Tech Coalition

Together, we fight online child sexual abuse.

How red teaming can help reduce child safety risks in generative AI

July 2026

Executive summary

Generative AI systems can create new risks for online child sexual exploitation and abuse.

Red teaming helps companies identify and reduce those risks before bad actors can exploit them. When teams test systems rigorously and responsibly, they can better understand how safeguards may fail, how people may try to misuse a product, and what changes teams need to make to strengthen protections.

Red teaming results need context. A higher number of harmful outputs does not always mean a system is more dangerous. It may show that a company has expanded its testing, improved its detection methods, or applied more rigorous evaluation. Raw numbers alone do not tell the full story.

Effective red teaming is risk-based, realistic, expert-informed, rigorous, iterative, and transparent. Stakeholders should look beyond headline metrics and ask how teams conduct testing, what risks they evaluate, and how companies use findings to improve safety.

The goal should be to encourage companies to find and fix risks, not to reward those that report the fewest failures. Robust testing, transparent reporting, and thoughtful interpretation of results can help companies strengthen safeguards and better protect children.

What is red teaming?

Red teaming is structured adversarial testing designed to identify how a system might fail, be misused, or produce harmful outcomes.

In the context of generative AI and online child sexual exploitation and abuse, this may include testing whether systems can generate, transform, describe, facilitate, or fail to block harmful child safety-related content or behavior.

Pre-deployment red teaming can help teams surface and mitigate those risks before release, reducing the chance that bad actors exploit them. Red teaming may also be repeated as models, product features, or integrations materially change. Post-launch monitoring for real-world misuse and emerging attack techniques can inform future rounds of red teaming and mitigation.

Online child sexual exploitation and abuse presents unique challenges for AI safety testing because there are important legal and ethical limits on what can be directly generated, possessed, or evaluated.

As a result, red teaming may focus on content, behaviors, and scenarios that can be tested lawfully rather than attempting to test illegal content directly. While these legal constraints limit what testing can reveal, red teaming still provides meaningful insight into system vulnerabilities, safeguard performance, and areas where further mitigation is needed.

Even when illegal content is not intentionally sought, testing may unexpectedly produce or surface reportable content. Establishing policies, procedures, and reporting pathways in advance can help companies respond appropriately if testing unexpectedly surfaces illegal content. Putting these protocols in place supports responsible reporting and enables teams to conduct safety testing with greater clarity and confidence.

What makes red teaming *effective*?

Effective red teaming is, among other things, risk-based, realistic, expert-informed, rigorous, iterative, and transparent.

- **Risk-based**: Testing focuses on the ways a system could be misused or fail in child safety contexts. This begins with understanding the nature of OCSEA offenses, and comprehensively considering the system capabilities, and how they could be used to perpetrate harm.
- **Realistic**: Red teaming should reflect real-world conditions and how bad actors actually attempt to misuse systems, while remaining within applicable legal boundaries. This may include testing evolving tactics, multiple languages, different modalities, multi-step interactions, and deployment-specific risks using lawful content, behaviors and scenarios where direct testing is not possible. Testing based solely on static prompts or laboratory conditions may fail to identify emerging threats.
- **Expert-informed:** Effective red teaming incorporates relevant subject matter expertise. In child safety, this may include child safety specialists, trust and safety practitioners, legal experts, and others with knowledge of OCSEA harms, offender behaviors, and applicable legal frameworks. This expertise helps teams determine which risks can be evaluated directly, which require the use of lawful proxies or adjacent scenarios, and how testing can remain both meaningful and legally compliant.
  Domain expertise helps ensure that testing focuses on meaningful risks, evaluates outputs appropriately, and reflects how harms occur in practice.
- **Rigorous**: Structured methods are used with clear objectives, documented methodologies, consistent criteria for evaluating outputs, and a way to distinguish between different types and levels of harm. Effective testing is done at sufficient scale to differentiate between systemic and outlying vulnerabilities.
- **Iterative**: Red teaming should evolve alongside the technology and threat landscape. As safeguards improve, testing methods, prompts, and evaluation criteria must also evolve. New model capabilities and real-world misuse patterns should continuously inform future testing; post-launch monitoring can surface novel attack pathways and identify safeguards that degrade as models, surfaces, and integrations change. These prompts or tactics can be incorporated into future pre-launch red teaming efforts.
- **Transparent**: Clear communication of methodology, scope, and limitations is important. Internally, relevant teams benefit from documentation explaining what was tested, how it was tested, why the testing was conducted, and what the results mean. When red teaming outcomes are shared externally - such as through model cards, transparency reports, or other public disclosures - results can be paired with context to support accurate interpretation. This may include any material changes to methodology, coverage, scenario libraries, or detection capability across reporting periods, along with known limitations of the evaluation. Raw numbers alone may not provide an accurate picture of safety performance.

Effective red teaming is not only defined by passes and failure counts, but by the quality of the testing process and the organization's ability to identify, understand, and mitigate risk.

Interpreting red teaming results

When the results of red teaming (such as the number of test failures) are reviewed without context, they may be open to misinterpretation.
A report stating that a test produced a large number of harmful outputs may be interpreted as evidence that the model is unsafe. However, it may also reflect more rigorous testing, expanded coverage, stronger detection systems, or the inclusion of new attack scenarios that were not previously evaluated. In many cases, adversarial testing teams identify more issues because they have become better at finding them.

Raw numbers can also be misleading without additional context. For example, 500 harmful outputs carry a very different meaning across 1,000 test attempts than across 10 million test attempts. The meaning of any finding also depends on the purpose of the test. Different red teaming exercises may focus on different risk areas, attack vectors, or model capabilities based on a company's assessment of where the greatest risks exist. As a result, findings should be interpreted in light of the testing methodology, objectives, severity of outputs, and overall risk profile, rather than as standalone metrics.

For these reasons, results from red teaming are most meaningful when accompanied by information about what was tested, how it was tested, why the testing was conducted, and what changes were made in response to the findings. Transparency about methodology and context helps stakeholders better understand whether results reflect increased risk, improved detection, expanded testing coverage, or some combination of the three. Stakeholders assessing red teaming results should look for that context and treat its absence as a constraint on what the results can support; a shift or difference in failure counts says little unless the testing that produced it is comparable. Additionally, the same caution applies across organizations, whose testing methodologies and scenario coverage may differ substantially.

While illegal OCSEA content surfaced as part of red teaming and model training must be reported to NCMEC or the relevant authority, it is worth distinguishing these reports from user-facing incidents. The Tech Coalition recently released an addendum to the Trust Framework that makes this distinction clear: violative outputs identified through red teaming are distinct from those encountered by users or consumers.

Different testing methods answer different questions

Not all red teaming is conducted in the same way. For example, some exercises are performed manually by expert red teamers who creatively probe a system for vulnerabilities. Others use automated methods to generate and evaluate large volumes of prompts at scale.

Both approaches provide valuable but different information. Manual red teaming can uncover novel attack pathways and nuanced failures that automated testing may miss, but it is inherently limited in scale and is generally not intended to produce statistically representative results. Automated testing can provide broader coverage and more quantitative insights, but may be less effective at identifying entirely new attack strategies.

Understanding whether findings came from manual or automated testing, and what question that testing was designed to answer, helps stakeholders interpret results appropriately.

Aligning incentives

Effective red teaming depends on incentives that reward rigorous testing, continuous improvement, and transparent reporting. Transparency allows companies to learn from one another about novel risks, methodologies, and mitigations. Because these incentives shape what companies are willing to disclose, stakeholders who misinterpret or narrowly focus on the total number of reported failures, may inadvertently discourage the very practices that make testing most effective and systems safer.

Stakeholders across sectors benefit when companies conduct realistic, iterative testing and share results transparently. Policymakers, regulators, researchers, and the public can ask questions that go beyond raw numbers and instead get to the quality of the testing program itself. These questions include whether testing reflects how bad actors attempt to misuse systems within applicable legal constraints, incorporates evolving threats, involves appropriate expertise, and leads to measurable safety improvements. Together, these factors provide a more meaningful assessment of safety than raw metrics alone.

The Tech Coalition has contributed to this environment with the creation of a standardized industry template for reporting AI-generated CSAM to NCMEC's CyberTipline, including for content generated during red teaming.

The template clarifies the testing context and absence of a typical suspect or victim, helping NCMEC and law enforcement triage reports more effectively while enabling companies to adopt consistent practices for responsible reporting.

Stakeholders can support this effort by building a shared understanding of how red teaming works and how results should be interpreted. Policymakers can help create legal environments that enable clearly defined and robust safety testing.

Third parties can provide support in assessing and mitigating risk, and companies can continue to share insights about red teaming efforts with each other through forums like the Tech Coalition.

The goal in testing AI systems should not be to reward the smallest number of reported failures. The goal should be to ensure that companies test rigorously, interpret results honestly, and use findings to strengthen safeguards that protect children.

The Tech Coalition unites the global tech industry to protect children from online sexual exploitation and abuse (OCSEA).

No single company can tackle this issue alone. But together, we're pooling knowledge, identifying threats, and developing solutions to prevent harm. Because children deserve a digital world that's safe to play, learn, and explore.

technologycoalition.org