] and the BigCode project[113] took a similar approach to develop a code LLM, including a new form of Governance Card to report relevant information including funding, main development choices and reasoning, handling of cybersecurity risks and personal data, and annotator pay for data augmentation work.[114]

[101] Alpaca: A Strong, Replicable Instruction-Following Model
[102] Dialect prejudice predicts AI decisions about people's character, employability, and criminality
[103] On Hate Scaling Laws For Data-Swamps
[104] 📚 Training Data Transparency in AI: Tools, Trends, and Policy Recommendations 🗳️
[105] How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN
[106] On Hate Scaling Laws For Data-Swamps
[107] Multimodal datasets: misogyny, pornography, and malignant stereotypes
[108] A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity
[109] The ROOTS Search Tool: Data Transparency for LLMs
[110] Data Governance in the Age of Large-Scale Data-Driven Language Technology
[111] Social Context of LLMs - the BigScience Approach, Part 3: Data Governance and Representation | Montreal AI Ethics Institute
[112] BigScience: A Case Study in the Social Construction of a Multilingual Large Language Model
[113] BigCode Project
[114] The BigCode Project Governance Card

Documentation approaches include model cards[115] and datasheets[116] that can be shared broadly. Accompanying weights with information necessary to assess their usability for a given use case helps ensure they are used in safer contexts. Data quality is broadly recognized to directly affect ML system performance,[117] and dataset limitations, such as toxicity and harmful social biases, are typically reflected in a trained model.[118] Datasets are often the most practical artifact of study to understand the capabilities, risks, and limitations of a model. Sufficient access to a training dataset[119] can help answer relevant questions about a model's characteristics[120].

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?

We need significantly more investment in evaluation to foster practices of safety, privacy, and fairness by design. Results from reliable evaluation methodologies and design decisions such as data selection can determine whether a model should be used, and in what conditions.

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

Critical variables for assessing risks and benefits include type of model and design process and data selection. Model evaluations are important in determining usability and use case, and in shaping moderation decisions, but should be complemented with sufficient documentation of both the development process and governance mechanisms (see responses to 3e and 4) and the specific evaluation methodology (see response to 3a). Past moderation decisions on Hugging Face about whether sharing specific models was consistent with our terms of service has depended on all of those factors, as model evaluations by themselves can fail to sufficiently address social dynamics and the context of use and development.

Models trained specifically for sensitive applications, such as identifying personal data (e.g. the StarPII model[121] used for private information redaction in the BigCode project training data) or producing hate speech,[122] require different governance and access to more specific stakeholders. Processes for determining release method will differ by deployer organization and calls for collective decision making boards are unanswered, although discourse is ongoing.[123][124]

[115] Model Cards for Model Reporting
[116] Datasheets for Datasets
[117] The Effects of Data Quality on Machine Learning Performance
[118] A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity
[119] 📚 Training Data Transparency in AI: Tools, Trends, and Policy Recommendations 🗳️
[120] The ROOTS Search Tool: Data Transparency for LLMs
[121] StarPII release decision
[122] Handling and Presenting Harmful Text in NLP Research
[123] Publication Norms for Responsible AI - Partnership on AI
[124] The Time Is Now to Develop Community Norms for the Release of Foundation Models

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

Privacy by design includes steps throughout development including removing unnecessary personal data from a training dataset (see BigScience and BigCode examples in 2c) and following data minimization principles[125] helps reduce risks described in 2c. However, privacy risks also come from the extent to which current deployed systems encourage users to share personal data during their interactions.[126] Ensuring that the product design minimizes instances where this information is stored or transmitted helps prevent damaging breaches.

Interventions at the fine-tuning level, such as in Constitutional AI or reinforcement learning with human feedback to aid instruction following, can help steer a model towards more desirable behaviors. While these interventions have an important role to play in adding friction to misuses of models, they are also limited in their robustness and effectiveness[127] and insufficient to address fundamental issues with models.[128][129] Input and output monitoring can also be used to automatically block queries that are incompatible with a model's term of services. However,this approach presents all of the familiar limitations and potential discriminate impacts of automatic content moderation. In particular, the social cost of developing these safeguard datasets[130] and of live content moderation[131] should be considered and minimized.

Model weight encryption is an ongoing research area, with proposals[132] but no wide deployment.

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

In addition to building on 5b, safety by design, tight scoping of objectives, and intentional and robust data curation all are paths forward. Using AI as safeguards on AI systems at the deployment level may mitigate some risks, can easily compound biases and have disproportionate impact on minority groups using the systems.[133]

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

Clearly defined platform governance frameworks and disclosure mechanisms can most effectively limit availability for systems that require this guardrail due to vulnerabilities or incompatibility with existing regulations. For models found to present critical vulnerabilities, AI can draw on cybersecurity practices to notify users of updates. More research is needed to

[125] Artificial intelligence: the CNIL opens a consultation on the creation of datasets for AI
[126] "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conv
[127] Jailbreaking Black Box Large Language Models in Twenty Queries
[128] Dialect prejudice predicts AI decisions about people's character, employability, and criminality
[129] The Male CEO and the Female Assistant: Probing Gender Biases in Text-To-Image Models Through Paired Stereotype Test
[130] Inmates in Finland are training AI as part of prison labor - The Verge
[131] Inside Facebook's African Sweatshop | TIME
[132] NN-Lock: A Lightweight Authorization to Prevent IP Threats of Deep Learning Models | ACM Journal on Emerging Technologies in Computing Systems
[133] Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection

scope what constitutes a critical vulnerability and to standardize evaluations. For models that violate existing regulations or platform content policies, removal from the platform can help limit their spread and mitigate their negative uses. On the Hugging Face platform, we have taken actions to limit the use of models trained in a way that leads them to disproportionately produce harassing text (e.g. GPT4chan[134]) or models that were trained to reproduce a person's voice or likeness without their consent.

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

Approaches can include a trusted research environment that provides researchers full access to model components through an isolated virtual machine (e.g. HathiTrust Data Capsules[135]) to ensure security, but are resource-intensive in terms of both computational and human support to adapt the environment to specific research needs. FMs and LLMs are notably computationally intensive. Public repositories with access management (such as Hugging Face gated repositories[136]) can conveniently balance easy access for legitimate researchers and stakeholders and limited release of model weights. In staged or gated releases, researchers may enter legal agreements with terms of use including preventing model weight distribution.

f. Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, 14 please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.

Different tasks and level of access needed per task may not be obvious just as risk landscapes for dual-use FMs cannot be fully taxonomized at a given time. While this mapping exercise has not and likely cannot be exhaustively conducted, some examples include bias evaluations listed in 3a and Table 1 of the paper Black-Box Access is Insufficient for Rigorous AI Audits details audit tasks and levels of access needed.[137]

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

Hugging Face has used a simple model hash to verify whether two models are the same item. For open-weight models, verification is as simple as for any other large files. For APIs, verification is significantly more complex given the stochastic nature of the outputs and complex stack. As outlined above, this can be a source of security vulnerabilities for critical infrastructure that relies on externally managed APIs.

[134] ykilcher/gpt-4chan · Hugging Face
[135] Data Capsules
[136] Gated models
[137] [2401.14446] Black-Box Access is Insufficient for Rigorous AI Audits

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?

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

Popular measures include terms of licenses with use clauses[138] and gating mechanisms in addition to safeguards enumerated in recent work.[139]

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

As noted in 2, 3, and 5, wide availability of open foundation model weights facilitates government action in AI regulation by supporting research on evaluation and standards, fostering more transparent and accountable development, and promoting fair competition. As noted in 2, governance of widely available open-weights models can also require different interventions than API-only models, with different benefits and challenges.

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

Entities deploying AI systems should typically disclose that fact, as noted in the AI Bill of Rights.[140] In general, extending the concept of a Software Bill of Materials (SBOM)[141] to AI systems, including identifying which versions of foundation models (open or closed) are being used would have beneficial implications for cybersecurity and reliability.

d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights? i. Should other government or non-government bodies, currently existing or not, support the government in this role? Should this vary by sector?

Evaluation presents two main challenges. First, specific risk evaluations is an open research area, and risk priorities vary by organization. Evaluations as a field require significant investment; widely used benchmarks for both performance and risks are criticized for often providing insufficient or even misleading results.[142] Second, model evaluation often depends on the type of models evaluated and type of risk[143], especially for topics like privacy risks. Continual investment is needed to ensure evaluations reflect risks as models evolve. Sectoral contexts improve metric robustness.

[138] OpenRAIL: Towards open and responsible AI licensing frameworks
[139] The Gradient of Generative AI Release: Methods and Considerations
[140] Blueprint for an AI Bill of Rights | OSTP | The White House
[141] Software Bill of Materials (SBOM) | CISA
[142] Dialect prejudice predicts AI decisions about people's character, employability, and criminality
[143] Evaluating the Social Impact of Generative AI Systems in Systems and Society

The U.S. government action can include establishing standards for best practices building on existing work[144] and prioritize requirements of safety by design across both the AI development chain and its deployment environments.[145] General conditions should treat models that are broadly commercialized and broadly shared similarly. **Actions should foster a research ecosystem that has sufficient access to the artifacts and infrastructure to conduct research, incentives to share lessons and artifacts for reproducibility, access to broad subject matter experts including social scientists.**

e. What should the role of model hosting services (e.g. HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed? 16

At Hugging Face, use and utility are contextual. Our platform provides features for access management to enable developers to tailor model availability and breadth of access to strike the best balance between benefits and risks. We universally encourage extensive and transparent documentation, mandating documentation for artifacts over 10,000 downloads and providing guides and resources. Content guidelines[146] address specific harms, including breach of consent[147] and use of models for intentional harm. Sharing platforms enable research on systems that may serve different purposes, and in many cases purely as research artifacts.

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

Given its responsibility to stakeholders, the government should meet sufficient procurement requirements to ensure accountability, rights-respecting deployment of technology, and to ensure that AI adoption does provide a net benefit for considered use cases. We appreciate the EO directive to do so and the work of the OMB in this direction.

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

International collaboration is key, especially among U.S. allies such as the UK, Singapore, and France. Vulnerability and security incident sharing is best raised among allies, especially for those with dedicated AI bodies such as the U.S. and UK AI Safety Institutes.

International standards, including among countries with differing values and regulatory approaches, are needed. An urgent need is disclosure standards, outlining what should be disclosed broadly, which can be fine-tuned per jurisdiction to meet requirements for example in

[144] NIST Risk Management Framework (RMF)
[145] Hugging Face Response to OMB RFC on federal use of AI
[146] Announcing our new Content Guidelines and Policy
[147] Announcing our new Content Guidelines and Policy

adhering to local laws. Evaluation standards can help build cross-cultural and multilingual approaches to model safety.

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

Lessons can be drawn from the EU's established AI Office, which maintains standards and operational definitions as AI evolves, and is empowered to make case-by-case definitions about which models might present a systemic risk to reflect the highly contextual nature of risk assessments. The U.S. should designate an accountable organization for metrics, standards, and requirements, and sufficiently resourcing them to keep pace with AI development. The National Institute of Standards and Technology is well positioned and should be well resourced.

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

As discussed in 5a, no definitive processes or standards exist. In addition to discourse referenced in 5a, ongoing cross-sectoral collaborations are sharing best practices.[148] Strict guidelines about safety by design processes can be helpful for developers as they have control over relevant properties, including what data selection, whether the data covers high risk topics such as nuclear devices, and whether the model is trained to produce malware. Limitations in the state of evaluation development makes quantifying model behavior technically difficult.

As discussed in 5d, models that are trained specifically on PII/malware/hate speech may have different release conditions. Non-regulatory practices such as vulnerability sharing can draw from cybersecurity. In high risk settings, existing controls on information flow should apply.[149] Narrow models for specific domains or applications should use practices from that domain.

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

Research institutions, especially those operating in public interest and outside of commercial product development, should not only have access but also infrastructural support. Government resourcing[150] can not only provide financial and technical support, but also ensure that researchers with sufficient breadth of expertise and external stakeholders have access to technical artifacts that may be used in rights-impacting settings. As noted in 3a, third party auditing is most effective when entities who can conditions without developer engagement.[151] Open-weight models can help create these conditions, along with varying

[148] PAI's Deployment Guidance for Foundation Model Safety

[149] How AI Can Be Regulated Like Nuclear Energy

[150] National Artificial Intelligence Research Resource Pilot | NSF

[151] Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance

access to components such as datasets, including e.g. extensive documentation and query access.[152]

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?

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

Policy decisions can jointly advance interests of innovation, competition, and security. Work across technological settings (e.g. in the contexts of election systems[153] or biosecurity in protein design[154]) has shown that a priori antagonistic goals can often both benefit from openness given appropriately tailored approaches. In specific settings where the tensions are harder to resolve, specificity in how these tensions are managed and narrowly tailored policies are particularly important.[155] Across both settings of widely available weights and API-only development, **extensive documentation, replicable evaluation, and transparency into design choices of FMs all contribute positively to all of these interests.**

b. Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 1026 integer or floating-point operations used in the 17 Executive Order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?

Concerns with thresholds include their robustness, purpose/trigger, enforcement, and verifiability. Methods for determining and proxies for model capability have differing effectiveness. **Thresholds along any singular variable, such as training compute, will not give robust insight to model capability and will adapt with time and incentives.** Additionally, while model evaluations should eventually play a major role in helping assess proper use and governance of foundation models, our scientific understanding of these artifacts and evaluation systems are not currently mature or robust enough to serve this purpose; and will require significant additional investment and access to open models as outlined in our response to 3a.

For determining what potential thresholds should trigger, we recommend starting with a voluntary reporting framework. Rather than red-teaming results, disclosure should include what replicable evaluations have been run, including on a spectrum of risks. This will build regulatory capacity and foster a stronger evaluation ecosystem. Thresholds should not be singular hardlines.[156] **Levels of mandates should apply by contextual use case and mode of**

[152] 📚 Training Data Transparency in AI: Tools, Trends, and Policy Recommendations 🗳️
[153] Transparency versus security: early analysis of antagonistic requirements
[154] Protein design meets biosecurity | Science
[155] Openness versus Secrecy in Scientific Research Abstract - PMC
[156] Drawing Lines: Tiers for Foundation Models

deployment, whether in research or commercial applications. Robust transparency requirements should apply to all commercial products.

Disclosure is critical and lessons can be drawn from cybersecurity to determining levels of disclosure and what can be made public. Meeting a base level requirement should mandate some level of disclosure. Models whose deployment is supported by extensive cloud compute capacity, due to model size or volume of users, should warrant additional scrutiny including cybersecurity audits given the scale of their impact. Robust documentation can also be beneficial for commercial usability and trust.

Conclusion

AI innovation would not be possible without open access and open science. Openness broadly continues to benefit the field. Since the AI field is a fast-evolving space with many arising considerations, expanding scope to many system artifacts can help to better weigh risks and benefits. Many risks apply across model weight availability and tailored threat models can narrow intervention options. Hugging Face greatly appreciates the opportunity to provide insights and we remain eager to support NTIA and provide our expertise moving forward.