ANTHROP\C

EU Digital Services Act

Transparency Report - Overview

February 2026

Anthropic is an AI safety and research company that builds reliable, interpretable and steerable AI systems. Our mission is to develop AI systems that are safe, beneficial, and understandable. This commitment extends to our responsibility as a provider of digital services in the European Union.

User safety is core to Anthropic's mission of creating reliable, interpretable, and steerable AI systems. We are actively investing in and experimenting with additional safety features to supplement our existing model safety efforts and are working to provide helpful tools to a wide audience while also doing our best to mitigate harm.

Provided below is a link to the completed quantitative and qualitative templates required by Commission Implementing Regulation (EU) 2024/2835 (the "**Implementing Regulation**").

Scope of the report

The report concerns Anthropic's Claude.ai service, some aspects of which fall within the scope of the EU's Digital Services Act (**DSA**), which is provided in the European Union by Anthropic Ireland, Limited. The report covers the period **May 1, 2025 to December 31, 2025**.

The data in the report concerns content moderation engaged in by Anthropic in connection with the use of those aspects of the Claude.ai service within the scope of the DSA, and does not concern those aspects of the Claude.ai service which fall outside the scope of the DSA. While the format of the detailed templates set out in the Implementing Regulation is not always suitable for a service like Claude.ai, Anthropic has attempted to complete the templates and provide relevant data to the extent applicable and reasonably practicable.

Safety-by-design

In addition to the content moderation described in the report for those Claude.ai features that are within scope of the DSA, Claude.ai also includes inherent safety-by-design protections. Although these protections apply to aspects of the Claude.ai service which fall outside the scope of the DSA, they are relevant because they reduce the likelihood that users will encounter potentially illegal or harmful content and they improve overall safety for users of Claude.ai.

Claude models are trained to protect users from harmful and illegal content through a combination of Claude's constitution, human feedback, and prompt classifiers. Claude's constitution is part of our model training process, shaping Claude's behavior to be helpful, honest, and harmless. Claude's constitution gives Claude a set of core ethical principles that it learns to apply through its own reasoning — rather

than simply memorising a list of banned topics, Claude develops judgment about why certain content is harmful and how to recognise user requests for harmful content even when they are disguised through clever framing or roleplay scenarios. Human trainers also review Claude's responses during training, reinforcing better and safer outputs over thousands of iterations. This process gives Claude contextual judgment skills to handle nuanced situations (such as recognising when a user might be distressed and being able to respond with appropriate care) while maintaining a set of firm safety commitments and behaviours. Anthropic monitors Claude's real-world behavior after deployment, using that feedback to improve safety over time, treating it as an ongoing responsibility rather than a problem solved once during training.

Safety by operation

In addition to providing safe outputs through Claude's training, Anthropic detects and enforces against violations of our Usage Policy across user interactions in real-time. Our systems for detection and enforcement are powered by a set of prompted or specially fine-tuned Claude models called "classifiers," which are designed to detect specific types of policy violations. We can deploy a number of different classifiers simultaneously, each monitoring for specific types of harm while the main conversation flows naturally. Along with our classifiers, we also employ specific detection for child sexual abuse material (CSAM), in which we compare hashes of uploaded images against databases of known CSAM on our first-party products.

Finally, for services within the scope of the DSA, Anthropic operates a dual-pillar framework comprising proactive detection systems and reactive user reporting mechanisms, designed to ensure compliance with applicable laws and our Usage Policy. Anthropic and its third-party search provider apply URL-based enforcement against illegal and policy-violating content, such as CSAM. In addition, Anthropic takes account-level actions for violative accounts. We have reported on Anthropic's proactive detection systems in our report, as well as URL-based enforcement engaged in by our third-party search engine provider at our direction. You can find our completed quantitative and qualitative templates required by Implemented Regulation concurrently with this Overview on our Transparency Hub.