SAFER PREDICT

safer BUILT BY THORN

Advanced AI-Powered Text Detection

Safeguard your platform and users from child sexual exploitation

Empower your trust and safety team to proactively combat text-based child sexual exploitation (CSE) at scale. Access powerful AI technology to identify text-based interactions indicating or potentially leading to CSE, ensuring user safety and a secure digital environment.

Harnessing the power of AI

Safer Predict leverages a machine learning (ML) classification model (a.k.a. "text classifier") to detect CSE. Send lines of text to the classifier, which analyzes the conversation line by line. The classifier then categorizes content and assigns a risk score based on the content's likelihood of being associated with harmful behavior such as:

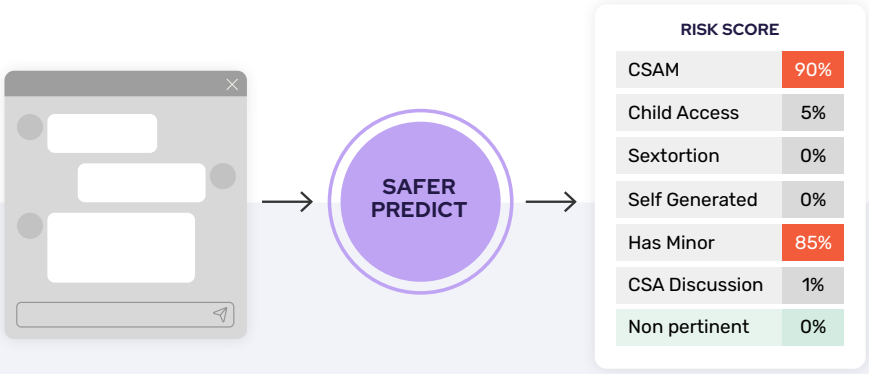
CSAM
Messages related to asking for, transacting in, and sharing CSAM

Child access
Messages discussing offline access to and harm of children

Sextortion
Messages related to sextortion activities

Self-generated content
Requests for and discussions about self-generated content

CSA discussion
Messages about sexually harming children in an offline setting



> "Safer Predict's text classifier significantly improves our ability to prioritize and escalate high-risk content and accounts. The multiple labels and risk scores help our team focus on problem accounts. Thorn's expertise in detection of child sexual abuse and exploitation makes Safer Predict far more effective than anything we could have built in-house."
>
> **NILES LIVINGSTON, CHILD SAFETY MANAGER, MEDIALAB**

Strengthen your platform's defenses

With Safer Predict, identify potential threats, such as discussions about child sexual abuse material (CSAM), sextortion threats, and other forms of sexual harm to minors.

Proactive detection
Detect and flag potentially harmful conversations, empowering your trust and safety teams to take swift action before it can cause harm.

Tailored safeguards
Tailor detection strategies and customize workflows in alignment with your policies and operational needs.

Efficient moderation
Streamline processes by empowering moderators to investigate pertinent results and prioritize harmful content for removal.

Flexible configuration
"Stack" multiple labels to zero in on problem areas or quickly target abuse content to support enforcement of your policies.

What sets Safer Predict apart?

Safer Predict is at the forefront of safeguarding digital environments against emerging threats of sexual harm to minors.

Built on trusted, targeted data

Safer Predict's text detection models are trained on real conversations, validated by trust and safety pros, discussing:

- Sextortion
- Requests for, transacting in, and sharing of CSAM
- Requests for a minor's self-generated sexual content
- Minors generating content
- Having access to and sexually harming children offline

Sophisticated model training

Trained on datasets that include various forms of harm, language models understand complex language patterns and discern subtle nuances that might indicate harmful behavior.

Contextual analysis

By examining the broader context of entire conversations, line-by-line, Safer Predict captures the full scope of interactions, identifying potential threats that might otherwise be missed.

Get started today

Ready to elevate your platform's safety?
Contact us today to build a safer digital future together.

safer.io | info@getsafer.io