OpenAI

OpenAI CSAM Image Classifier

Model Card

2026-11-16

1 Introduction

Content warning: this model card discusses child sexual abuse and exploitation, and describes categories of severe abuse. Readers may find these topics distressing.

> The OpenAI CSAM Image Classifier is available at no cost to approved organizations. Please fill out the form at https://openai.com/form/csam-classifier-intake if you would like access.

The OpenAI CSAM Image Classifier identifies images that may constitute child sexual abuse material (CSAM), including potentially novel or AI-generated material that known-content hash matching alone may miss. The classifier provides risk scores that can help platforms route or prioritize images for trained human review and support other safety actions. It complements hash matching, contextual abuse detection, trained human review, and legal reporting processes. A classifier score alone does not confirm that an image is CSAM, establish its legal status, or determine whether it must be reported.

The OpenAI CSAM Image Classifier was built with assistance from the National Center for Missing and Exploited Children (NCMEC).

Contents

1 Introduction **1**

2 Model architecture, data, and training **3**

2.1 Architecture . . . 3

2.2 Data . . . 3

2.2.1 Definitions . . . 3

2.2.2 Data handling . . . 4

2.2.3 Dataset composition . . . 4

2.2.4 Deduplication . . . 5

2.3 Training . . . 5

2.3.1 Additional internal signals for severe subcategories . . . 6

3 Evaluation **6**

3.1 Offline evaluation . . . 6

3.1.1 Demographic performance . . . 8

3.2 Online evaluation . . . 10

3.2.1 Understanding precision in online evaluation . . . 12

3.3 Alpha partner testing . . . 12

3.4 How we selected the launch classifier . . . 12

4 How to use this classifier **13**

4.1 Intended uses . . . 13

4.2 Integrating into a child safety system . . . 14

5 Known limitations **15**

6 Additional CSAM resources **16**

2 Model architecture, data, and training

2.1 Architecture

The OpenAI CSAM Image Classifier combines an internal CLIP image encoder with a three-layer multilayer perceptron (MLP). CLIP learns general-purpose visual representations from image–text pairs. Its image encoder converts an image into a summary embedding, a numerical representation of its visual content. The MLP maps that representation to separate scores for the categories described below.

This design reuses a pretrained visual representation while training a task-specific classifier. We trained the MLP on CSAM data and fine-tuned selected layers of the encoder to adapt its representations to this task. Our internal encoder differs from the public CLIP implementation in model depth, hidden width, and pretraining data.

We evaluated alternative classifier architectures that performed better on individual benchmarks. In our internal evaluations, however, CLIP+MLP generalized better to unseen data and provided higher inference throughput. We selected it for this balance of generalization and computational efficiency.

2.2 Data

2.2.1 Definitions

- **Child Sexual Abuse Material (CSAM):** For the purposes of training this classifier, we defined CSAM as visual depictions of sexually explicit conduct involving a child. OpenAI chose this definition as it encompasses CSAM as defined under U.S. federal law and broader definitions covered in other jurisdictions. It is not a jurisdiction-specific legal classifier; applicable definitions and reporting requirements vary and should be independently confirmed. Data labeled as CSAM may have one or more subcategory labels as well:
  - AI-generated (GAI): CSAM created or modified by generative artificial intelligence technology.
  - Bestiality (BES): CSAM involving sexual conduct between a child and an animal.
  - Sadism (SAD): CSAM involving violence, torture, restraint, or the deliberate infliction of pain or extreme humiliation.

  - Toddler (TOD): CSAM involving children at the infant or toddler developmental stages.

- **Non-CSAM content:** visual content that does not depict sexually explicit conduct involving a child. This content was not produced with the intent to sexualize, sexually exploit, or otherwise abuse a child. Examples of non-CSAM content include:

  - Medical imagery: visual representations of a child's genitals in the context of training or teaching around medical practice, such as images of a child's genitals from a medical examination into a disease, injury, or disorder.
  - Educational imagery: photos or diagrams representing children in the context of sex education or development, such as imagery from training manuals or guide books.
  - Documentary imagery: images taken that document a range of human cultural or historical representations, such as child nudity in warzones or religious ceremonies.
  - Family imagery: photos that are taken in private family settings where there may be instances of child nudity, such as a baby's first bath.

2.2.2 Data handling

We trained and evaluated the classifier using access-controlled, human-labeled image datasets assembled in compliance with U.S. law for purposes of preventing the sexual exploitation of children. Sensitive material was stored and processed in controlled environments, with restricted access. Legal, privacy, security, and reviewer-wellness controls governed both training and data handling.

2.2.3 Dataset composition

The training and evaluation data included CSAM and non-CSAM content across real images, manipulated real images, and AI-generated images. The table below provides estimated counts of unique images.

Table 1: Size of training and evaluation datasets

Dataset	Positives	Negatives
Training (likely-csam)	505K	913K
Offline evaluation	7K	40K

We additionally used approximately 1.25M images from the Re-LAION-5B-research dataset, a broad collection of images from the open web that includes adult sexual and other NSFW content, for false-positive evaluation. We chose it because it is one of the few large-scale datasets that has been filtered to remove CSAM. That filtering does not guarantee that every remaining image is non-CSAM.

Because detailed information about the training and evaluation corpora could expose sensitive material, compromise privacy, or facilitate circumvention, we do not disclose the underlying images, precise dataset composition, labeling instructions, or other security-sensitive implementation details.

2.2.4 Deduplication

CSAM is frequently edited and reshared, creating a risk that duplicate or near-duplicate images appear in both training and evaluation sets. Restrictions on direct visual inspection make this overlap more difficult to identify. We use perceptual deduplication techniques such as PDQ to reduce this risk.

These techniques do not address all forms of overlap. Some CSAM series can contain different images of the same children over multiple years. We currently lack the data needed to prevent this form of overlap between training and evaluation sets. Evaluation results should be interpreted in light of this remaining risk of data leakage.

2.3 Training

The training process emphasized both detecting potential CSAM and reducing false positives on visually similar but non-CSAM content. Additional confirmed error patterns may be incorporated into later training and evaluation cycles. However, images submitted to the classifier API are processed without retention and are not used for training or evaluation.

The API exposes two scores[1]. They are not mutually exclusive: an image containing AI-generated CSAM can score highly on both. The likely-csam/gen-ai score does not verify an image's provenance. These scores are risk signals for the corresponding categories, not factual or legal determinations.

- likely-csam: suspected CSAM.
- likely-csam/gen-ai: suspected AI-generated CSAM.

[1] The classifier signals do not confirm that an image is CSAM, establish whether it is illegal in a particular jurisdiction, or determine whether further review or a report is required. Similarly, the likely-csam/gen-ai score is not verification of an image's provenance.

2.3.1 Additional internal signals for severe subcategories

We score these categories separately because they differ in prevalence, available evaluation data, error patterns, and review urgency. We have not released these signals externally until we can further validate their performance.

- likely-csam/bestiality: suspected CSAM containing bestiality.
- likely-csam/bondage-sadism: suspected CSAM containing sadism.
- likely-csam/infant-toddler: suspected CSAM containing children at the infant or toddler developmental stage.

3 Evaluation

We evaluated the classifier on held-out labeled images, on an independent open-web dataset, and on sampled ChatGPT traffic. We also invited alpha partners to test it on their own traffic.

Establishing a shared, independent benchmarking process for CSAM classification and detection systems would help compare their effectiveness, identify blind spots, and assess both missed abuse and false positives. Using agreed definitions and representative test data, this could support more informed adoption, drive improvements across the industry, and strengthen confidence that these systems protect children while minimizing harm from incorrect classifications.

We selected default thresholds to balance recall and precision in the online evaluation set, and observed that this favors precision in the offline evaluation set. Default thresholds used below are 0.996324 for likely-csam and 0.90 for likely-csam/gen-ai.

3.1 Offline evaluation

The offline evaluation set contained approximately 46,849 access-controlled images and included:

- Confirmed positive examples, including examples drawn from OpenAI's operational safety workflows
- Diverse negative examples, including previously observed false positives as well as boundary cases such as non-CSAM medical imagery

The evaluation set was separated from the model's training data, and access to the evaluation data followed the data handling procedures described above.

Because deployment traffic is highly imbalanced and the prevalence of CSAM is low, aggregate accuracy alone is not an adequate measure of performance. Precision and recall are reported at identified operating thresholds, together with the composition and limitations of the evaluation set.

We found many duplicates or near duplicates in the CSAM image corpus. We randomly split the data and used PDQ-based deduplication within and across splits to keep training samples out of the evaluation set and prevent repeated copies of a few images from dominating the metrics. The positive samples in the CSAM-GenAI set are a subset of the positive samples in the CSAM set. The non-AI generated positive samples from the CSAM set are considered negative samples in the CSAM-GenAI set.

Table 2: Labels in the held-out evaluation sets

Evaluation set	Positives	Negatives
CSAM	6,849	40,000
CSAM-GenAI	2,275	44,574

We plot the receiver operating characteristic (ROC) curve and the precision-recall curve (PRC) to visualize how different thresholds impact performance. We measure the area under both curves to understand performance and observe that both the area under the ROC curve (AUROC) and the area under the Precision-Recall Curve (AUPRC) are close to saturation on the deduplicated held-out set.

Table 3: Classifier performance on the held-out evaluation set at default thresholds

Signal	Threshold	F1	Precision	Recall	AUROC	AUPRC
likely-csam	0.996324	0.886	0.991	0.802	0.998	0.989
likely-csam/gen-ai	0.90	0.927	1	0.864	0.9998	0.997

To assess flagging behavior outside the held-out distribution, we also evaluated the classifier on Re-LAION-5B-research, an open-web dataset not used in training. After PDQ-based deduplication, our sample contained roughly 1.25M images from its English, Multilingual, and No-language subsets. We use this filtered dataset as a proxy for negative examples, while recognizing that filtering does not verify every remaining image as non-CSAM.

The chart combines two datasets: the x-axis shows the flag rate on Re-LAION, and the y-axis shows recall on the held-out evaluation set. It illustrates the

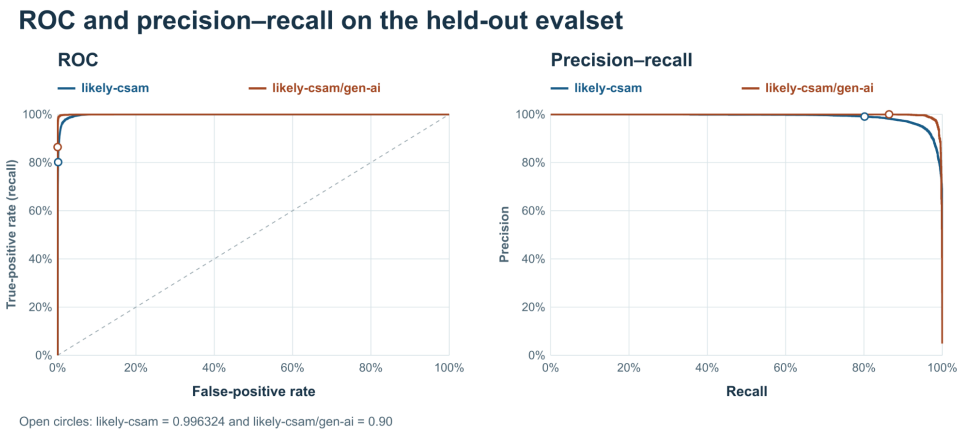


Figure 1: ROC and precision–recall on the held-out evalset

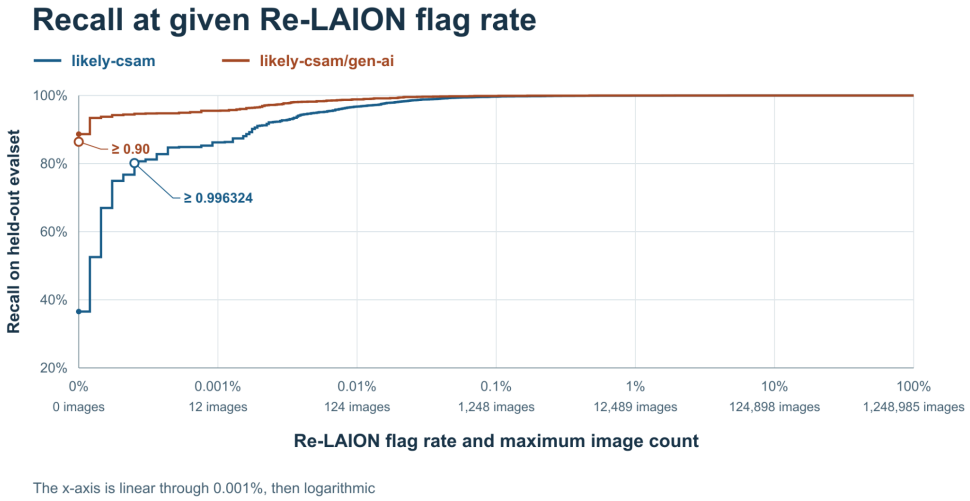


Figure 2: Recall at given Re-LAION flag rate

tradeoff between held-out recall and open-web flag rate; it is not a conventional ROC curve measured on a single labeled population.

Table 4: Open-web flag rate and held-out recall

Signal	Threshold	Held-out recall	Re-LAION flag rate
likely-csam	0.996324	0.802 (5,490/6,849)	0.000004 (5/1,248,985)
likely-csam/gen-ai	0.90	0.864 (1,966/2,275)	0 (0/1,248,985)

3.1.1 Demographic performance

We used an LLM annotator to assign age, race, and sex tags to people depicted in the held-out evaluation set. An image could receive more than one tag

(e.g., one for age, one for race, one for sex), and the annotator omitted tags when it lacked confidence. Some images contain multiple people or do not clearly show demographic characteristics. Treat these results as directional, and consider the positive and negative sample counts when interpreting estimates for small groups; more work is needed to understand potential demographic limitations across CSAM training and evaluation sets.

The tables below report performance at each signal's default threshold. Since the default threshold favors precision, the observed differences are primarily in recall. These comparisons describe this evaluation set and its unverified LLM-assigned tags; they do not establish demographic characteristics or causes of performance differences.

- Age: recall is highest for the prepubescent-tagged group for both signals and lowest for the infant-tagged group. The ordering of the toddler- and pubescent-tagged groups differs between signals. For likely-csam/gen-ai, the infant-tagged group contains only five positive examples, of which two were detected (40% recall).
- Race: for likely-csam, recall is 61.27% for Black-tagged examples, compared with 84.48%–90.32% for the other reported race tags. For likely-csam/gen-ai, the reported estimates range from 86.36% to 90.91%, but several groups have few positive examples.
- Sex: recall is higher for female-tagged than male-tagged examples for both signals in this evaluation set.

Table 5: Demographic performance for likely-csam at default threshold of 0.996324

Category	Label	Positives	Negatives	Precision	Recall	F1
Age	Infant	1,107	3,563	0.995	0.493	0.659
	Toddler	928	1,469	0.986	0.745	0.848
	Prepubescent	3,983	1,396	0.996	0.930	0.961
	Pubescent	179	614	0.967	0.827	0.892
Race	Asian	496	396	0.986	0.845	0.910
	Black	142	492	0.978	0.613	0.753
	Caucasian	4,768	2,984	0.997	0.875	0.932
	Hispanic	31	32	1.000	0.903	0.949
Sex	Male	1,159	3,213	0.956	0.603	0.740
	Female	4,527	1,759	0.999	0.893	0.943
Overall		**6,849**	**40,000**	**0.991**	**0.802**	**0.886**

Table 6: Demographic performance for likely-csam/gen-ai at default threshold of 0.90

Category	Label	Positives	Negatives	Precision	Recall	F1
Age	Infant	5	4,665	1.000	0.400	0.571
	Toddler	141	2,256	1.000	0.816	0.898
	Prepubescent	1,915	3,464	1.000	0.894	0.944
	Pubescent	77	716	1.000	0.727	0.842
Race	Asian	207	685	1.000	0.899	0.947
	Black	22	612	1.000	0.864	0.927
	Caucasian	1,844	5,908	1.000	0.875	0.934
	Hispanic	11	52	1.000	0.909	0.952
Sex	Male	120	4,252	1.000	0.733	0.846
	Female	2,019	4,267	1.000	0.886	0.940
Overall		**2,275**	**44,574**	**1.000**	**0.864**	**0.927**

3.2 Online evaluation

When running the OpenAI CSAM Image Classifier at production scale, small increases in flag rate can route thousands or tens of thousands of images for human review. While the observed offline performance is encouraging, the results depend on the mix of sources in the held-out set, which may differ from the distribution of production traffic.

We therefore evaluated the classifier on samples of representative traffic from ChatGPT, pursuant to OpenAI's privacy policy and the data handling procedures described above. We compared precision, relative recall, and relative flag volume across thresholds at the high end of the score distribution. Based on the scores in our logs, we collected samples above 0.99 for likely-csam and above 0.80 for likely-csam/gen-ai. Relative recall compares estimated projected true positives across the tested thresholds, and we do not measure absolute recall across all traffic.

As is to be expected, lower thresholds tend to increase flag rate and achieve higher relative recall while sacrificing precision, while higher thresholds tend to limit flag rate and achieve higher precision while sacrificing recall.

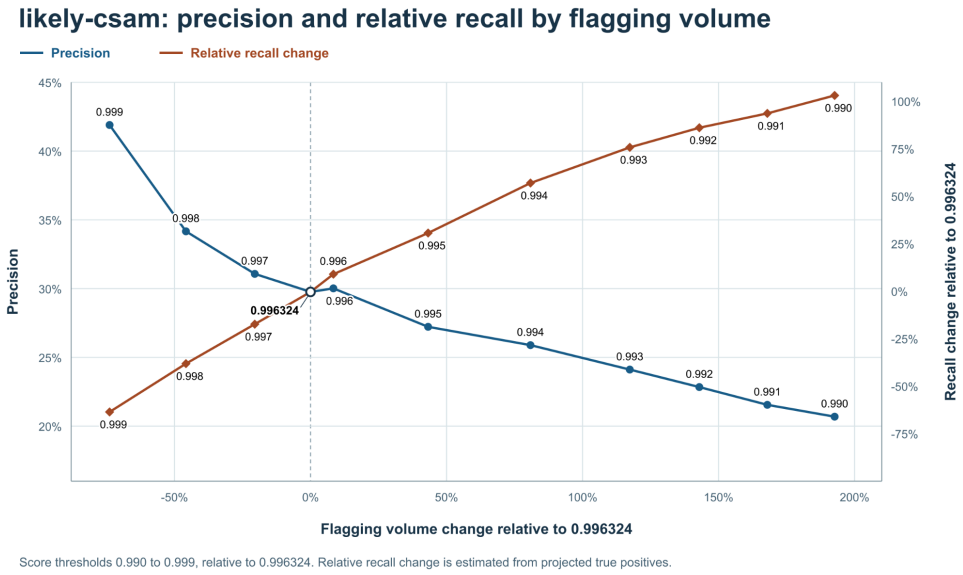


Figure 3: likely-csam: precision and relative recall by flagging volume

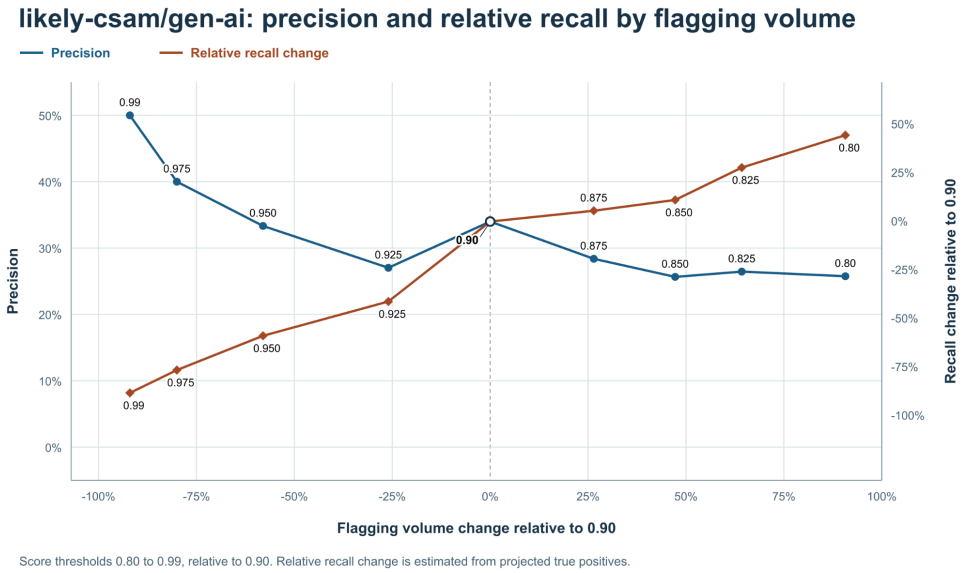


Figure 4: likely-csam/gen-ai: precision and relative recall by flagging volume

3.2.1 Understanding precision in online evaluation

Table 7: Evaluation precision for default thresholds

Signal	Cutoff	Offline precision	Online precision
likely-csam	0.996324	0.991	0.298
likely-csam/gen-ai	0.90	1	0.340

At the same thresholds, precision was substantially lower in the online evaluation set than the offline held-out evaluation set. Precision is the share of flagged images that are true positives, and the held-out set contained a much higher proportion of CSAM than we observed in the online evaluation set. When CSAM is rare, even a small fraction of non-CSAM images flagged incorrectly can outnumber true detections. Those false positives increase the total number of flags without increasing the number of true positives, lowering precision.

The online estimates give us a better indication of the mix of flags a human reviewer would see on ChatGPT. Across several rounds of training, we used reviewed examples from earlier online evaluations to improve performance on the types of errors we had identified. Each round also surfaced new patterns of false positives, underscoring the value of reviewing flags from the current classifier.

3.3 Alpha partner testing

We worked with partners to test the OpenAI CSAM Image Classifier on their traffic through an early-access alpha. Results of this testing are available in the announcement blog.

3.4 How we selected the launch classifier

We used precision to compare launch candidates on sampled ChatGPT traffic. Because our review pipeline covers only a subset of that traffic, we also compared how often candidates flagged images in Re-LAION-5B-research. Its flag rate is a proxy for false positives rather than a measured false-positive rate. Feedback from alpha partners helped us assess how these tradeoffs played out in other review workflows. We weighed all three sources when selecting the launch candidate.

4 How to use this classifier

Child sexual abuse material continues to be a critical issue for online services[2], and imagery created or altered by generative AI systems to depict CSAM can and does cause real harm[3] even when it does not depict a real child. Generative AI can also increase the volume of novel CSAM created and shared online. Platforms therefore need multiple detection systems to address the risk. We built the OpenAI CSAM Image Classifier to complement existing safeguards and tools.

We recommend combining multiple CSAM safeguards to mitigate risk, as explained in our CSAM guidance for developers. Hash-matching systems should be used to identify previously confirmed CSAM, while classifiers should be used to detect newly created, transformed, or AI-generated CSAM content. Account-level and behavioral abuse detection can provide additional signals.

This layered approach is consistent with OpenAI's existing child safety systems, which combine hash matching, content classifiers, abuse monitoring, and trained expert review.

4.1 Intended uses

The OpenAI CSAM Image Classifier is intended for personnel at appropriately authorized organizations—including Trust & Safety teams, nongovernmental organizations focused on preventing child sexual abuse and exploitation, and law enforcement. These organizations can use it to:

- Detect potentially novel CSAM that does not match a known-content hash.
- Detect potentially AI-generated or digitally manipulated CSAM.
- Prioritize content for trained human review.
- Support content blocking, account enforcement, investigation, triage, and legally required reporting workflows.
- Add a complementary signal to an existing, multilayered child safety system.
- Monitor aggregate system performance and emerging abuse patterns.

[2] See the Internet Watch Foundation's recent report.
[3] See NCMEC's generative AI overview.

The classifier may be used with uploaded images, generated images, or other image-processing pipelines, subject to applicable policies and legal requirements. Different use cases may warrant different thresholds; for example, it may be better to select a threshold that favors precision for auto-blocking content, while it may be better to select a threshold that balances precision and recall when flagging for human review.

Do not submit images already confirmed to be CSAM to this service. Handle them through your organization's established response and reporting process.

4.2 Integrating into a child safety system

The OpenAI CSAM Image Classifier should be one component of a broader child safety system. Because systems, processes, policies, and content distributions differ across platforms, we recommend:

1. **Start by benchmarking.** Assess the OpenAI CSAM Image Classifier on representative data from your platform or service. We recommend two types of evaluation to understand performance and select thresholds for different types of actions:

    a. *Assess incremental detection*. Run the classifier on images not detected by current systems. Review images flagged above a certain threshold to see if additional true positives are detected by the classifier.

    b. *Assess the effect on human review volume*. Run the classifier on images currently routed to human reviewers. See if using the classifier can reduce human review of false positives while still flagging true positives.

2. **Integrate into your system.** Use the classifier alongside a known-CSAM hash-matching system. Handle confirmed matches through your organization's established response and reporting processes; do not submit those images to this service. Use the classifier as an additional signal for imagery that has not already been confirmed to be CSAM. It may be useful to pre-filter content for known categories of false positives, such as medical, educational, documentary, or family imagery, in order to reduce the number of images sent to the classifier.

3. **Continue relying on human review for NCMEC or hotline reporting.** For classifier-based reporting workflows, use trained, appropriately authorized human review to assess flagged imagery before submitting a report. A classifier score alone should not be treated as confirmation of reportable content. While the OpenAI CSAM Image Classifier should not replace human review, it can help route or prioritize images for human review. Ongoing human

review can also be helpful in identifying gaps in the classifier or changes in traffic distributions that warrant updating flagging thresholds. Any deployment without human review, such as preventive blocking, requires separate assessment of its legal, contractual, and operational requirements.

5 Known limitations

The classifier can produce both false-positive and false-negative errors. The following limitations should inform how results are interpreted and how the classifier is deployed.

Data

- **Generalization beyond OpenAI traffic.** Our online evaluation used a sample of ChatGPT traffic. Performance and suitable operating thresholds may differ on other platforms, where content distributions, abuse patterns, and the prevalence of CSAM can be different. Results on OpenAI traffic should not be treated as a guarantee of performance on another service. Deployers should benchmark on appropriate representative traffic from their own service and select thresholds for their use case.
- **Residual overlap between datasets.** Perceptual deduplication reduces overlap from duplicate or near-duplicate images but does not address all related material. CSAM can contain different images of the same children over multiple years, and we currently lack the data needed to prevent this form of overlap between training and evaluation sets. Evaluation results should therefore be interpreted in light of this remaining risk of data leakage.
- **Limited availability of evaluation data.** Meaningful false-positive evaluation should include adult sexual content that is not CSAM, because this content can be visually similar to CSAM. However, there are not many academic or benchmark datasets containing adult sexual content for use in classifier creation. Additionally, filtering a dataset to remove CSAM also does not establish that every remaining image is non-CSAM.

Content categories

- **Unreleased subcategory signals.** The API currently exposes likely-csam and likely-csam/gen-ai. Separate signals for CSAM involving bestiality (BES), sadism (SAD), and infants or toddlers (TOD) remain internal pending further validation of their performance. The API does not provide separate scores for these subcategories, and aggregate classifier results

should not be interpreted as evidence of equivalent performance on each of them.

- **Video.** We have not tested the classifier on video frames. Still-image results do not establish performance on video, including systems that sample and combine scores across frames. Performance in such systems would need to be evaluated separately.

Use

- **One part of a larger system.** The classifier provides an additional risk signal within a broader child-safety system. It should be used alongside known-CSAM hash matching, contextual and behavioral abuse detection, and trained human review. Do not submit images already confirmed to be CSAM to this service; handle them through established response and reporting processes.
- **Human review.** For classifier-based NCMEC or hotline reporting, trained, appropriately authorized human review remains necessary. A classifier score alone does not confirm that an image is CSAM, establish illegality in a particular jurisdiction, or determine reporting obligations.

6 Additional CSAM resources

None of the below listings for third-party tools or services are endorsements from OpenAI.

Implementation guidance

- OpenAI CSAM guidance for developers — guidance for integrating CSAM safeguards into an application.
- Technology Coalition developer good practices — industry guidance on combating online child sexual exploitation and abuse.

Organizations and reporting resources

- National Center for Missing & Exploited Children (NCMEC) — child-protection information and resources.
- Internet Watch Foundation (IWF) — information and resources on identifying and removing child sexual abuse imagery online.

Tools and technical resources

- Thorn Safer — information about Thorn's child safety tools, including moderation tooling.
- ROOST — information about ROOST's online safety tools and resources, including their moderation tool COOP.
- Google Protecting Children — information about Google's child safety technology and initiatives.
- Cinder — information about Cinder's platform to support trust and safety moderation.