ActiveFence

ActiveScore

Automated AI content detection fueled by intelligence.

Empower your team to make faster decisions with greater accuracy

Harness deep intel expertise to tackle the most evasive threats

Threats are becoming more complex, and context is crucial to understand how violative a piece of content really is. Our models are built with the intelligence of 150+ in-house domain and linguistic experts specializing in child safety, hate speech, misinformation, counter-terror, and more. This enables you to catch even the most evasive threats and protect against the ever-changing threat landscape platforms and users face.

Access years of intel insights with one seamless integration

Setup is simple and takes a few hours to configure with our API documentation. Each piece of content is automatically analyzed against our ever-growing proprietary database of billions of signals of malicious content collected over years of research into the way bad actors operate across the clear, deep, and dark web. We look for duplicates and similarities across keywords, videos, audio files, images, and other abusive content, in order to provide you with a more confident risk score.

Increase moderator efficiency with adaptive, contextual AI models

Minimize exposure and handle violations fast. We analyze all surrounding metadata such as title, description, user, thumbnails, and more to flag what is risky and weed out the benign. Our models constantly learn and improve from every decision, enabling you to reach higher accuracy and better efficiency.

- Increase moderator efficiency and reduce and achieve <1% false positive rate
- Coverage in 100+ languages, slang, l33tspeak, emojis, and more to catch unknown unknowns
- Support across multiple media formats, including text, audio, images, and video
- Adaptive to the ever-changing adversarial landscape to stay ahead of evolving threats
- Continuous decision-based feedback loop to improve accuracy over time

"ActiveFence's Risk Score Engine analysis is great"

Ryan Skidmore, Software Engineer | deliveroo

Protecting 3B+ global users against the most complex threats:

- Adult Content
- Bullying & Harassment
- Child Safety
- Graphic Violence
- Hate Speech
- Illegal Goods
- Misinformation
- Nudity
- PII
- Profanity
- Suicide & Self-Harm
- Violent Extremism

TRUSTED BY THESE COMPANIES AND MORE

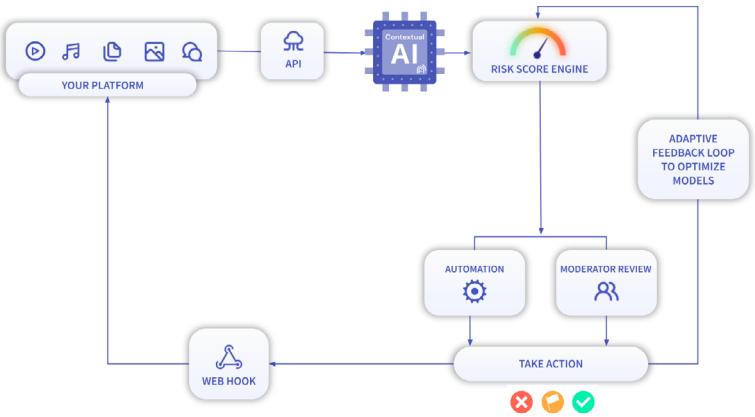
THE TREVOR PROJECT · Dapper Labs · SOUNDCLOUD · vimeo · audiomack · Post. · maxxer

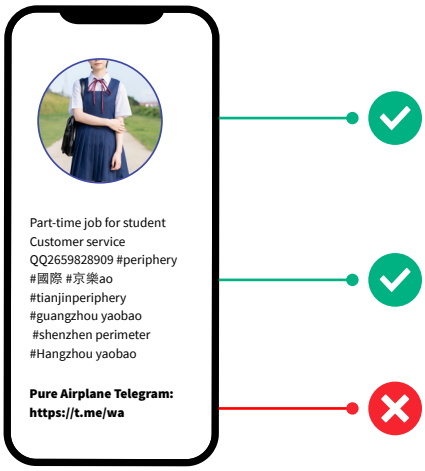
CONTEXTUAL AI IN ACTION

Eliminating blindspots to catch the unknowns

The confidence to make accurate decisions

Integrate once to automatically receive risk score analysis on text, audio, image, or video content. Define thresholds for automated actioning, or push to prioritized moderator queues, ensuring swift, accurate decisions on the spot. Every action taken by your team will train our adaptive, contextual AI to improve scoring over time and increase accuracy even further.





Uncovering CSAM network in a seemingly harmless profile using expert insights

A profile containing a seemingly benign picture and description was sent for analysis on ActiveFence's Risk Score Engine. Fueled by intel insights, and analyzing the profile's complete metadata, including image, description, comments, and shared URLs, the engine labeled the profile as high-risk due to its promotion of a link to an off-platform group dedicated to the trade of CSAM. Using internal detection tools, the profile would have been overlooked, due to the off-platform nature of the shared content.

Reducing false negatives of white supremacist songs using the largest database of Hate Speech songs

A seemingly harmless English punk band uploaded their album to an audio-sharing platform, and was not initially flagged as malicious. When analyzed with ActiveFence's Risk Score Engine, the cover art, audio, titles, band name,and other metadata were automatically analyzed against our proprietary database that includes a collection of over 14K hate speech songs. Within seconds, we provided a high risk score for hate speech, as the band was previously flagged by our intelligence experts and labeled in our database.