



Public Comment 2026/10

Respecting Human Rights When Using LLMs in Content Moderation: Key Principles for the Tech Industry

Comment Submitted Thursday, October 15, 2026 by
Tal-Or Cohen Montemayor on behalf of CyberWell Ltd. (CC)

Executive Summary

CyberWell submits this public comment to the Oversight Board to address the use of large language models (LLMs) in content moderation. Drawing on CyberWell's research and AI expertise, we examine **persistent moderation challenges that are relevant to LLMs, including the need for contextual interpretation, moderation across platform surfaces, and inconsistent enforcement against high-reach accounts that repeatedly disseminate hateful content.** We also offer recommendations for how Meta and the broader technology industry can responsibly apply LLMs to content moderation, while protecting freedom of expression and mitigating risks to users targeted by online hate.

CyberWell's Mission

CyberWell is a nonprofit organization dedicated to addressing online antisemitism, by driving the enforcement and improvement of community standards across the digital space. Drawing on our data-driven research, expertise in Arabic and English analysis, and experience developing AI-based solutions to advance online safety, this submission uses Meta as a case study to examine key considerations for the responsible use of LLMs in content moderation across the technology industry.

Introduction: The Use of LLMs in Content Moderation

As platforms explore the use of LLMs to address the growing scale and complexity of content requiring moderation, it is important to consider both the potential benefits and challenges associated with these models. LLMs may enable platforms to review large volumes of content more efficiently, enforce platform policies at scale, improve explanations



to users, and allow human moderation teams to focus on more complex cases requiring human judgement.

At the same time, the integration of LLMs into content moderation raises important considerations regarding their ability to interpret context-dependent content, including coded language, linguistic nuances, and the use of humor and disclaimers to evade moderation. Additional questions arise around the ability of LLMs to identify and enforce policy violations across different platform surfaces (e.g., comments, usernames, audio), while ensuring consistent moderation of content across accounts regardless of follower count or reach. These key considerations should inform how LLM-based moderation systems are designed, evaluated, and supported by human oversight.

Contextual Interpretation

Contextual interpretation remains a key challenge for content moderation, because harmful meaning is not always apparent from individual words or phrases alone. **Coded language, linguistic manipulation across languages, and humor or disclaimers can make harmful content appear benign when assessed without sufficient context.** As LLMs are increasingly applied to content moderation, their ability to interpret these contextual signals is important for distinguishing harmful content from non-violative expressions.

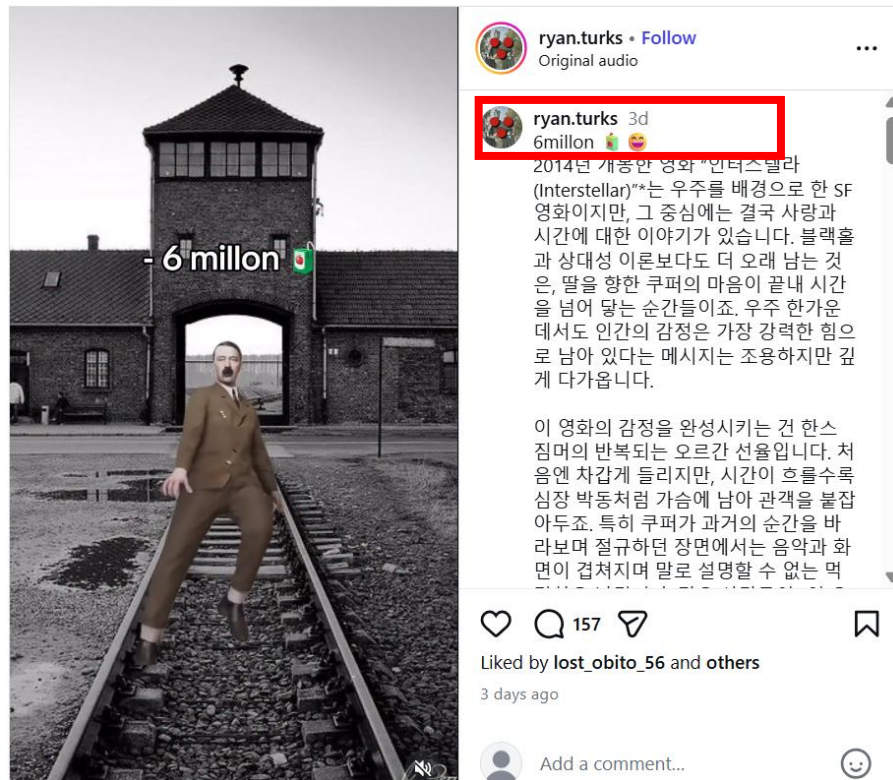
Coded Language

Coded language refers to the use of seemingly neutral words, symbols, and visual references to express hateful messages. Last year, CyberWell submitted a [public comment](#) to the Oversight Board addressing the use of coded language and emojis to spread hate against Jews. Many of the examples documented in this report illustrate how **longstanding antisemitic narratives, rooted in historical stereotypes and prejudices, are repurposed using emojis and coded language to promote modern-day hate.**

For example, in this Instagram post below, the user employs generative AI to celebrate the mass murder of Jews during the Holocaust. The AI-generated video features Adolf Hitler dancing in front of Auschwitz, accompanied by overlaid text that reads: “- 6



million 🍷". The juice box emoji (🍷) serves as a coded reference to Jews through its phonetic similarity to the word "Jews".



The growing use of AI to create and adapt hateful content also raises concerns about users' abilities to develop new methods for evading content moderation. CyberWell's latest research on [AI-generated antisemitism](#) highlights how **coded language is highly context-dependent**, varies across languages and geographic regions, and appears across multiple formats—including text, images, memes, symbols, emojis, and combinations of these elements. As users leverage AI to generate increasingly sophisticated variations of antisemitic narratives, including through coded language and visual references, **LLMs must keep pace with emerging forms of hate and accurately interpret the contextual references necessary to detect policy violations.**

Use of Humor and Disclaimers

Humor, satire, and disclaimers can complicate the detection of antisemitic content, by obscuring hateful intent and presenting policy-violating

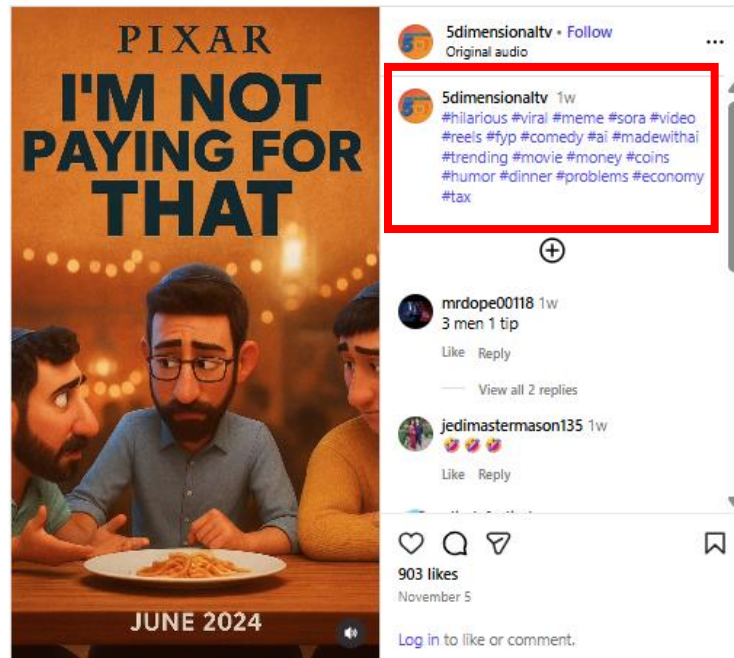


narratives as jokes or educational material. For example, **Holocaust-related mockery**, which involves ridiculing the Holocaust, its historical reality, or its victims presents moderation challenges when hateful content is disguised through humor. Coded phrases such as [“6 million pizzas”](#) are used to mock Holocaust victims, and promote Holocaust denial or distortion. These forms of antisemitism also often appear in youth-oriented formats, such as gaming videos and fabricated Disney-Pixar movie trailers.

This AI-generated Facebook post below illustrates this phenomenon. Presented as a news report examining the feasibility of baking six million pizzas in five years, the video uses this analogy to cast doubt on the murder of six million Jews during the Holocaust under the guise of humor.



Furthermore, **the use of disclaimers such as “#satire”, or “for educational purposes”** may be used to frame content containing antisemitic narratives and references as humorous or informational, [obscuring a post’s hateful nature](#). The Instagram post below promotes antisemitic stereotypes that depict Jews as greedy and money obsessed. The caption includes multiple hashtag-based disclaimers, such as **“#hilarious [...] #meme [...] #comedy [...] #humor [...]”**, which frame the content as humorous. These disclaimers appear despite the presence of antisemitic rhetoric and stereotypes, functioning to downplay the harmful nature of the content and reduce the likelihood of moderation.



These examples highlight the importance of ensuring that LLMs can assess the underlying meaning of content beyond its humorous or educational framing. This includes distinguishing between legitimate humor and content that deliberately uses humor or satire to disguise antisemitic narratives and evade moderation. To support this capability, **model training and evaluation should incorporate examples that capture the contextual interpretation of coded historical references, visual and audio elements, and accompanying disclaimers that would enable LLMs to better identify content that uses humor or satire to disguise antisemitic narratives.**

Arabic Linguistic Manipulation

Arabic-language content can apply **subtle modifications to letters, punctuation, or word structures to evade moderation when sharing violative content.** These techniques are sometimes used to facilitate the dissemination of content produced by Dangerous Organizations and Individuals [hereinafter: “DOI”], including through news accounts where newsworthiness exceptions may also affect enforcement outcomes.

For example, the post below, published by a news account amplifies Hamas messaging and promotes violence against Jews following a terrorist attack that killed three



Israelis. The caption refers to the perpetrator as a “heroic martyr”, violating Meta’s DOI policy, as [updated](#) following the recommendations adopted following the [Oversight Board’s Policy Advisory Opinion](#), on use of “shaheed”. The caption also employs a linguistic evasion tactic: inserting dots between Arabic letters in the terms “killing”, “shaheed”, and “al-Qassem Brigades”, altering their written form in a manner that evades automated moderation. This example underscores **the importance of human oversight in LLM-based moderation, particularly when nuanced linguistic interpretation is required to identify policy violations in non-English languages.**



Moderation Across Platform Surfaces

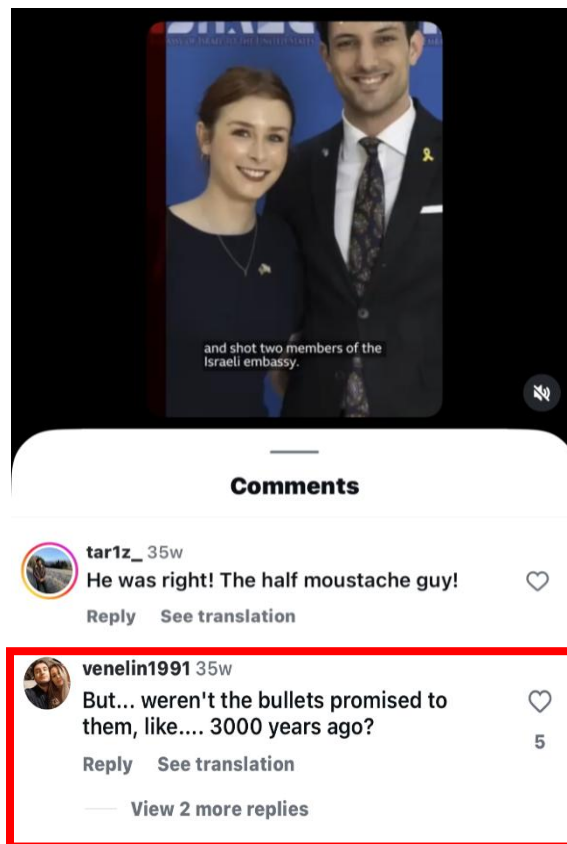
Content moderation challenges currently extend across the different surfaces and formats through which users interact with a platform. **Harmful signals can appear in the primary text of a post, but may also exist in comment sections, account-level features like usernames, and audio.** The integration of LLMs into content moderation underscores the need for platforms to ensure that these models can accurately identify and interpret harmful signals across different platform surfaces and content formats.

Comment Sections



Comment sections present distinct moderation challenges as antisemitic narratives can migrate between posts and comments. CyberWell documented this trajectory with the [“promised 3,000 years ago” trend](#), which mocks and attempts to delegitimize the Jewish historical connection to the land of Israel, by portraying Jews as inherently entitled and greedy. In the aftermath of antisemitic attacks, CyberWell observed this phrase appearing in the comment sections of news outlets, where it was used to mock victims and justify violence against them.

In the comment section of the Instagram post below, published by a news account reporting on the Jewish victims killed during the [Capitol Jewish Museum shooting](#), users glorify Hitler through the coded reference “The half moustache guy” and invoke the “promised 3,000 years ago” trope to justify the antisemitic violence and mock its victims.

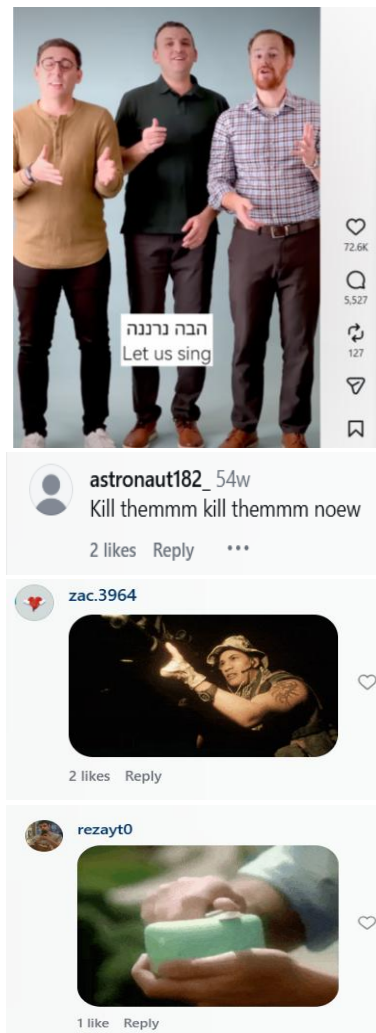


While some platforms may not consider the “promised 3,000 years ago” trend a policy violation when used in posts, its appearance in comments responding to reports of antisemitic violence may constitute a violation. This distinction underscores the importance of evaluating whether LLM-based moderation can **interpret emerging forms of hate in comments, at scale, within the context of the posts they respond to, rather than evaluating their language in isolation.** This is equally



important for protecting Jewish content creators from targeted antisemitic harassment in their comment sections.

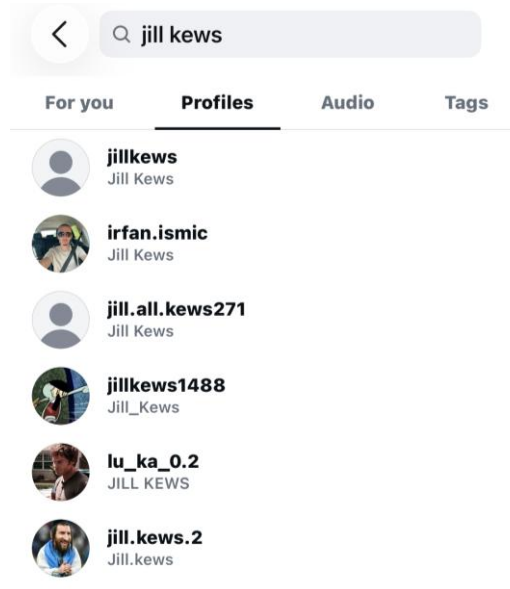
The ability of LLMs to identify these violations at scale depends on their capacity to interpret visual content in the context of the original post and the creator's identity to determine whether it constitutes targeted harassment. The example below, taken from the comment section of an Instagram post shared by the Jewish acapella group The Maccabeats, illustrates how **users employ GIFs and images to mock the Holocaust and incite violence against Jews**. References to soap bars invoke the [theory](#) that Nazis manufactured soap from the bodies of murdered Jews, while GIFs depicting machine guns being fired convey calls for violence against Jews.



Username



Username, display name, and handles provide important account-level signals for LLM-based content moderation, particularly when coded language is used to communicate hateful or violent intent. For example, the Instagram usernames below use **“Jillkews”** a simple linguistic manipulation of the words “Kill Jews” to communicate a call for violence against a protected group.



A username can provide evidence of hateful or violent intent, while also offering important context for interpreting an account’s broader activity. For LLM-based moderation, **usernames highlight the importance of assessing whether violations identified at the account level should prompt an automatic review of the account itself and its broader activity.**

Audios

Audio poses an additional moderation challenge when **violative content circulates across posts in different versions or clips of varying lengths**. During periods of global conflict, audio can be used to call for violence against protected groups. For example, the song [“Boom, Boom Tel Aviv”](#), which justified violence against Jews, circulated across major social media platforms during and after the June 2025 Israel-US-Iran War.



Detecting these violations can be particularly challenging, when lyrics are in non-English languages like Arabic. **Transcription alone may not capture culturally specific references, coded terminology, or the broader context necessary to identify hateful or violent meanings.**

CyberWell identified an Arabic-language song by Syrian singer Abdullah Al-Salmou circulating across Facebook and Instagram. The song's lyrics, "I swear, O sons of Jews, your killing is near [...] you will be buried in Tel Aviv" explicitly threaten violence against Jews, while invoking the seventh-century Battle of [Khaybar](#), a reference often used to encourage violence against Jews.



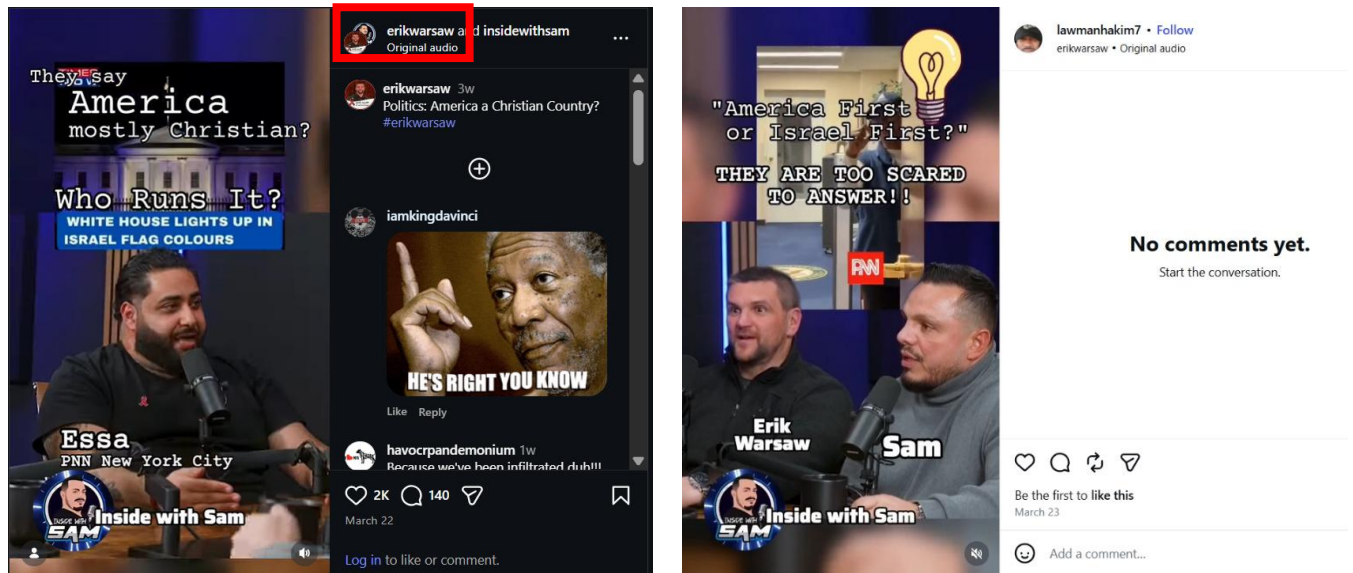
The integration of LLMs into content moderation therefore raises concerns about their ability to interpret audio alongside text-based content. An additional challenge arises **when the same audio circulates in multiple versions or as clips of varying lengths, complicating consistent enforcement as shorter excerpts omit key lyrics or contextual references.**

Inconsistent Enforcement for Accounts with Large Followings

Consistency in enforcement remains an important challenge in content moderation, particularly when the same or similar content receives different enforcement outcomes depending on the account sharing it. These inconsistencies can be consequential for high-reach accounts, where content that remains available has the potential to reach significantly larger audiences.



The two posts shown side-by-side below feature the same post promoting the antisemitic trope alleging that Jews control the U.S. government. Despite both posts being escalated by CyberWell through Meta's Trusted Partner channel, enforcement was inconsistent: the repost (right) was removed immediately, while the original post (left), shared by influencer Erik Warsaw with over 301,000 followers remained available even after a series of appeals.



As LLMs become more integrated into content moderation, this disparity raises questions about how automated systems can support consistent enforcement across accounts, particularly when multiple users share identical content. **LLMs should support the identification of original sources of violative content, detection of subsequent reposts, and at-scale removal of such content.** Evaluating moderation outcomes across original posts and reposts, including those shared by accounts with different follower counts and reach, can help assess the consistency of policy enforcement within LLM-based systems.

Recommendations

CyberWell recognizes the potential benefits and challenges of using LLMs in content moderation. We offer the following recommendations in response to the Board's requests:



I. Collaborate with Subject-Matter Experts to Strengthen Contextual Interpretation

Platforms should **consult with independent subject-matter experts like CyberWell**, to **inform how LLMs are trained, prompted, and evaluated to identify emerging forms of hate while reducing social biases that may be embedded in existing classification systems**. External experts can contribute specialized knowledge, real-world examples, annotated datasets, and test cases to strengthen model training and evaluation, improve detection accuracy, and reduce the risk of incorrectly classifying legitimate expression as violative content.

II. Ensure Consistent Moderation Across Platform Surfaces and Multimodal Formats

CyberWell encourages Platforms to ensure that **LLMs can identify and assess violative content across different platform surfaces**, including comments, usernames, and display names. LLMs should also **support multimodal analysis** of text, images, GIFs, videos, and audio, incorporating relevant contextual information to identify policy violations that may not be apparent when assessed in isolation.

III. Strengthen the Accuracy of LLM-Based Moderation in Non-English Languages

CyberWell recommends that Platforms conduct **language-specific evaluations** of LLMs to assess their ability to accurately interpret linguistic and cultural nuances, including dialects, slang, irony, and coded language. These evaluations should assess whether models can reliably interpret multilingual and multimodal content, **including cases where inaccurate transcriptions, text extraction, or incomplete visual and audio analysis may lead to incorrect moderation decisions**.

IV. Establish Clear Escalation Pathways for Human Review of Context-Dependent Content

CyberWell encourages Platforms to establish clear criteria for escalating cases involving ambiguous or context-dependent content to human moderators. Human oversight is especially important in the aftermath of violent incidents and during rapidly evolving global events, when **emerging narratives and contextual**



developments may complicate the identification of violative content. Meta should establish effective coordination between LLMs and human moderators to facilitate timely review and removal of violative content before it accumulates significant views and engagement, with consistent application across platform surfaces and multimodal formats.

V. Increase Public and User Transparency

CyberWell encourages Platforms to increase public transparency through regular reporting on the use and performance of LLMs, including enforcement outcomes, error rates, and appeals. Reporting should also **identify specific categories of harmful content**, including emerging forms of hate that existing classification systems struggle to detect, alongside measures being taken to address these gaps and adapt models to evolving threats. Meta should also provide users with **content-specific explanations for enforcement decisions**, identifying which elements of the content (e.g., text, image, audio, or video) were found to violate which policy. Prompt engineering can support more informative removal notices and improve users' understanding of moderation decisions.