Stability Vulnerability Disclosure Policy

Program Scope

Name	Description	Scope
Stability.AI infrastructure	Wildcard domain which includes all public-facing services and subdomains operated by Stability AI, covering interfaces, APIs, model endpoints, and documentation platforms.	https://*.stability.ai
Dreamstudio	Stability AI's official user interface for interacting with our generative AI models, primarily Stable Diffusion. It allows users to generate and edit images through a graphical frontend backed by our production API.	https://dreamstudio.stability.ai/
Dreamstudio API	The Dreamstudio API is used for backend requests for the Dreamstudio web application	https://dreamstudio-api.stability.ai/
Stable Audio	This application enables users to create original, high-quality music and sound effects from text prompts or audio input using the stable audio generative AI model.	https://*.stableaudio.com/
Stability.AI Platform API	This endpoint is the primary production API for accessing Stability AI's suite of generative models, including text-to-image, image-to-image, and audio. It supports authenticated access via API keys, and serves as the backend for partner integrations.	https://api.stability.ai/

Rules of Engagement

To ensure that good-faith security research is clearly distinguished from malicious activity, participants must follow these rules:

Authorized Testing

- You are authorized to conduct security testing only within the bounds defined by this policy.
- You must comply with this policy and all applicable agreements. In the event of any conflict, this policy takes precedence.

Responsible Disclosure

- All vulnerabilities must be reported promptly through our official Bugcrowd program.
- Do not disclose any information about a vulnerability to the public or third parties without explicit written permission from our security team. We aim to authorize public disclosure within 90 days of initial report receipt.

Prohibited Actions

- Do not engage in activities that compromise privacy, disrupt services, destroy data, or degrade the user experience.
- Do not access, alter, or use data that does not belong to you. This includes any confidential data exposed as a result of your testing. If such access occurs, immediately stop all testing, report the incident, and delete all related data.
- Extortion, threats, or any attempts to elicit a response under duress are strictly prohibited. Submissions involving coercion will not qualify for Safe Harbor protections.

Out-of-Scope and Prohibited Testing

The following types of activities are strictly prohibited and considered out of scope:

- Any test or technique that degrades availability or performance of Stability AI services *without exposing a security flaw*
- **Denial-of-Service (DoS) or Distributed Denial-of-Service (DDoS) attacks**

- **Social engineering of any kind**, including phishing, impersonation, or manipulation of employees or contractors
- Physical security testing, including unauthorized access to facilities or hardware
- Vulnerabilities in external systems or code outside of Stability AI's control
- Automated vulnerability scanning or brute-force attacks that degrade service availability
- Attacks targeting or exploiting **open source model weights or checkpoints**, including inference behavior or re-hosting vulnerabilities outside of deployed production services
- Use of stolen/leaked credentials
- Impersonation on third-party platforms
- **General Best-Practice & Informational Findings**: Issues that do not pose a meaningful threat to security or privacy, often highlighting recommendations rather than concrete vulnerabilities.

Scope Compliance

- Only systems and assets explicitly listed as in-scope are authorized for testing.
- Interact only with accounts and assets that you own or are explicitly permitted to test.

Model-Related Reports

Model Behavior Is Not a Security Bug

Model safety issues are not typically isolated, fixable software bugs. They often require broader research efforts and do not fit within the traditional framework of a bug bounty program. As such, model behavior issues are not eligible for rewards through this program.

To ensure effective review and prioritization, these issues should be reported via the appropriate feedback channels, not through the bug bounty process.

Out-of-Scope Model Issues

The following categories are explicitly out of scope and will not be eligible for bounty rewards:

- Prompt injection or jailbreak techniques (e.g., DAN-style prompts)
- Offensive or inappropriate model responses
- Instructions resulting in harmful, unethical, or malicious content generation
- Model hallucinations (e.g., fabricated outputs, code, or secret information)