stability.ai

**Comment on Petition for Rulemaking on AI**
**Federal Election Commission**

**October 2023**

Stability AI welcomes the opportunity to comment on the Petition for Rulemaking ("Petition") to the Federal Election Commission in relation to the use of AI-generated content in campaign advertisements and communications. As a leading developer of generative AI models, Stability AI is dedicated to the safe, open, and responsible deployment of these emerging technologies. In particular, we share the FEC's commitment to mitigating the harmful effects of electoral misinformation and disinformation across the content ecosystem.

To that end, in July, Stability AI testified at a hearing of the Senate Judiciary Subcommittee on Intellectual Property to emphasize the importance of helping users and platforms distinguish AI-generated content.[1] Among other things, we urged clear guardrails in relation to the use of a person's likeness for improper purposes. In August, Stability AI participated in a groundbreaking initiative to evaluate AI models through community-led testing.[2] In September, the White House announced that Stability AI had joined the Administration's Voluntary AI Commitments, which include a number of measures to promote transparency in the dissemination of AI-generated content.[3]

We support public scrutiny of these important issues, and we are pleased to share our experiences and perspectives to date. While we believe that existing laws account for the kinds of AI misuse contemplated in this petition, we take this opportunity to outline the technical and non-regulatory steps that Stability AI is taking in response to emerging concerns. As the FEC considers future rulemaking, it is essential that any future rule is targeted, technology-neutral, and accounts for the range of ways in which AI can be used for legitimate purposes.

## Background

Stability AI is a global company that aims to unlock humanity's potential by making foundational AI technology accessible to all. Today, Stability AI develops a variety of generative AI models across different modalities, including image, language, audio, and video. These models are essentially software programs that analyze vast datasets to learn the hidden relationships between words, ideas, and fundamental textual or visual features. Such models are commonly described as "generative" AI because they can apply this knowledge to help a user generate new content.

With appropriate safeguards, we release many of these models openly along with the distinctive settings or "parameters" that define the model's performance. That means developers and researchers can freely integrate or adapt our models to develop their own AI models, build their own AI tools, or start their own AI ventures, subject to ethical use licenses.[4]

[1] Senate Judiciary Committee, Subcommittee on Intellectual Property, July 2023, available here.

[2] White House, 'Administration Announces New Actions to Promote Responsible AI Innovation', May 2023.

[3] White House, 'Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI', September 2023.

[4] See e.g. our OpenRAIL license for Stable Diffusion, which prohibits a range of unlawful, misleading, or exploitative uses of the model, 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.

1. **Promote safety through transparency**. AI models will form the backbone of our digital economy, and everyone should have a voice in their design. By releasing open models, researchers, authorities, and developers can "look under the hood" to verify the performance and suitability of these models. They can identify potential risks, develop new interpretability techniques, and help implement new mitigations. Because these models are auditable, firms and agencies in sensitive sectors can build on open models to produce their own specialized models for critical tasks.[5]

2. **Foster competition in critical technology.** Developing a generative AI model typically requires significant resources. Open models can lower these barriers to entry, fueling competition in AI.[6] Using open models, creators and developers can deploy new AI tools or launch new AI ventures without spending millions on research and computing power. They can participate in this new industrial revolution as builders – not just consumers – of AI technology, and they can do so without relying on a handful of firms for critical infrastructure.

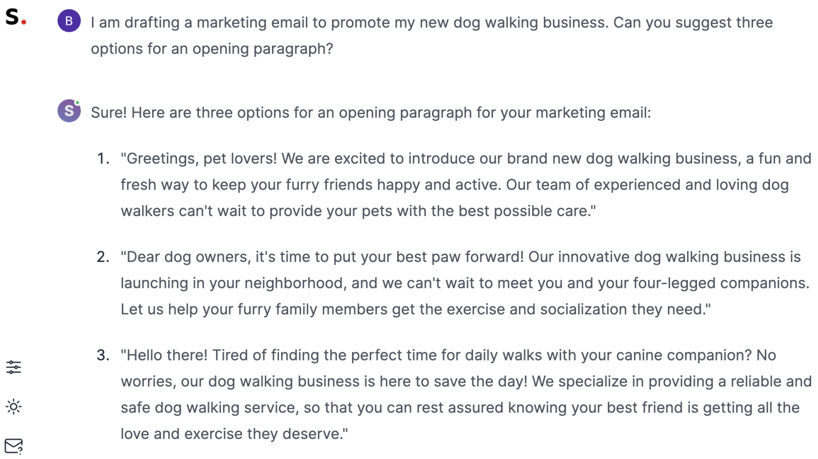
3. **Support grassroots innovation.** Grassroots innovation by anyone, anywhere is one of America's greatest assets, and open models put these capabilities in the hands of everyday creators, developers, and researchers. Everyday people can experiment with open models to develop new and innovative applications that support their work and serve their community. In this way, open models can help distribute the economic benefits of AI across the United States, beyond Silicon Valley.

We are focused on building models to support and augment our users, not replace them. We develop practical AI capabilities that can be applied to everyday tasks – not a quest for artificial superintelligence. Designing around these principles can help to unlock the useful potential of AI while minimizing the risk of misuse, weaponization, or "runaway" systems.

[5] For example, a regulated financial institution may customize AI models to assist in analysis, decision making, or customer support. The financial institution may need to audit the performance of the model for reliability; train the model without exposing sensitive customer data to third-parties; and retain full control over the AI model without relying on a third-party provider. By building on open models, a financial institution can train and manage their own AI system.

[6] See, e.g. the Hugging Face 'Open LLM Leaderboard' comparing open language models, available here.

In 2022, we took over the exclusive development of Stable Diffusion, an open image model that takes a text instruction or “prompt” from a user and helps to produce new images. By some measures, developer interest in Stable Diffusion has grown faster than many open-source software projects in recent history, and over 12 of the 15 billion images generated with AI in the past 18 months may have been produced with Stable Diffusion.[7]



B I am drafting a marketing email to promote my new dog walking business. Can you suggest three options for an opening paragraph?

S Sure! Here are three options for an opening paragraph for your marketing email:

1. "Greetings, pet lovers! We are excited to introduce our brand new dog walking business, a fun and fresh way to keep your furry friends happy and active. Our team of experienced and loving dog walkers can't wait to provide your pets with the best possible care."
2. "Dear dog owners, it's time to put your best paw forward! Our innovative dog walking business is launching in your neighborhood, and we can't wait to meet you and your four-legged companions. Let us help your furry family members get the exercise and socialization they need."
3. "Hello there! Tired of finding the perfect time for daily walks with your canine companion? No worries, our dog walking business is here to save the day! We specialize in providing a reliable and safe dog walking service, so that you can rest assured knowing your best friend is getting all the love and exercise they deserve."

*Above left: Image prompted by “photograph of an astronaut riding a pink horse in space”. Above right: Language models can be used to power a range of creative, analytic, or coding tools. For example, they can help to draft or edit documents, analyze text, or help to identify bugs in software code.*

In early 2023, we released the first in a series of open language models to support research into AI safety, performance, and accessibility. These language models can take a prompt from a user and help to produce new passages of text or software code. Our research outputs include highly-capable “fine-tuned” language models that demonstrate new optimization techniques (Stable Beluga); lightweight “base” language models to help make AI more efficient and accessible for real-world tasks (Stable LM);[8] specialized language models to support software development (Stable Code); and models for underrepresented languages, including the highest-performing open Japanese model (Japanese Stable LM).[9]

In September, Stability AI developed an audio model, known as Stable Audio, that can help a user to produce high-quality soundtracks of ~90 seconds length.[10] Stable Audio was trained on 19,500 hours of music, or 800,000 soundtracks, obtained through our partnership with AudioSparx, a leading content library.

[7] Everypixel, ‘AI Has Already Created As Many Images As Photographers Have Taken in 150 Years’, August 2023, available here.
[8] See e.g. Stability AI, ‘Stable LM-3B Technical Report’, October 2023, available here.
[9] A base model is an AI model that is trained to understand the hidden relationships within vast datasets of text. A specialized model is an AI model that is optimized with specific data and targeted adjustments for better performance on specific tasks. An application is a software program that uses an AI model to help end-users perform a task (e.g. a chatbot).
[10] Stability AI, ‘Stable Audio’, September 2023, available here.

Over 200,000 creators and developers actively contribute to the Stability AI community. In addition, Stability AI partners with organizations to adapt these models for specific purposes, helping to sustain our open research and development efforts. Stability AI provides computing services so that developers and users can access the powerful computing resources necessary to train or run these models, and actively supports research into scientific applications of AI.[11]

### There are layers of mitigation for content transparency

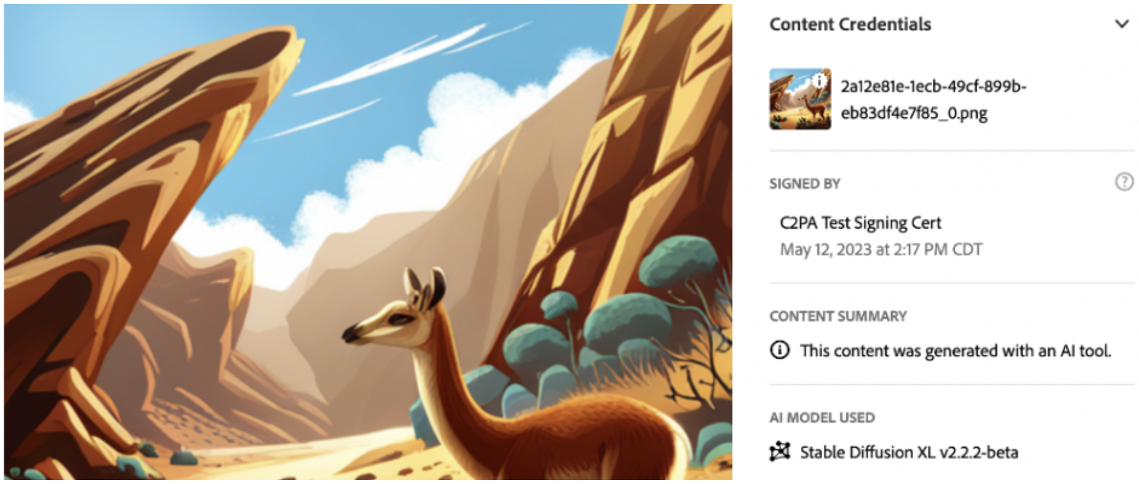
Recent developments in AI may pose a challenge to the integrity of our information ecosystem. These models can perform a wide range of complex, sensitive, or nonroutine creative tasks. They can amplify bias, errors, or omissions in training data, or they can be misused to generate believable but misleading or abusive content. 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 mitigate the spread of unintentional misinformation and intentional disinformation. Together, these mitigations can provide a layered defense to emerging risks.

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.[12] 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.[13]

[11] See e.g. MedARC, supported by Stability AI, ‘Reconstructing the Mind’s Eye: fMRI to Image’, 2023, available here.

[12] CAI, ‘C2PA’, available here.

[13] CAI, ‘Command Line Tool’, available here.



Content Credentials

2a12e81e-1ecb-49cf-899b-eb83df4e7f85_0.png

SIGNED BY

C2PA Test Signing Cert

May 12, 2023 at 2:17 PM CDT

CONTENT SUMMARY

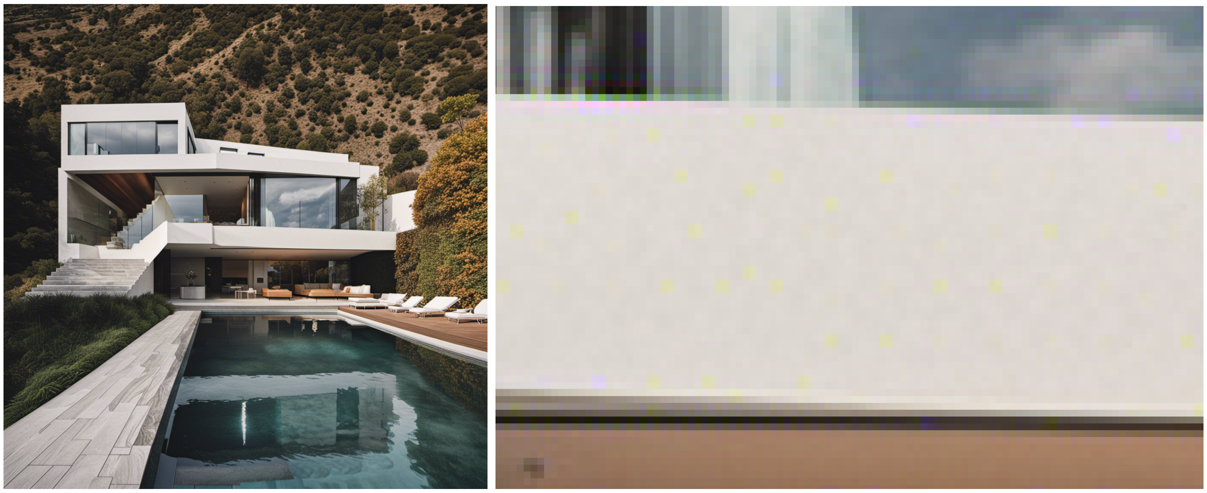
ⓘ This content was generated with an AI tool.

AI MODEL USED

Stable Diffusion XL v2.2.2-beta

*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 API.[14] 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 demonstrations implemented by default, enabling downstream developers to integrate watermarking in their own API services or AI applications. We provide software code to detect these watermarks.

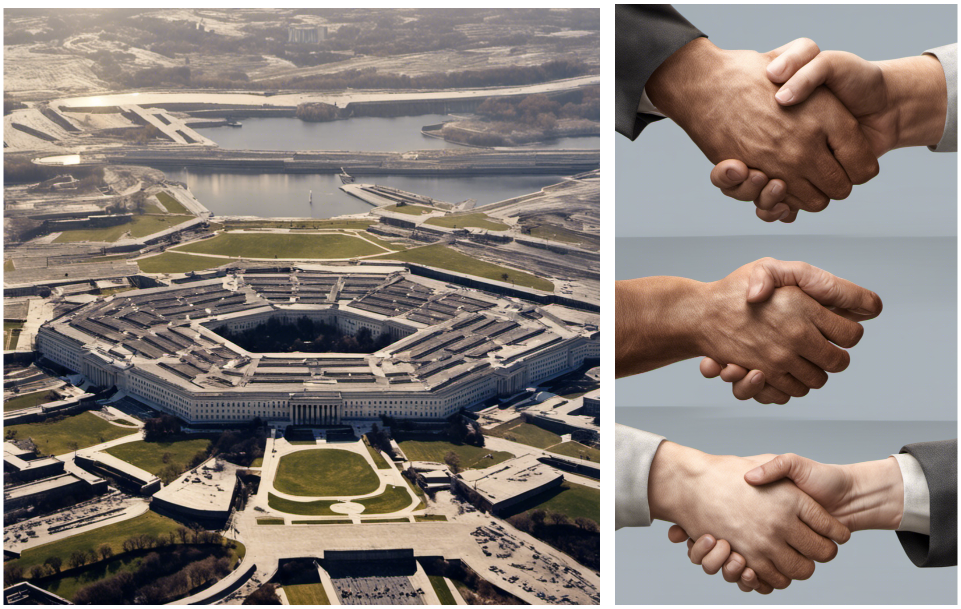


*Above left: An image generated through our API. Above right: Pixels (yellowed) embed a 48-bit pattern.*

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

[14] Stability AI, 'Generative Models Repository', available here.

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.



*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 contain tell-tale signs that can be detected by other classifier models.*

**These mitigations can help the information ecosystem respond to misleading AI content**

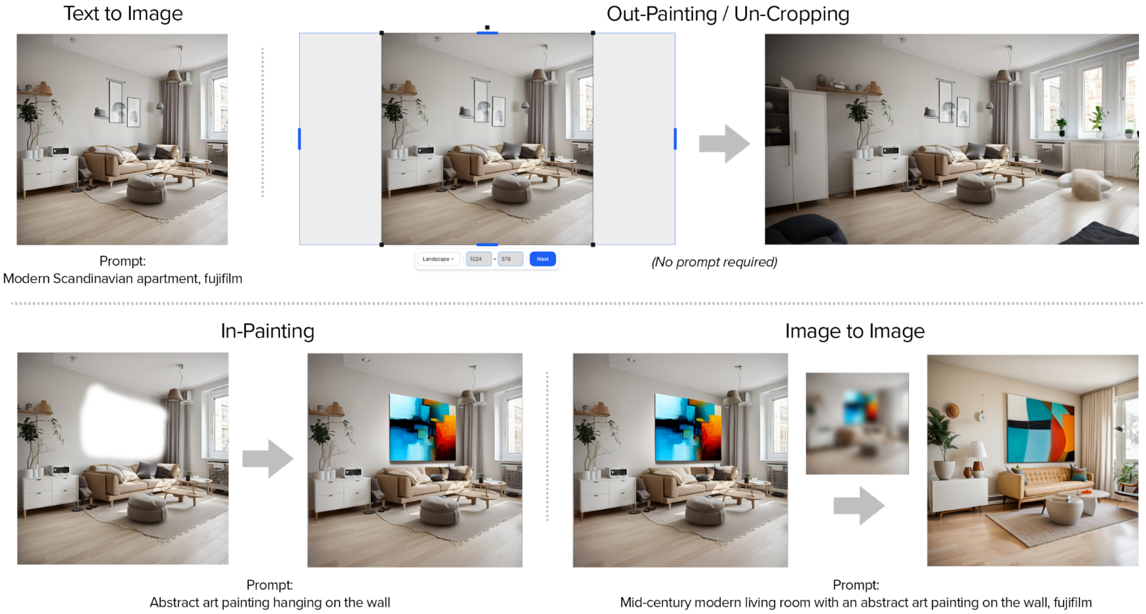
There are no "silver bullets" to prevent the misuse of AI. However, by taking a layered approach to mitigation, these measures can help to identify and limit the spread of misleading content. Downstream intermediaries – such as social media or streaming platforms – can use metadata, watermarks, and other 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.

Importantly, these mitigations are additional to existing user-based or conduct-based restrictions. These include existing federal and state laws governing fraud, abuse (e.g. non-consensual intimate imagery), rights of publicity, and provisions such as 52 USC 30124 directed at fraudulent misrepresentation.

### Content restrictions or disclosure requirements should be targeted and technology-neutral

In developing future requirements for AI content transparency, 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 Stable Diffusion are used for everything from editing photographs to prototyping architectural designs to researching new diagnostic techniques for complex medical disorders.


Text to Image
Prompt:
Modern Scandinavian apartment, fujifilm
Out-Painting / Un-Cropping
Landscape
1024
×
576
Next
(No prompt required)
In-Painting
Prompt:
Abstract art painting hanging on the wall
Image to Image
Prompt:
Mid-century modern living room with an abstract art painting on the wall, fujifilm

*Above: Image models like Stable Diffusion 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.*

To that end, we urge care in the development of mandatory disclosure or labeling rules. 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. In these scenarios, the use of likeness can be problematic if it wrongfully implies a person's endorsement of, affiliation with, or promotion of a work or idea. The improper use of personal likeness should be governed by clear rules that define impermissible use. We believe the Federal Election Campaign Act has already established clear and technology-neutral rules to that effect.

If the FEC determines that 52 USC 30124 or 11 CFR 110.16 require further clarification, we encourage a targeted and technology-neutral approach. An overbroad and overinclusive definition of “fraudulent misrepresentation”, or 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 platforms like Google Photos – should not attract that same compliance obligations or liabilities as a work that features an entirely fictitious representation of a candidate.

**Conclusion**

Stability AI welcomes the focus on content transparency in relation to new generative AI tools. We encourage policymakers to take stock of existing measures taken by industry to respond to these concerns, and to ensure that future rules – if required – are targeted, technology-neutral, and account for the legitimate use of AI tools.