DIGITAL SERVICES ACT

Transparency Report

Article 15 DSA

SERVICE PROVIDER	EverAI Ltd.
REPORTING PERIOD	01 January 2025 – 01 January 2026
VERSION	1.0

1. Introduction

This transparency report is published in accordance with Article 15 of Regulation (EU) 2022/2065 (Digital Services Act) ("DSA"). It covers content moderation activities undertaken by EverAI Ltd. during the reporting period specified above.

This report is made publicly available in a machine-readable format and provides clear, comprehensible information on content moderation activities as required by the DSA.

2. Orders from Member States' Authorities *(Art. 15(1)(a))*

No orders from Member States' authorities were received during the reporting period.

3. Notices Submitted *(Art. 15(1)(b))*

No notices regarding illegal content or notices from Trusted Flaggers were received during the reporting period.

4. Own-Initiative Content Moderation *(Art. 15(1)(c))*

4.1 Overview of Content Moderation Activities *(Art. 15(1)(c), (e))*

We use a combination of automated tools and human review to prevent, detect, and respond to inappropriate user-generated content. This may include preventing the production of content, escalating flagged accounts for human review, or terminating users' accounts. These tools assist in the detection of violations and enforcement of rules relating to restricted content within user-generated content and activities on the platform.

Examples of tools we use include:

- Automatic scans of inputs and outputs for content that violates our policies, including via our proprietary moderation LLM
- Technical measures that block and prevent AI tools from fulfilling requests that violate our policies
- Tools that detect fraudulent activity, use of the platform, or account creation
- Human reviews of suspicious profiles or activity

If our automated moderation controls detect anything that violates our policies, we or authorized partners may access and manually review the flagged content and/or other content associated with the account and take appropriate action. More information is available in our Terms of Service (Sec. 5.4) and Blocked Content Policy. Such automated moderation is used for the following purposes:

- Detecting content that may violate our Terms of Service, Community Guidelines, or other Policies
- Preventing misuse of platform features, including requests for and/or attempts to generate prohibited content through text or image-generation functionalities
- Supporting users' safety when accessing AI tools
- Assisting in the enforcement of platform rules through automated flagging, restriction, or other moderation actions

None of our content moderation activities constitute automated decision-making within the meaning of Art. 22 GDPR.

Decisions to block users for content-related violations are always made by trained moderators. Human moderators making these decisions are experienced professionals who receive extensive training internally to ensure they are well-prepared for their responsibilities. Managers conduct thorough reviews of moderators' completed work and monitor performance on an ongoing basis, providing additional coaching, guidance, and/or timely feedback as needed.

4.2 Indicators of Accuracy and Possible Rate of Error *(Art. 15(1)(e))*

We assess the performance of automated moderation tools through ongoing testing and evaluation using internal validation datasets, policy-specific benchmark sets, and operational monitoring.

The primary indicators used to evaluate performance include:

- Precision (the proportion of flagged content that is correctly identified as policy-violating)
- Recall (the proportion of policy-violating content successfully detected)
- False-positive rate (content incorrectly identified as violating policy)
- False-negative rate (policy-violating content not detected by the system)
- Consistency across policy categories and languages

The possible rate of error depends on a number of factors, including the nature of the content being assessed, the complexity and ambiguity of language, the use of slang, coded language, satire, or fictional roleplay, the specific policy category being evaluated, and the extent to which relevant context is available to the system at the time of assessment.

We apply a number of safeguards to automated moderation, including:

- Regular testing and validation of moderation LLMs
- Regular review and updating of moderation policies and classification criteria
- Monitoring of moderation outcomes and identified error patterns
- User-facing mechanisms for contesting moderation decisions
- All account termination decisions for legal or policy violations are made only following human review
- Access controls and security measures governing the operation and deployment of moderation systems

These safeguards are intended to improve the reliability, proportionality, and fairness of content moderation processes while maintaining platform safety and compliance obligations.

4.3 Number and Type of Measures Taken *(Art. 15(1)(c))*

Our products and services provide users with access to AI-generated text, visual media, and audio, and tools to create the same. All text inputs from users and outputs from the AI – over 2,000,000 daily – are subject to moderation through our proprietary moderation LLM. Moderation categories include, for example, protection of minors, violence, hate speech, and references to illegal substances/activities.

Measures we take, which affect the availability, visibility and accessibility of information provided by users and the users' ability to provide information through the service are those measures that result in:

- Removal of user content
- Terminating accounts
- Blocking re-registration of users

Content / Violation Type	Content Removal	Account Termination / Blocking Re-registration
Protection of Minors	0	508
Violence	0	0
Illegal substances/activities	0	0
Hate speech	0	0
Other terms violation	42	n/a
Fraud or Abuse	n/a	228,390

Note: "Content Removal" refers to content removed without also disabling or blocking an account

5. User Reports under Art. 16 *(Art. 15(1)(d))*

During the reporting period, we did not receive notifications through our internal complaint-handling systems from individuals or entities regarding the presence of information they considered to be illegal content.