stability.ai

Response to Request for Information on Artificial Intelligence from the National Institute of Standards and Technology

February 2024

Introduction

Stability AI welcomes the opportunity to respond to the National Institute of Standards and Technology (NIST) request for information (RFI) under Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. As a leading developer of generative AI models, Stability AI is committed to the safe, open, and responsible deployment of these emerging technologies, and we welcome the ongoing efforts of NIST to improve risk evaluation, mitigation, and assurance in AI systems. The following response shares our emerging perspective on several areas of inquiry, including content provenance, supply chains, risk mitigation, and future standards.

Background

Stability AI is a global company working to amplify human intelligence by making foundational AI technology accessible to all. Today, we develop AI models across a range of modalities, including image, language, audio, and video. 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 openly, sharing our software code along with the billions of distinctive settings or "parameters" that define the model's performance. 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 ethical use licenses.[1]

To date, our models have been downloaded over 100 million times by developers, and nearly 300,000 developers and creators actively contribute to the Stability AI online community.[2] Our family of image models, Stable Diffusion, underpin up to 80 percent of all AI-generated imagery.[3] These models can take a text instruction or "prompt" from a user and help to create a new image. In addition, we develop a suite of language models that can interpret, summarize, or generate text. These include highly capable large language models, compact language models, specialized models for software development, and models for underrepresented languages, including Japanese and Spanish. Our audio model, Stable Audio, generates high-quality soundtracks and was recently listed on the *TIME* Best Inventions of 2023. Building on this experience, we have developed video models that demonstrate new breakthroughs in video generation.[4] Further, we support academic research into scientific applications of AI. Stability AI provides a range of services to help partners customize and deploy our models, sustaining our open research efforts.

[1] See e.g. the Open Responsible AI License (OpenRAIL) for Stable Diffusion, prohibiting a range of unlawful or misleading uses, available here. We use the term "open" to refer to any models with publicly-available parameters.
[2] Figures from Hugging Face and Discord, November 2023.
[3] Everypixel, 'AI Image Statistics', August 2023, available here.
[4] See e.g. Stability AI, 'Improving Latent Diffusion Models', July 2023, available here; Stability AI, 'Stable LM-3B Technical Report', October 2023, available here; Stability AI, 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', November 2023, available here.

Why we develop open models

With appropriate safeguards, open models can help to improve safety through transparency, foster competition in critical technology, and support grassroots innovation in AI:

- **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, develop new mitigations, and correct for bias. By comparison, closed models may not disclose how they are developed or how they operate. Closed models may be comparatively opaque, and risk management may depend on trust in the developer.

- **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 tens of millions of dollars on research and computing.[5] In this way, the economic benefits of AI accrue to a broad community of developers and firms across the United States, not just Silicon Valley.

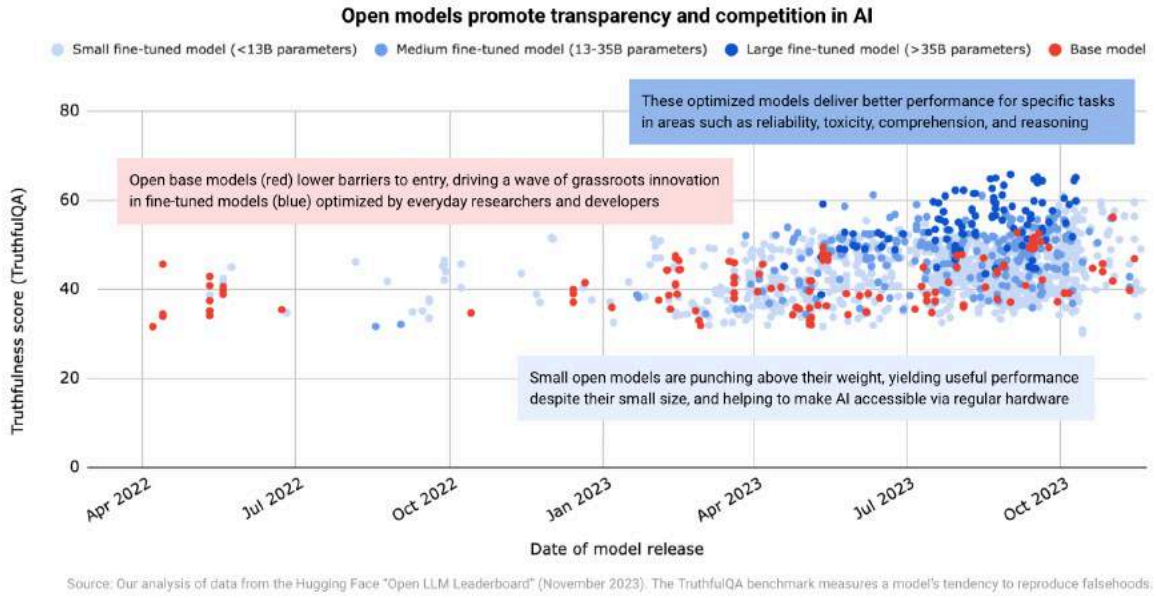
- **Open models drive innovation in safety.** Developers can refine open models for improved safety and performance in specific tasks. For example, open models can be optimized through a range of techniques to mitigate undesirable behavior such as bias, misinformation, or toxicity. These techniques can yield significant improvements in the behavior of a model without requiring extensive computing resources. That means ordinary developers can build safer and more effective models to better support their real-world applications.

- **Open models foster strategic independence.** Open models enable public and private sector organizations to build independent AI capabilities without relying on a handful of firms for foundational technology. They can develop these AI capabilities securely "in house" without exposing their confidential data or ceding control of their distinctive model parameters to third parties. Operational independence will be important for organizations in sensitive or regulated sectors, such as healthcare, finance, law, and public administration.

- **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.

[5] OpenAI disclosed that it cost USD 100 million to train the closed-source GPT-4 model: Wired, 'Open AI's CEO says the age of giant models is already over', April 2023, available here.

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 across America who are helping to build safer AI models and useful AI tools. Open models offer a transparent, competitive, and secure alternative to "black box" technologies owned and operated by a small number of firms.

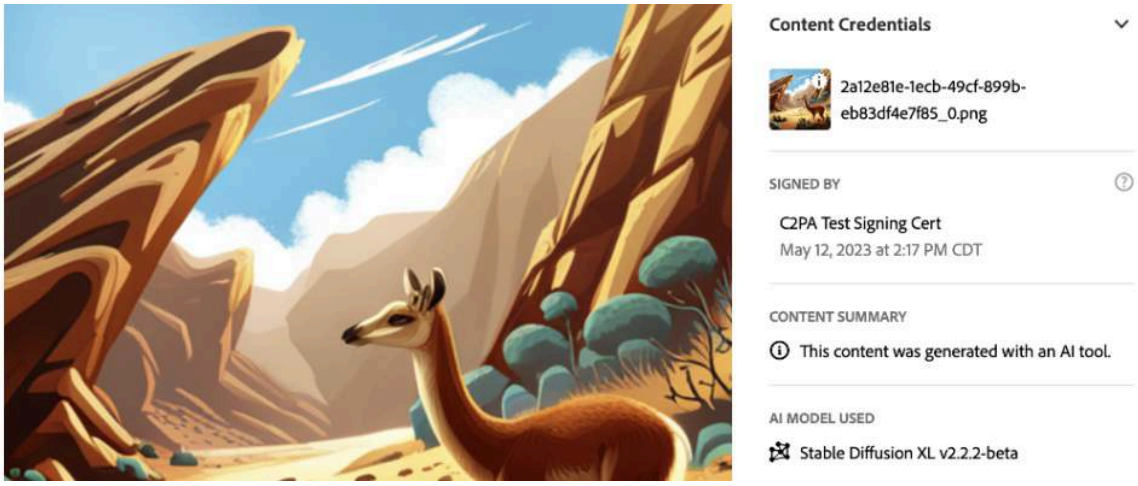


We are committed to the safe development of AI. To that end, we are signatories to the White House *Voluntary AI Commitments* and the British Government's *Joint Statement on Tackling Child Sexual Abuse in the Age of AI*; we participated in the first large scale public evaluation of AI models at DEF CON, facilitated by the White House, and the UK AI Safety Summit; we have testified before the US Senate and UK Parliament; and we continue to engage with agencies in the US and around the world. We focus on building models to support and augment our users, not replace them, and we develop practical AI capabilities that can be applied to everyday tasks. Designing around these principles can help to unlock the useful potential of AI while minimizing the risk of misuse, weaponization, or "runaway" systems.

1. Content authenticity, provenance, and detection
Section 2(a) points 1-3 and Section 1(a)(1) point 6

Recent developments in AI may pose a challenge to the integrity of our information ecosystem. These models can perform a wide range of complex, sensitive, or nonroutine creative tasks. They can amplify bias, errors, or omissions in training data, or they can be misused to generate believable but misleading or abusive content. Deployed AI systems can produce content quickly and on a large scale, which may exacerbate these risks. Stability AI is alert to these challenges, and we are proactively implementing a range of features to improve transparency in the production and dissemination of AI-generated content.

For example, we are implementing content credentials to help users and content platforms better identify AI-generated content. Images generated through our API will be tagged with metadata to indicate the content was produced with an AI tool. In partnership with the Content Authenticity Initiative (CAI) led by Adobe, we are adopting the "Coalition for Content Provenance and Authenticity" (C2PA) standard for metadata.[6] This metadata will indicate the model used to generate an output image. Once the metadata is generated, it will be digitally sealed with a cryptographic Stability AI certificate and stored in the image file. This process uses a C2PA tool to ensure the correct implementation of standards.[7]



Above: An example of content authenticity metadata indicating an image was generated with an AI tool.

In addition, we have implemented an imperceptible watermark for AI-generated content produced through our core API. The watermark is a 48-bit pattern discreetly embedded in pixels. This pattern is distributed across the image to improve the robustness of the watermark to manipulation or removal. Further, we share our open models with watermarking implementations included by default, enabling downstream developers to integrate watermarking in their own API services or user-facing applications.[8] We provide software code to detect these watermarks.

[6] CAI, 'C2PA', available here.
[7] CAI, 'Command Line Tool', available here.
[8] Stability AI, 'Generative Models Repository', available here.

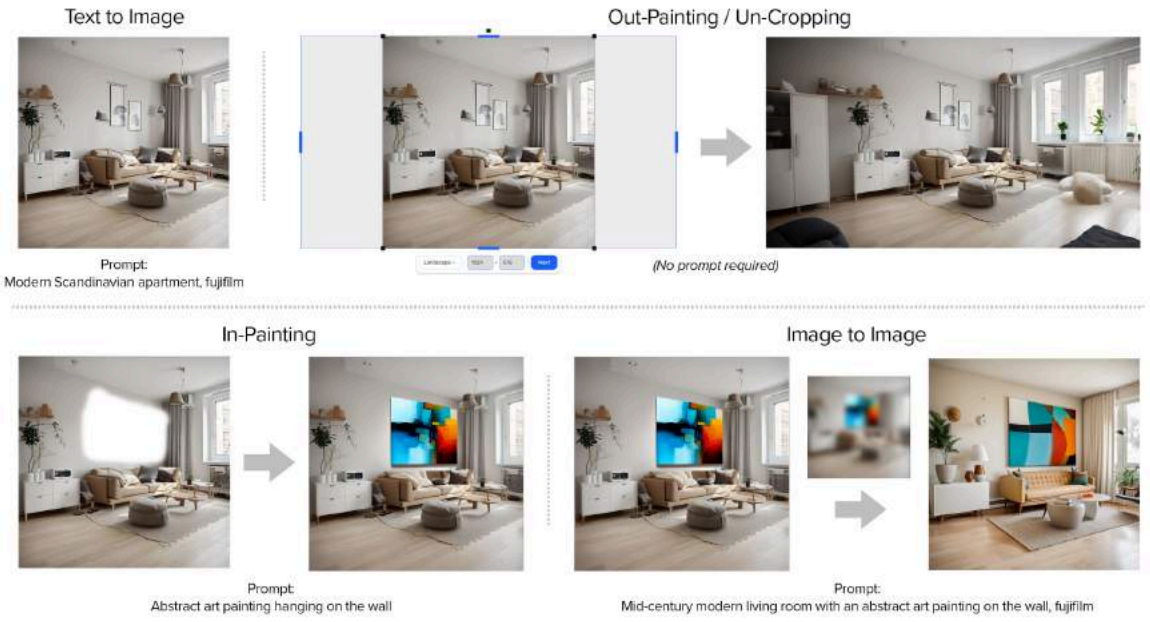
Above left: An image generated through our API. Above right: Pixels (yellowed) embed a 48-bit pattern.

Finally, we expect that deepfake detection technology will play a vital role in helping to identify unmarked AI-generated content. Image models review billions of images to learn the hidden relationships between words, ideas, and fundamental visual features or structures. They can apply this knowledge to help a user generate new content. However, this knowledge is applied imperfectly, and AI-generated content may contain a number of perceptible and imperceptible artifacts indicating that it was generated or edited through AI tools. These artifacts can be identified by other classifier models that can help to detect unseen AI-generated content for purposes such as content disclosure or moderation. We continue to engage with social media platforms and other content intermediaries to support their deployment of these systems.

Above: An AI-generated "Pentagon building in Washington" or "handshake" may appear to be authentic at first glance; on closer inspection, however, the Pentagon has six sides, not five, and the hands may have two thumbs, or an irregular number of fingers. These are small but illustrative examples of how AI-generated images often contain tell-tale artifacts that can be detected by other classifier models.

We expect these measures to complement other content transparency and platform safety mitigations. Downstream intermediaries – such as social media, search, or streaming platforms – can use metadata, watermarks, and other signals to assess the provenance of content before amplifying it through their network. 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 and unverified accounts by default, unless the image has trusted metadata that confirms its origin. Together, these features can help platforms to distinguish AI content; enable users to exercise appropriate care when interacting with AI content; and help to limit the spread of misleading content produced with AI tools.

As the Executive Order acknowledges, the development of standards and practices for content provenance is essential to inform potential requirements for content transparency. Already, there are several bills that seek to impose mandatory labeling requirements on all or certain kinds of AI-generated content. In developing these requirements, we encourage policymakers to acknowledge the range of ways in which AI is used for legitimate purposes. The use of AI does not, by itself, make content misleading, objectionable, or dangerous. In many cases, AI is simply a tool within a creative workflow, and the contribution of AI to the final work may vary. AI may be used to create significant portions of a work, or to edit, augment or transform an existing work in more subtle ways. Today, models like those in the Stable Diffusion family are used for everything from editing photographs to prototyping architectural designs to researching new diagnostic techniques for complex medical disorders.



Above: Image models like the Stable Diffusion family can be used in a range of ways as part of a design workflow. They can help to produce new images based on a text description, fill in or replace parts of an existing image, incrementally extend parts of an existing image, or subtly transform an existing image.

In particular, we urge care in the development of mandatory disclosure or labeling rules that rely on content provenance standards or practices. We support clear rules governing the use of AI in sensitive contexts, such as election campaigns, or the use of a person's physical or vocal likeness for improper or exploitative purposes. We have previously urged Congress and other agencies to strengthen the guardrails for improper use of likeness, such as nonconsensual intimate imagery, election disinformation, and commercial misuse. However, a suite of new mandatory disclosure requirements for all AI-generated content, in all circumstances, could have a chilling effect on legitimate artistic expression and legitimate economic activity. For example, a photograph that has been subtly adjusted for aesthetic purposes using AI tools – such as those commonly found in tools like Google Photos – should not attract that same compliance obligations or liabilities as a work generated *de novo* for illegal or improper purposes.

2. Distribution of responsibility across the supply chain
Section 1(a)(1) point 2

Models are just one component in an AI system such as a chatbot or image generator. The model must be hosted or deployed within a user-facing system in order to analyze or generate content. In that environment, different actors perform different functions, ranging from: training a base model (the raw "engine" that understands complex patterns and relationships within a textual, visual, musical, or scientific dataset); fine-tuning 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. Each actor may have limited visibility or control over downstream activity.

Future standards and risk management frameworks should account for diversity in these supply chains. The relationships between actors in a vertically-integrated, closed-source environment may be different to the relationships in a disaggregated, open-source environment. In particular, the risk profile of an AI system will vary depending on how and where the system is deployed. For example, an AI system deployed in higher-stakes domains such as healthcare, finance, education, or public administration may attract more rigorous obligations than an AI system deployed in a lower-stakes domain such as entertainment, with different requirements for reliability, interpretability, and robustness. One model may be deployed in a range of such environments, and responsibility for risk mitigation and assurance may be shared by different actors in different ways. "One size fits all" frameworks for evaluation and performance could set back open innovation by imposing disproportionate or ill-adapted requirements on every AI system without accounting for their specific risks.

3. Mitigations for emerging risks across the supply chain
Section 2(a) points 6-9

We acknowledge that open models pose unique challenges for certain risks, such as the prevention of misuse. For example, language models may be misused to generate intentional disinformation, exploit software vulnerabilities, or summarize dangerous information. Audiovisual models may be misused to generate misleading or unlawful deepfakes. As with other digital technologies, there are no silver bullets to eliminate the risk of misuse. However, there are layers of effective mitigations that help to make it easier to do the right thing with AI, and harder to do the wrong thing:

- 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 our image-based training data, helping to prevent the model from producing unsafe content. We continue to invest in ongoing efforts to improve the quality and reliability of datasets to mitigate these risks. Following initial pre-training, we evaluate and fine-tune our models to help eliminate undesirable behaviors, such as bias. We disclose known risks and limitations in standardized formats, such as model cards, to help downstream deployers decide on additional mitigations.[9] Our most capable models are subject to ethical use or acceptable use licenses that prohibit a range of unlawful or misleading applications.[10]

- 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 (see above). We include watermarking modules by default in certain model repositories so that developers can easily implement and detect these watermarks.

- As a third line of defense, users are governed by technology-neutral rules – state and federal – that apply to the misuse of AI models (e.g. fraud, abuse, defamation, non-consensual intimate imagery, election interference, intrusion software, or privacy). Where necessary, these can and should be fortified to account for novel types of misuse or increased prevalence of misuse.

- As a fourth line of defense, AI countermeasures can be integrated across the digital economy to detect and defend against misuse. Today, AI models are used to detect unsafe content on social media and identify 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 for AI will become increasingly effective in detecting, intercepting, and remediating various kinds of AI misuse.

No mitigation is watertight, but together, they provide a layered defense to emerging risks. In that environment, we encourage NIST to continue developing holistic frameworks for risk evaluation, mitigation, and assurance. They should account for a variety of mitigations, applied by a variety of actors, in a variety of deployment environments. They should identify outcomes – e.g. transparency, reliability, robustness, or interpretability – but avoid prescribing specific means of compliance, and they should be sufficiently adaptable to respond to emerging evaluation techniques or novel mitigations.

[9] See 'Stable Diffusion V2-1 Model Card' available at https://huggingface.co/stabilityai/stable-diffusion-2-1.
[10] 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.

4. Companion materials for the Risk Management Framework (RMF)
Section 1(a)(1) and Section 1(a)(2)

Stability AI welcomes the development of the RMF for AI. The RMF offers a descriptive framework that can help different actors in the supply chain establish risk management processes across the AI lifecycle. However, the RMF does not identify specific evaluation benchmarks, techniques, or practices, many of which have emerged since the publication of the RMF. Evaluation through standardized testing will be essential to verify that a model operates as expected, and that it demonstrates the required level of reliability and robustness for a particular task. Evaluation will be particularly important for AI deployment in regulated sectors (e.g. finance, healthcare, education) as agencies develop minimum performance requirements, as well as AI deployments in less regulated sectors (e.g. entertainment) where performance requirements may be nascent or indeterminate. We welcome NIST's commitment to study and help standardize different benchmarking techniques, adversarial testing practices ("red-teaming"), and human evaluation processes. We encourage NIST to ensure that all modalities are represented in these studies – image, video, and audio in addition to language models – and to release these studies as companion materials to the RMF. Standardized evaluation will help to provide confidence that AI systems deliver the expected or required performance in any deployment environment.

Conclusion

Stability AI welcomes the opportunity to share our experiences with NIST. We encourage NIST to ensure that future standards account for diversity in the AI ecosystem, from grassroots innovation in open models to closed-source products from corporate labs. We support the efforts of the Biden Administration to resource and accelerate NIST's work in these vital areas, and we look forward to continued engagement.