ANTHROPIC

Anthropic’s Transparency Hub

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

System Trust and Reporting

Last updated August 7, 2025

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.

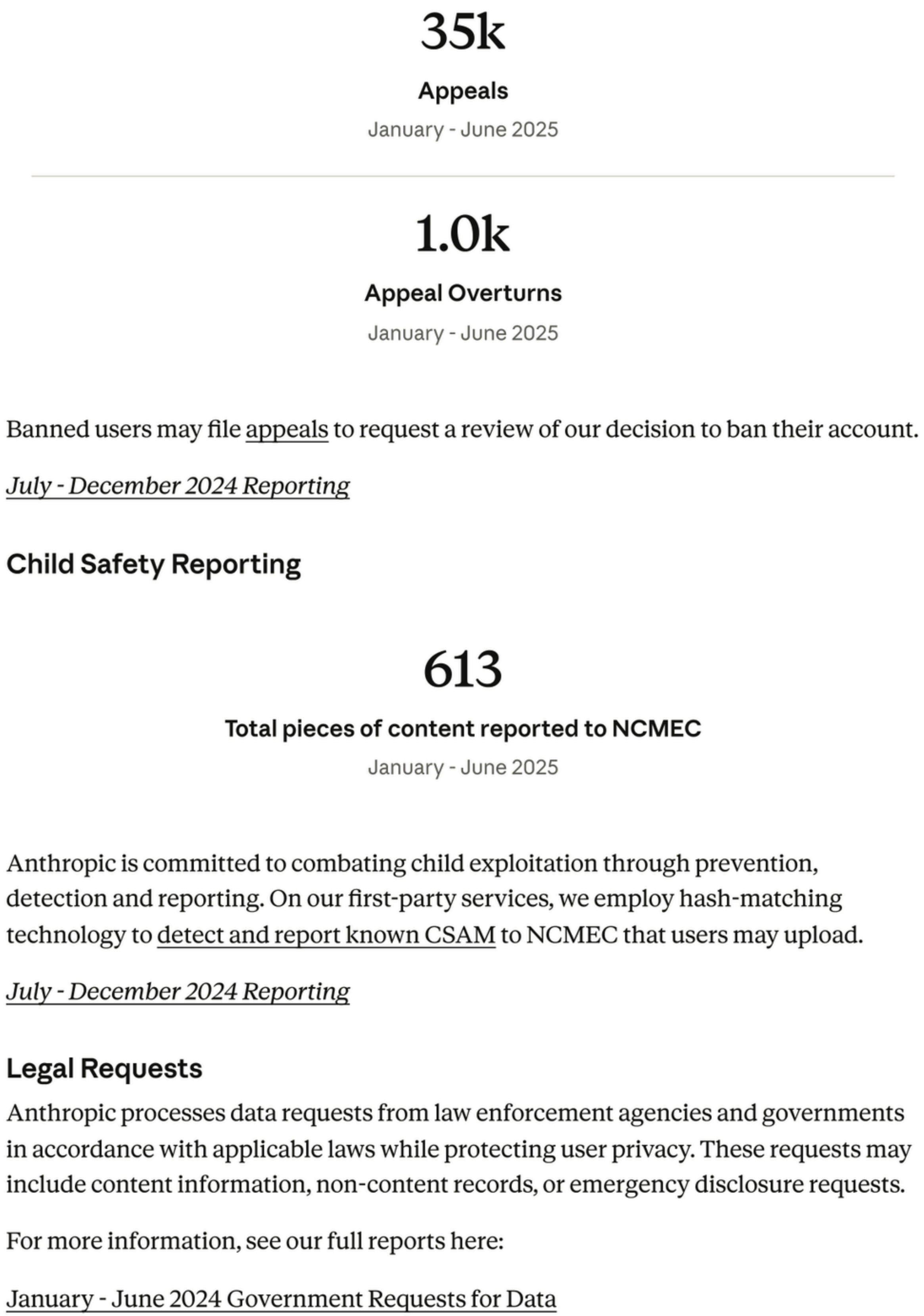
Banned Accounts

690k

Banned Accounts

January - June 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 Usage Policy, we may take enforcement actions such as warning, suspending, or terminating their access to our products and services.



35k

Appeals

January - June 2025

1.0k

Appeal Overturns

January - June 2025

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

July - December 2024 Reporting

Child Safety Reporting

613

Total pieces of content reported to NCMEC

January - June 2025

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

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 2024 Government Requests for Data

July - December 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.