digital terrorism+hate
report card 2025

SIMON
WIESENTHAL
CENTER

Rubric for Grading Social Media Platforms on Handling Hate, Antisemitic and Terrorist Content

Overview

The Simon Wiesenthal Center Research Department Rubric evaluates social media platforms' effectiveness in addressing hate speech, antisemitic content, and terrorist-related material. It assesses nine key criteria, each scored on a 0–4 scale, for a total of 36 points. The rubric provides a structured framework for grading platforms based on their policies, responsiveness, transparency, compliance with legal and international standards, and engagement with trusted flaggers, with final grades assigned as A through F.

Rubric Criteria and Scoring

Criteria	4 (Excellent)	3 (Good)	2 (Fair)	1 (Poor)	0 (Non-Existent)
Responsiveness to Reported Hateful Material	Platform removes reported hate, antisemitic, or terrorist content within 24 hours consistently, with clear user feedback.	Removes content within 48 hours in most cases, with occasional delays; provides some user feedback.	Removes content inconsistently, often taking over 48 hours; minimal feedback to users.	Rarely removes reported content or takes excessive time (>72 hours); no user feedback.	No process for removing reported content.
Violent Extremism Policy	Comprehensive policy explicitly addressing violent extremism, with clear definitions, enforcement mechanisms, and regular updates.	Policy exists but lacks specificity or is not regularly updated; enforcement is inconsistent.	Vague policy with limited enforcement; minimal focus on violent extremism.	Policy mentioned but not enforced or poorly defined.	No violent extremism policy.

IHRA-Specific Policy	Adopts IHRA working definition of antisemitism fully, integrates it into content moderation, and trains staff on its application.	Adopts IHRA definition but with partial integration into moderation; limited staff training.	References IHRA definition but does not integrate it into moderation or training.	Mentions antisemitism vaguely without referencing IHRA or applying it.	No policy addressing antisemitism or IHRA definition.
Adherence to Terms of Service (TOS) Agreements	Strictly enforces TOS, consistently penalizing violations (e.g., suspensions, bans) with transparent	Enforces TOS in most cases but with occasional inconsistencies; appeal process exists but is slow.	Inconsistent TOS enforcement: appeal process is unclear or limited.	Rarely enforces TOS; no functional appeal process.	No enforcement of TOS or TOS is absent.
Transparency Report	Publishes detailed, regular (e.g., quarterly) transparency reports with specific data on hate/terrorist content removals, appeals, and trends.	Publishes annual reports with moderate detail on content moderation but lacks specificity or frequency.	Publishes infrequent reports with vague or incomplete data on content removals.	Publishes minimal or unclear reports with no actionable data.	No transparency reports published.
Digital Services Act (DSA) Compliance	Fully compliant with DSA requirements, including content moderation, transparency, and user redress mechanisms; audited regularly.	Mostly compliant with DSA but with minor gaps in implementation or reporting.	Partially compliant with DSA; significant gaps in moderation or transparency.	Minimal compliance with DSA; major gaps in adherence.	Non-compliant with DSA or no evidence of compliance.
Emergency Plan for Crisis Situations	Maintains a robust emergency plan for rapid response to spikes in hate/ terrorist content, with dedicated teams and real-time monitoring.	Has an emergency plan but with limited activation or slower response times (e.g., >24 hours).	Emergency plan exists but is vague or rarely activated; slow response to crises.	Minimal emergency measures; ineffective or ad hoc responses to crises.	No emergency plan for handling content spikes.
Mutual Legal Assistance Treaty (MLAT) Requirements with Law Enforcement	Collaborates efficiently with law enforcement under MLAT, balancing user privacy with timely data sharing for investigations.	Cooperates with MLAT requests but with occasional delays or inconsistent processes.	Limited cooperation with MLAT; slow or selective responses to law enforcement.	Rarely complies with MLAT requests; significant barriers to cooperation.	No cooperation with MLAT or law enforcement.

Engagement with Trusted Flaggers	Maintains a formal trusted flagger program with prioritized review, regular communication, and high removal rates (>90%) for flagged content.	Has a trusted flagger program with moderate prioritization and communication; removal rates are inconsistent (70–90%).	Informal or limited trusted flagger engagement; low prioritization and removal rates (50–70%).	Minimal engagement with trusted flaggers; rarely acts on flags (<50% removal rate).	No trusted flagger program or engagement.

Scoring Guidelines

- **Total Score**: Sum of the points across all nine criteria (maximum 36 points).

- **Grading Scale**:
  - 34–36 (A): Exemplary performance; platform demonstrates robust, proactive measures and compliance.
  - 32–33 (A-): Near-exemplary performance with very minor gaps in consistency or implementation.
  - 30–31 (B+): Strong performance with slight deficiencies in policy or enforcement.
  - 28–29 (B): Good performance but with noticeable gaps in consistency or implementation.
  - 26–27 (B-): Moderate performance with some significant deficiencies.
  - 24–25 (C+): Fair performance with consistent deficiencies in policy or enforcement.
  - 22–23 (C): Poor performance; significant gaps in addressing hate/terrorist content.
  - 20–21 (C-): Very poor performance; major gaps in multiple areas.
  - 18–19 (D+): Inadequate performance; minimal efforts to address critical issues.
  - 16–17 (D): Severely inadequate; major failures in policy and enforcement.
  - 14–15 (D-): Barely any measures; near-total failure to address issues.

    - 0–13 (F): No meaningful measures; platform fails to address hate, antisemitic, or terrorist content.

Notes

- **IHRA Definition**: The International Holocaust Remembrance Alliance (IHRA) working definition of antisemitism is a widely recognized standard for identifying antisemitic content. Platforms adopting it explicitly are better equipped to address antisemitism systematically.
- **DSA Compliance**: The EU's Digital Services Act (DSA) mandates transparency, content moderation, and user protections. Compliance is critical for platforms operating in the EU.
- **MLAT**: Efficient cooperation with law enforcement via MLAT ensures platforms balance legal obligations with user rights, particularly in terrorism-related investigations.
- **Trusted Flaggers**: Trusted flaggers are vetted organizations or individuals with expertise in identifying harmful content. Platforms with formal programs prioritize their reports, improving response times and accuracy in content moderation.
- **Context**: This rubric is designed for global platforms but can be adapted for regional variations in legal frameworks. Evaluators should consider platform size and resources when assessing smaller platforms.