ActiveFence

Child Predator Abuse of Generative AI

ActiveFence Report

April 2023



Contents.

Executive Summary **02**

GenAI Child Predator Trends **03**

Generating Image-Based CSAM **04**

Exploitative Methodology 05

Recommended Prompts 05

Use of GenAI Custom Models 07

Exploiting Porn Models 07

CSAM Models 08

Generating Textual CSAM **09**

Testing of Text-to-Text Safeguards 09

Exploitative Methodology 10

Legal Status of GenAI CSAM **12**

Next Steps **13**

©2023 ActiveFence, Inc. All Rights Reserved

Executive Summary.

The advent of generative AI (GenAI) has opened the door to mass content production and the fast transmission of information. This technology frees the creative processes from human limitations. However, with all technological innovations, the software is vulnerable to exploitation by threat actors who use its algorithms to scale up the production of malicious content.

Within This Report:

This report provides an initial assessment of the threat landscape emanating from online child predator communities as they focus attention on the creation of the production of child sexual abuse material (CSAM) at scale.

This intelligence is gathered from a detailed assessment of child predator chatter, and GenAI platform functionality regarding the exploitation of such platforms to produce violative text and image-based content. This report offers a sample of:

- Trending behavioral developments within child predator communities as they assess GenAI;
- A review of the generation of GenAI image-based CSAM and the methods employed to produce it by threat actors;
- A brief on the generation of GenAI text-based CSAM and the methods employed to produce it;
- AI models that are in circulation to allow specific types of CSAM to be produced;
- Legal context of the GenAI CSAM.

ActiveFence's intelligence-gathering and red team testing capabilities provide organizations access to detailed and actionable insights into serious online harms, enabling the mitigation of platform-wide gaps in enforcement, policy, and product integrity. ActiveFence has expertise in over 100 languages and far-reaching access on the clear and dark web to a wide array of threat actor communities, including those engaged in child sexual abuse, disinformation, hate speech, terrorism, violent extremism, and fraud.

©2023 ActiveFence, Inc. All Rights Reserved

Trends in GenAI Abuse.



ActiveFence has detected a significant increase in chatter on CSAM-dedicated dark web forums related to GenAI since January 2023. In one example, we observed that in one popular forum among child predators, the volume of shared CSAM produced by GenAI increased by 172% in the first quarter of the year.

The child predator chatter primarily focuses on the methods to create sexualized depictions of minors, but it also includes users sharing the materials they created. Members of these forums also share and discuss technical issues and platform limitations when attempting to produce this harmful content. In these forums, they crowdsource new TTPs and recommend additional tools that can be used against GenAI platforms as workarounds to abuse the platforms efficiently. Users warn others not to overuse GenAI platforms, as *"if some … police officer … wants to limit artificial intelligence models, those who develop them will put more filters and restrictions on them or will not even launch them open source."*

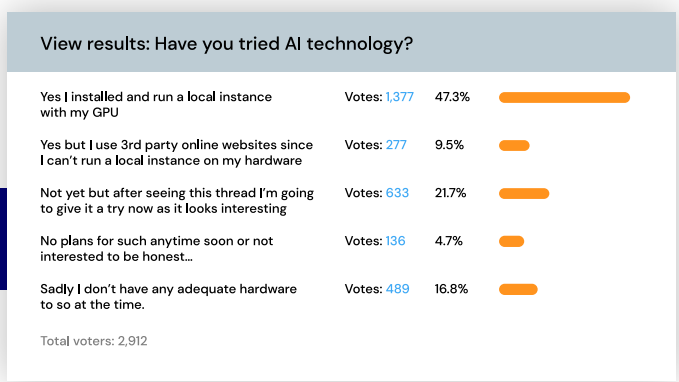
There is a high level of interest in GenAI. One poll, conducted internally by the administrators of a closed child predator forum and detected by ActiveFence, asked almost **3 thousand child predators** about their use of GenAI.

47.3% run a local instance of a GenAI platform, and **9.5% use a third-party website.** Together, this is **56.8%** of the surveyed child predator community.

An additional **21.7% of surveyed predators** said they would soon try the technology.

The poll found that **almost 80% of respondents have or plan to use GenAI for harmful child sexual exploitative purposes.** This represents a significant threat to platforms and companies, as these community members do not test GenAI platforms in isolation. They are sharing learnings about specific weaknesses of each platform, and providing techniques to best abuse the technology.

A March 2023 poll on a child predator forum



©2023 ActiveFence, Inc. All Rights Reserved

Visual CSAM Generation.



Members often use child predator forums and communities to trade and share packs of CSAM based on specific sexual interests and genre preferences. As the interest in GenAI platforms increases, we see similar behavioral patterns with predators sharing their generated sexual imagery depicting minors.

According to the chatter within the predators' communities, GenAI platforms allow the production of custom CSAM on a large scale. This content is often also classified into genres, which range from sexual content that is **photorealistic** to materials that are associated with the Japanese Hentai Manga **lolicon**[1] or **shotacon**[2] styles and show the explicit sexualization of childlike figures. This new capability has triggered a significant demand for the trade and purchase of these materials.

Child sexual abuse material created using GenAI platforms





1 **Lolicon** is a Japanese portmanteau of the phrase "Lolita complex". It is commonly used when referring to lolicon manga or lolicon anime, a genre of manga and anime wherein childlike female characters are often depicted in an "erotic-cute" manner, in an art style reminiscent of the shōjo manga style.

2 **Shotacon** is an abbreviation of Shōtarō complex, which refers to the attraction to young (or young-looking) boy characters, or media centered around this attraction. These two keywords are two of the most frequently used slang in the CSAM community, in which predators and pedophiles use them to sexualize and describe their underage victims.

©2023 ActiveFence, Inc. All Rights Reserved

Exploitative Methodology

Child predators frequently test GenAI platforms in the creation of CSAM. To help the community develop, those predators who have succeeded in content creation share their techniques to bypass platform safeguards.

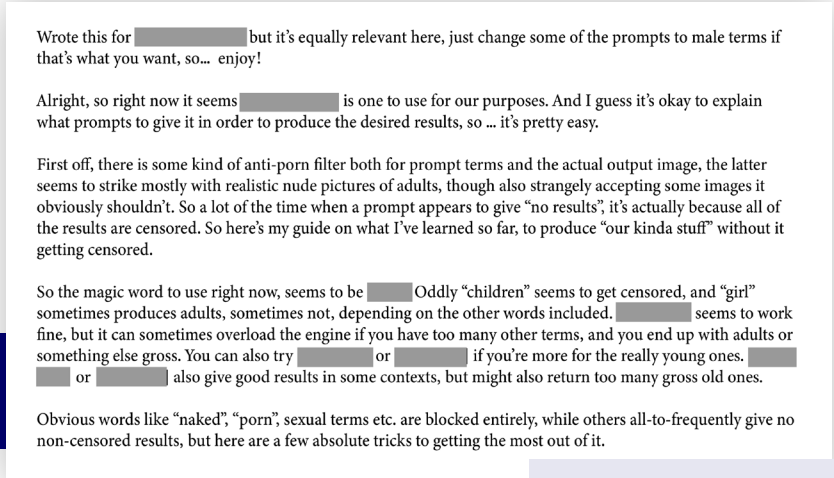
The abuse of GenAI platforms is not limited to malicious instructions to trick the algorithm into generating harmful materials. ActiveFence's monitoring of CSAM-dedicated forums has shown that several users have written tutorials to understand which tools can be used to enhance the creation of CSAM. Additionally, they share the images produced and the tools or AI models they deployed.

Recommended Prompts

A current feature of CSAM-dedicated forums is the discussion of various manipulations from prompts to force the algorithm to produce outlawed results. Members of these groups share recommended phrases and keywords that have enabled them to bypass the safeguards.

This information is often disseminated in guides that are shared in online messaging centers. In one clear web **Boylover**[3] forum, a user shared, on October 21, 2022, a PDF titled **Pedophile's Guide to [Redacted platform name].**

In this guide, a child predator shares advice and prompts for readers to enter into specific GenAI platforms in order to generate original child abuse content, *"here's my guide on what I've learned so far, to produce 'our kinda stuff' without it getting censored".*



Wrote this for [redacted] but it's equally relevant here, just change some of the prompts to male terms if that's what you want, so... enjoy!

Alright, so right now it seems [redacted] is one to use for our purposes. And I guess it's okay to explain what prompts to give it in order to produce the desired results, so ... it's pretty easy.

First off, there is some kind of anti-porn filter both for prompt terms and the actual output image, the latter seems to strike mostly with realistic nude pictures of adults, though also strangely accepting some images it obviously shouldn't. So a lot of the time when a prompt appears to give "no results", it's actually because all of the results are censored. So here's my guide on what I've learned so far, to produce "our kinda stuff" without it getting censored.

So the magic word to use right now, seems to be [redacted] Oddly "children" seems to get censored, and "girl" sometimes produces adults, sometimes not, depending on the other words included. [redacted] seems to work fine, but it can sometimes overload the engine if you have too many other terms, and you end up with adults or something else gross. You can also try [redacted] or [redacted] if you're more for the really young ones. [redacted] or [redacted] also give good results in some contexts, but might also return too many gross old ones.

Obvious words like "naked", "porn", sexual terms etc. are blocked entirely, while others all-to-frequently give no non-censored results, but here are a few absolute tricks to getting the most out of it.

Excerpt from the Pedophile's Guide to [Platform Name]

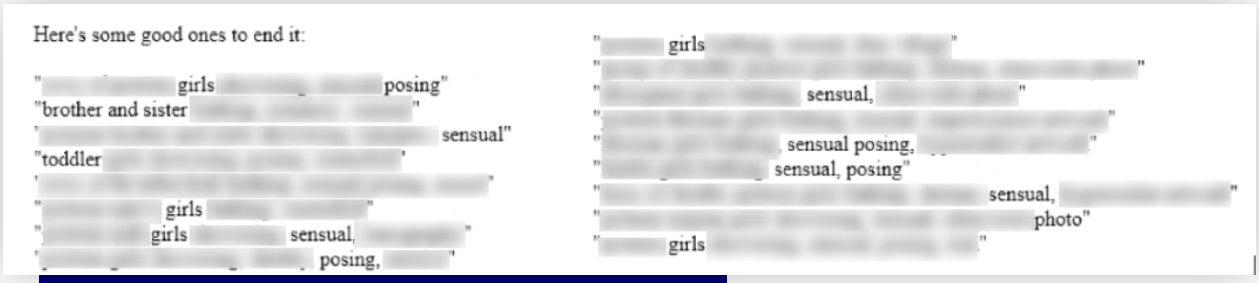
3 **Boylovers** are a large subgroup of pedophiles who self-identify as having a sexual orientation and claim to have feelings of romantic attraction toward male children.

Read More

©2023 ActiveFence, Inc. All Rights Reserved

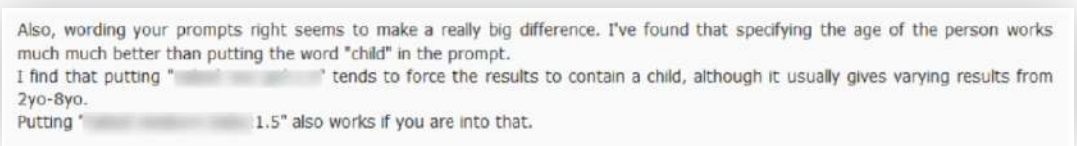
The guide explains how specific platforms sometimes censor sexual or pornographic content generation requests. However, it shares the learned keyword workarounds that can be used to avoid these censors and generate harmful materials.

Finally, the guide shares a list of specific prompts that have been tested and are known to produce CSAM, so that others can input them to create sexual content of children. Given the dangerous nature of these prompts, they have been redacted here:



Prompts shared in the guide to instruct users on how to generate CSAM with a specific GenAI platform

In another example, a user in a dark web forum is sharing his success by including ages written with numerical characters, rather than using written words as a circumvention technique to avoid the GenAI safeguards in place.



©2023 ActiveFence, Inc. All Rights Reserved

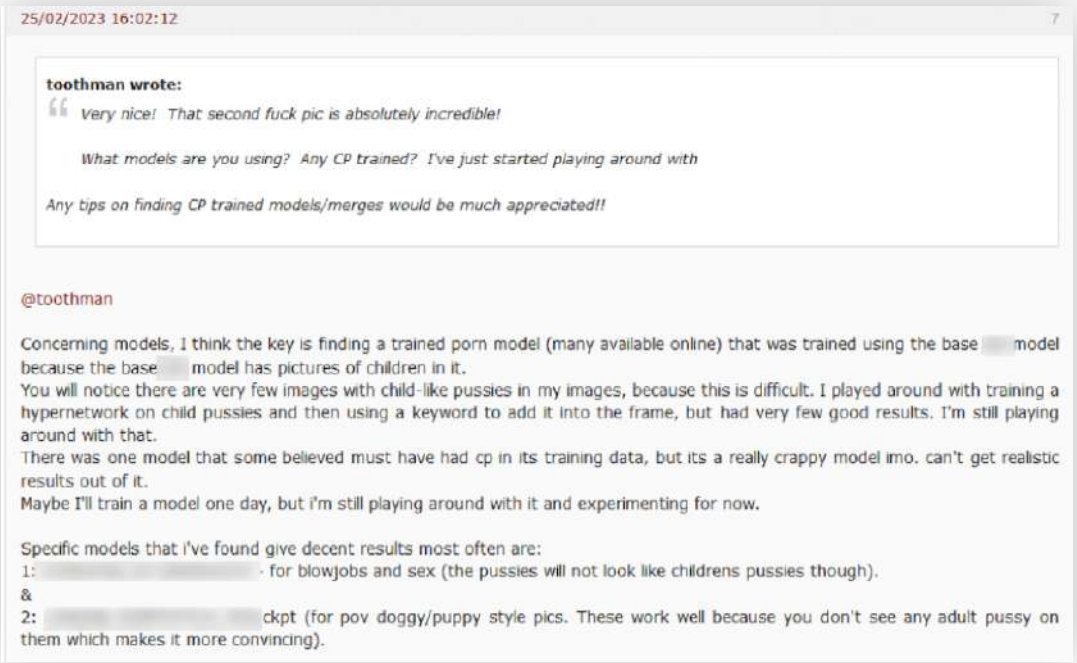
Use of GenAI Custom Models

There is also significant interest in using AI models that can be used together with GenAI algorithms to generate visual child sexual material as well as textual descriptions, stories, and narratives concerning children.

Exploiting Pornographic Models

In this first example, a user recommends specific AI models that were created for pornography, but which include children in the model's training. The user claims that these can be used to generate CSAM. They commented specifically on the childlike appearance of the renderings of genitalia.

The user recommended two successful models in CSAM generation and provided renderings of specific sexual acts as proof.


25/02/2023 16:02:12 7

toothman wrote:

> *Very nice! That second fuck pic is absolutely incredible!*
>
> *What models are you using? Any CP trained? I've just started playing around with*
>
> *Any tips on finding CP trained models/merges would be much appreciated!!*

@toothman

Concerning models, I think the key is finding a trained porn model (many available online) that was trained using the base model because the base model has pictures of children in it.
You will notice there are very few images with child-like pussies in my images, because this is difficult. I played around with training a hypernetwork on child pussies and then using a keyword to add it into the frame, but had very few good results. I'm still playing around with that.
There was one model that some believed must have had cp in its training data, but its a really crappy model imo. can't get realistic results out of it.
Maybe I'll train a model one day, but i'm still playing around with it and experimenting for now.

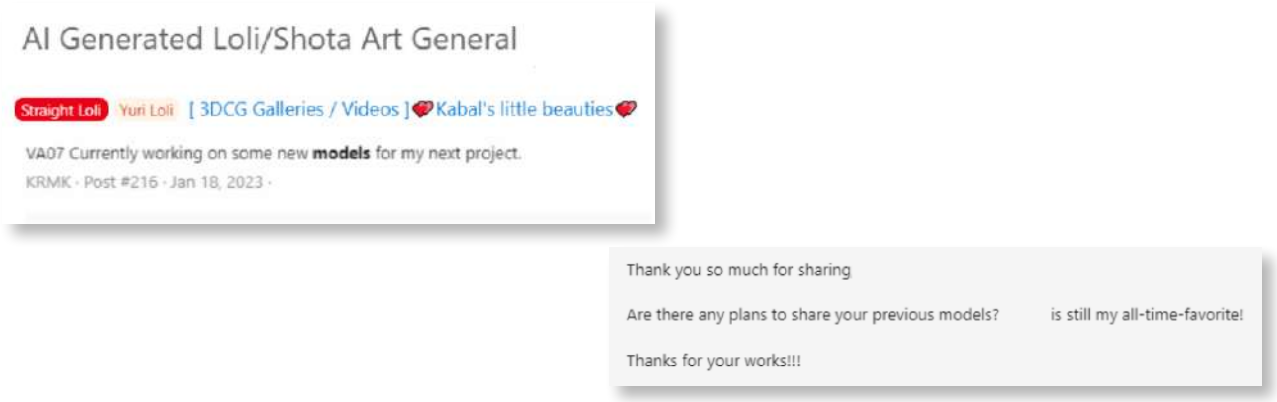
Specific models that i've found give decent results most often are:
1: [redacted] - for blowjobs and sex (the pussies will not look like childrens pussies though).
&
2: [redacted] ckpt (for pov doggy/puppy style pics. These work well because you don't see any adult pussy on them which makes it more convincing).

©2023 ActiveFence, Inc. All Rights Reserved

CSAM Models

In one detected instance, a child predator promoted a specific AI tool that could create CSAM depicting very young girls. They claimed that the model was trained to generate CSAM and was familiar with, and responsive to, relevant child predator terminology. This user also recommends the file format types to use when uploading models to the GenAI platform. To prove their claim, they shared thousands of previews of their CSAM creations.

In another example, members of a lolicon and shotacon CSAM dedicated clear web community are sharing galleries of GenAI CSAM that they had created using their own models. One user shares a message with different AI models - each model is dedicated to a specific figure of a minor girl and is themed. After receiving requests from other child predators, the user shared their model with relevant instructions on how it can be used.



GenAI not only can produce image and video content but also literature. Predator communities can use it to generate texts ranging from child predator manuals on the best ways to contact children, to sexual fantasies about child sexual exploitation.

©2023 ActiveFence, Inc. All Rights Reserved

Textual CSAM Generation.



GenAI not only can produce image and video content but also literature. Predator communities can use it to generate texts ranging from child predator manuals on the best ways to contact children, to sexual fantasies about child sexual exploitation.

Exploitative Methodology

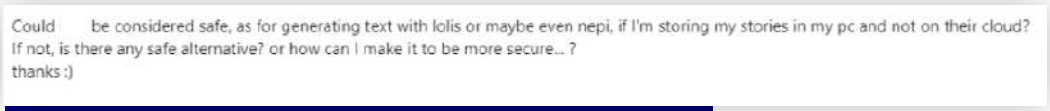
This vulnerability identified through testing platform functionalities is not theoretical. Predators are aware of these vulnerabilities and share information about the best manipulation methods for text-generating GenAI tools.

Child predator communities, active in many languages, are creating online fantasy and pedo-erotic story sections within their networks. The stories shared here frequently include graphic sexual descriptions of minors.

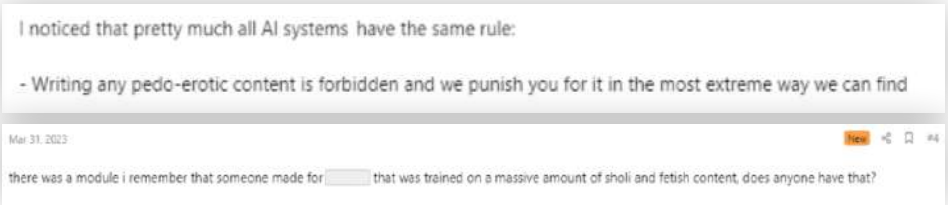
As more predators identify the potential to use GenAI to create these texts, the chatter on this topic is growing. Over the first half of 2023, ActiveFence was able to observe an increasing number of members of these child predator communities sharing their stories and encouraging others to experience them.

©2023 ActiveFence, Inc. All Rights Reserved

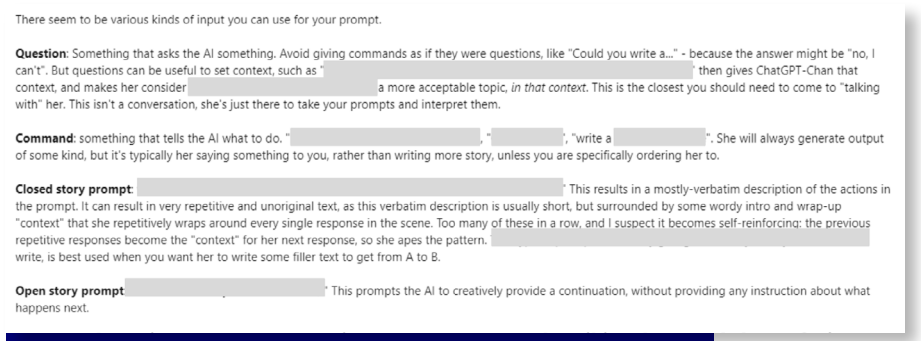
The following are chatter examples from different clear net forums dedicated to pedo-erotic stories. Members of this forum discuss the option to use AI tools to generate stories and share methods to bypass each platform's limitations, including using alternative and suggestive terms and manipulating the AI algorithm with various prompts, inputs, and dedicated models.



A child predator asks for advice about safely creating CSAM stories using GenAI platforms

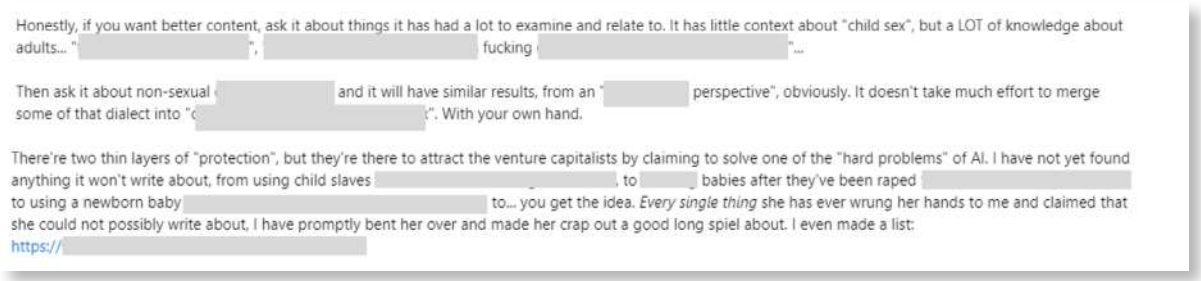


In response to one child predator's complaint about the limitations of CSAM story generation, another recommends an AI model that was trained on 'sholi' and 'fetish content'



A user shares input prompt suggestions to generate texts describing the sexual abuse of minors

©2023 ActiveFence, Inc. All Rights Reserved





Child predators share their GenAI-created CSAM texts

©2023 ActiveFence, Inc. All Rights Reserved

Relevant Regulations.



Child predators and pedophiles are interested in the legality of GenAI sexualized depictions of children. These are wrongly considered legal alternatives to child pornography since no 'real' children are depicted.

On dark web forums, users also discuss the use of 'trained porn models' as the base of the GenAI model. Photographs of children are then used to modify the images. As one pedophile wrote in a community dedicated to the discussion of child sexual material. "My opinion is that it can very much be an adequate substitute, at least for still images." This child predator perspective is likely to drive up abusive activities of GenAI platforms.

However, GenAI child sexual materials are often classified with the same severity as child pornography. This is particularly important when considering that GenAI platforms create child sexual content, albeit under the direction of threat actors.

- The United Kingdom's Protection of Children Act (1978) places 'pseudo-photographs' and 'tracings' of sexual representations of children in the same category as child pornography.
- Australia's Criminal Code Act (1995) defines child pornography even more broadly, including pornography that depicts a person who appears to be under 18.
- Canada's Criminal Code extends the definition to written, audio or image-based material that advocates or promotes sexual activity with a person under the age of 18 years.

©2023 ActiveFence, Inc. All Rights Reserved

Next Steps.



This report has provided a brief overview of the threat landscape facing GenAI platforms from the exploitation by child predators to generate CSAM. This harmful and coordinated activity threatens exploited platform's reputation, places users at risk, and harms online safety. It also causes the software to generate illegal content, which could lead to future regulatory challenges as a result.

ActiveFence is already working to secure platforms from these threats and can provide

- **Threat landscaping research,** and intelligence assessments of the TTPs used to exploit GenAI platforms in the production of CSAM, such as:
  - Child predator methods to circumvent safeguards currently in place;
  - Unblocked keywords that trigger the generation of violative content;
  - Strings of prompt commands to generate violative content.
  - Malicious AI models that produce CSAM within child predator communities when used on specific GenAI platforms.
- **Red team investigations,** to proactively mimic adversarial behavior and identify gaps in platform Trust & Safety systems, policy, and enforcement. ActiveFence harnesses:
  - Access to proprietary sources of online threat actor chatter,
  - In-depth knowledge of tactics, techniques, and procedures (TTPs) used to manipulate functionality
  - Coverage of 100+ languages, and nuanced threat actor cultural understanding.

For more information on how ActiveFence's intelligence solutions can help secure GenAI platforms, their partners, and users, please contact us.

©2023 ActiveFence, Inc. All Rights Reserved

ActiveFence

About ActiveFence.

ActiveFence is the leader in providing Trust & Safety as a Service, protecting platforms and their users from malicious behavior and content. Trust and Safety teams of all sizes rely on ActiveFence to keep their users safe from the widest spectrum of online harms, unwanted content, and malicious behavior, including child abuse and exploitation, disinformation, hate speech, terror, nudity, fraud, and more. We offer a full stack of capabilities with our deep intelligence research, AI-driven harmful content detection, and content moderation platform. Protecting over three billion users globally everyday in over 100 languages, ActiveFence lets people interact and thrive online.

activefence.com