DSA Transparency Report
Cici - 2025

Introduction

Cici is an AI chat assistant designed for intelligent conversations, writing, translation, and programming.

In line with our obligations under the Digital Services Act (**DSA**), we are pleased to publish our first DSA transparency report for the reporting period of 17 February 2024 to 31 December 2024.

Report index

Section 1. Content moderation

We review content uploaded by our users proactively (through systems we have in place which detect illegal and harmful content, including content which may be in violation of our Terms of Service or our Community Guidelines (together, our **Policies**)) and reactively (for example, on receipt of notice from users or authorities). To do this we deploy a combination of technology and human moderators.

- We use automated moderation technology to identify content that violates our Terms of Service or Community Guidelines. This technology looks at a variety of signals across content, which may include, for example, keywords or images to detect potential violations.

- Human moderators work alongside our automated moderation systems to review and assess content that may violate our Terms of Service or Community Guidelines. Our moderators undergo regular training on our content moderation processes and policies.

We may remove or restrict access to content if we reasonably believe it is in breach of our Terms of Service or our Community Guidelines.

Key Principles

Content that is uploaded on Cici is typically first reviewed by our automated moderation technology, which aims to identify content that violates our Policies before it is viewed or shared by other people on Cici or reported to us. While undergoing this review, the content is visible only to the uploader.

If our automated moderation technology identifies content that is a potential violation, it will either be automatically removed from Cici or flagged for further review by our human moderation teams.

Automated Review

We use a variety of automated tools, including:

- Computer Vision models, which help to detect objects (for example visual signals, emblems, logos objects that are known to be associated with extremist and hate groups) so it can be determined whether the content likely contains material which violates our policies.
- Keyword lists and NLP models are used to review text content to detect material in violation of our policies.

Human Moderation

Cici currently adopts proactive risk detection models that surface content that potentially violates our Community Guidelines for human moderation review. These models are trained on content that have been tagged to specific policies by our human moderators. Tagged content undergoes random quality assurance samples to ensure tagging accuracy and alignment to policy. This ensures our datasets remain diverse, fair and unbiased.

Moderation data is continuously monitored through random sampling, with regular meetings held to discuss potential grey area cases. If required, interpretation guidelines are published for moderators to revise and align with policy.

If proactively detected content has been confirmed by human moderators to be in violation of our Community Guidelines, users will be informed of the fact.

Section 2. Illegal content reports

Our Policies apply to all accounts and content on Cici, and they often align with, and sometimes go beyond, local law requirements. While we primarily enforce our Policies at our own initiative through automated and human moderation, users can also alert Cici to content they believe violates our Policies or is illegal. Cici received no illegal content reports in the European Union during the period from 17 February 2024 to 31 December 2024.

Section 3. Our Moderators

To ensure a consistent understanding and application of our Policies, all content moderator personnel receive training across our relevant Policies. All content moderators undergo training on Cici's content moderation systems.

Content moderation training materials are kept under review to ensure that they are accurate and current.

Members of our Trust & Safety teams attend regular internal sessions dedicated to knowledge sharing and discussion about relevant issues and trends.

Section 4. Orders from government authorities

We may receive requests from government authorities in the European Union to remove content. During the period from 17 February 2024 to 31 December 2024, we received no requests from government authorities in the European Union to remove content.

We may also receive requests from government authorities in the European Union for user information disclosure. During the period from 17 February 2024 and 31 December 2024, we received no information requests from government authorities in the European Union.

Annex A - Cici's own-initiative content moderation

This Annex A provides the number of moderation actions we took against content and accounts under our Policies.

This table sets out the number of the content-level moderation actions taken where content is found to violate our Policies, broken down by the moderation action taken.

Number of moderation actions taken under our policies by type	
Content Removed	**Content Restricted**
2,640	456

This table sets out the number of the content items removed where content is found to violate our Policies, broken down by the sub-policy under our Community Guidelines. Content may violate multiple policies and each violation is reflected in the breakdown of each of the respective sub-policies. The total number of profiles removed by automation was 438.

Policy Category	Number of profiles removed
High Risk & Regulated Activities	192
Harmful Misinformation	1
Violent Behaviors & Dangerous Actors	3
Mental Health	100
Deceptive Behaviors	830
Harassment & Hateful Behavior	141
Shocking & Graphic Content	24
Other Bots	188
Nudity & Sexual Activity	871
Exploitation & Abuse	146
Total	**2,640**

This table sets out the number of the content items restricted where content is found to violate our Policies, broken down by the sub-policy under our Community Guidelines. Content may violate multiple policies and each violation is reflected in the breakdown of each of the respective sub-policies.

Policy Category	Number of Profiles restricted
Deceptive Behaviors	3
Mental Health	1
Shocking & Graphic Content	10
Nudity & Sexual Activity	439
Other Bots	3
Total	**456**