OpenAI DSA Transparency Report
Qualitative Information

1. Introduction

The qualitative information below and the quantitative data posted alongside it on our Trust and Transparency page constitute OpenAI's Transparency Report (the "Report"), provided in accordance with the EU Digital Services Act ("DSA"). The Report includes data for content, users and reporters, as applicable, from EU member states, covering the period from February 17, 2024 through December 31, 2024 (the "reporting period"). Our policies and practices continue to evolve in conjunction with our services themselves, as well as environmental factors and patterns of potential abuse. We welcome the opportunity to discuss such changes through future reports.

Unless it is reported here or in the quantitative data, OpenAI has no data to report for certain categories of content moderation contemplated by Article 15 DSA (for example because it received 0 requests from Member States pursuant to articles 9 and 10 DSA in the relevant period).

OpenAI's mission is to ensure that artificial general intelligence benefits all of humanity. Avoiding the spread of illegal and harmful content via our services is a key part of this commitment.

Relevant to the DSA, we invest in a range of policies, processes and tools that address instances where users may provide and/or publish content that violates the law and/or our policies, as well as violative third-party content that may appear in our ChatGPT Search service. These combine algorithmic and human interventions, as well as a mix of proactive detection and responses to external notifications.

For context, it may be helpful to understand the following OpenAI services that are discussed in this Report:

- ChatGPT: The advanced conversational AI service offered by OpenAI. This Report covers cases where user accounts are terminated for content-related violations of law or policy, as well as feature-specific moderation.
- ChatGPT Search: A tool made available within ChatGPT that allows users to access real-time information from the web.

- Third-party “GPTs”: Tailored versions of ChatGPT that users can create by providing specific instructions, knowledge and—in some cases—access to third-party content or APIs.
- The “GPT Store”: A service allowing users to access different GPTs via our directory at https://chatgpt.com/gpts .
- The OpenAI Developer Forum: An online message board that allows developers to ask questions and get help building with the OpenAI platform.
- The OpenAI Forum: An invitation-only online message board that brings together domain experts and students to discuss and collaborate on the present and future of AI.

2. Content moderation engaged in at the provider’s own initiative

2.1. Summary

OpenAI’s Terms of Use and Usage Policies broadly define the behaviors that are expected of users, and on which their continued access to our services is conditioned. When we become aware of possibly violative content provided by a user—either through our own detection methods, through user reports, or via external reports—we evaluate whether the content violates those terms and policies. If it does not, and a user or external entity has alleged a legal violation, we evaluate it under the applicable law. Per this approach, all of the content moderation we performed at our own initiative during the reporting period was on the basis of terms and conditions.

Depending on the service, the illegality or policy at issue, the severity of the violation, and the user’s track record (*e.g.* if they are a repeat offender), we may apply one of a number of actions, including:

- For egregious violations, such as uploading of child sexual abuse material (“CSAM”) or repeatedly uploading other violative content, terminating the user’s OpenAI account.
- Preventing a user from sharing a ChatGPT conversation via our in-product link sharing functionality.
- Disabling the link to a conversation a user has already shared.
- Blocking a particular reported search result from appearing in ChatGPT Search.
- Disallowing a GPT from being accessed by anyone other than the user who created it.
- Disallowing a GPT from being accessible via the GPT Store.
- Disallowing a GPT from being featured on the GPT Store home page.
- Removing posts from and disabling users’ access to the OpenAI Developer Forum and OpenAI Forum.

2.2. Detection methods

As noted in the Introduction, OpenAI uses both algorithmic means and human intervention to detect and address content provided by users that may violate our Terms of Use and Usage Policies.

We apply a specially trained version of our public Moderation API across a number of surfaces, including shared ChatGPT conversations and GPTs. This API builds upon our foundation models to evaluate whether text or images fall into one of a number of categories of potential harm, *e.g.,* harassment, self-harm, sexual content involving minors, and violence. More information on the public version of this API can be found at https://platform.openai.com/docs/guides/moderation#content-classifications .

In addition, all video and image content uploaded to OpenAI consumer services is scanned with Thorn's CSAM classifiers and compared to a database of hashes of known CSAM, as part of a partnership with Thorn announced in April, 2024. (*See* https://openai.com/index/child-safety-adopting-sbd-principles/ .) When the classifiers identify potential CSAM, the relevant content is temporarily removed and reviewed by a human moderator. If the moderator determines it is illegal, we permanently remove it from our services and report it to the National Center for Missing and Exploited Children ("NCMEC"). If it does not meet the legal bar, we analyze it under our Usage Policies and act accordingly.

We also apply specialized classifiers to GPT Store entries designed to detect submissions that may be impersonating known brands in an effort to deceive users.

Human investigation and reporting is also an important means of detection. We believe that learning from real-world use is a critical component of creating and releasing increasingly safe AI systems. We cannot predict all beneficial or abusive uses of our technology, so we proactively monitor for new abuse trends. We have engaged human investigators in targeted efforts to monitor specific services for specific categories of potential harm. For example, we proactively search for activity that violates our usage policies and could be used to defraud or scam others and take action on the associated account or accounts to safeguard our users.

We also accept reports from external sources—both as required under the DSA and through voluntary means. For example, the ChatGPT and GPT Store user interfaces provide entry points for users to flag items as potentially illegal or in violation of our terms and conditions. In addition, we launched a comprehensive content reporting webform at https://openai.com/form/report-content/ .

3. Content moderation by automated means

Beyond the various automated detection methods discussed in Section 2.2 above, there are two specific areas where OpenAI takes automated action on user accounts and user-provided content:

- An automated system tracks each account’s history of policy violations, automatically sending the user warning messages indicating the possibility of account termination. In cases of repeated misuse of a given service, we may disable access to that service and/or terminate the account overall. The purpose of this automation is to provide users of the service with notice and education regarding our terms and conditions at scale, while also ensuring timely and consistent action on those users who nonetheless repeatedly violate those terms.

- GPTs may be automatically removed from the GPT Store and/or blocked from sharing due to policy violations, such as GPTs that are not appropriate for all ages, hateful or otherwise harmful. The purpose of this automation is to take timely action on GPTs that may violate our policies. By automating detection and action, we are able to catch violations at multiple key milestones, including when a GPT is first made shareable with other users and when a GPT is submitted for inclusion in the GPT Marketplace.

3.2 Indicators of accuracy and possible rate of error

The individual violations potentially leading to account termination are typically a result of manual review by human agents trained on our Usage Policies. The agents’ performance is regularly reviewed by internal experts, who provide additional training and guidance on any patterns of inaccuracy.

For automated moderation of GPTs, we largely rely on the same technologies powering our public Moderation API, which we routinely evaluate for precision and improvement. For example, in September 2024, we announced a new version of the model with 42% improvement measured by “area under precision recall curve” in our internal multimodal eval. This improvement was consistent across 98% of languages tested, with much of it focused on non-English languages. While the previous model had limited support for non-English languages, the performance of the new model in Spanish, German, Italian, Polish, Vietnamese, Portuguese, French, Chinese, Indonesian, and English all exceed even English performance from the previous model. *See* https://openai.com/index/upgrading-the-moderation-api-with-our-new-multimodal-moderation-model/ .

3.3 Safeguards

We periodically review the outputs of the relevant systems for accuracy using human raters. This process allows us to identify and remedy any areas where the classifiers may not be performing well. We also generally provide users whose accounts have been terminated with an opportunity to appeal the action. Such appeals are reviewed by a human agent.