ANTHROP\C

A look at Anthropic's key processes, programs, and practices
for responsible AI development.

Anthropic's Transparency Hub

System Trust and Reporting

Last updated January 29, 2026

We are sharing more detail on our Usage Policy, some enforcement data, how we handle legal requests, and our approach to user safety and wellbeing to enable meaningful public dialogue about AI platform safety.

Banned Accounts

1.45 million

Banned Accounts

July - December 2025

Anthropic's Safeguards Team designs and implements detections and monitoring to enforce our Usage Policy. If we learn that a user has violated our policies, including our Usage Policy, Consumer Terms of Service or Commercial Terms of Service (as applicable), or Supported Region Policy, we may take enforcement actions such as warning, suspending, or terminating their access to our products and services.

52k

Appeals

July - December 2025

1.7k

Appeal Overturns

July - December 2025

Banned users may file appeals to request a review of our decision to ban their account.

January - June 2025 Reporting

July - December 2024 Reporting

Child Safety Reporting

5,005

Total pieces of content reported to NCMEC

July - December 2025

Anthropic is committed to combating child exploitation through prevention, detection and reporting. On our first-party services, we employ hash-matching technology and detection classifiers to detect and report CSAM that users may upload.

January - June 2025 Reporting

July - December 2024 Reporting

Legal Requests

Anthropic processes data requests from law enforcement agencies and governments in accordance with applicable laws while protecting user privacy. These requests may include content information, non-content records, or emergency disclosure requests.

For more information, see our full reports here:

January - June 2025 Government Requests for Data

July - December 2024 Government Requests for Data

January - June 2024 Government Requests for Data

Protocol for Addressing Expressions of Suicidal Ideation, Suicide, and Self-Harm Risk

Anthropic is committed to the safety and wellbeing of users who interact with Claude. Please refer to our Protocol for Addressing Expressions of Suicidal Ideation, Suicide, or Self Harm for more information.

Threat Intel Reports

Anthropic monitors and analyzes emerging threats to AI systems and our services to enhance security and inform defensive measures. Our threat intelligence activities help identify attack patterns, malicious actors, and vulnerabilities that could impact AI safety and user security.

Public reports on our findings can be found below:

Disrupting the first reported AI-orchestrated cyber espionage campaign

Detecting and countering misuse of AI: August 2025

Detecting and countering malicious uses of Claude: March 2025

EU Digital Services Act (DSA) Transparency Reporting

Anthropic publishes transparency reports in accordance with its obligations under the EU Digital Services Act (DSA) for certain features of Claude.ai that fall within the scope of the DSA. These reports cover content moderation activities, enforcement actions, and complaint-handling mechanisms for those in-scope features.

For more information, see our full report and summary document here:

EU DSA Transparency Report Overview: May - December 2025

EU DSA Transparency Report: May - December 2025