frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 23 January 2026
DOI 10.3389/fcomm.2025.1729279

OPEN ACCESS

EDITED BY
Jurate Ruzaite,
Vytautas Magnus University, Lithuania

REVIEWED BY
Elena Negrea-Busuioc,
National School of Political Studies and Public Administration, Romania
Ledia Kazazi,
Aleksandër Xhuvani University, Albania

*CORRESPONDENCE
Matthias J. Becker
✉ mjb@decoding-antisemitism.eu
Jordan Blatter
✉ jordanb@bsqa.org
Oksana Stanevich
✉ oksana.stanevich@gmail.com

RECEIVED 21 October 2025
REVISED 25 November 2025
ACCEPTED 28 November 2025
PUBLISHED 23 January 2026

CITATION
Becker MJ, Blatter J and Stanevich O (2026) Decoding antisemitism online: linguistic and multimodal challenges in the age of AI. *Front. Commun.* 10:1729279. doi: 10.3389/fcomm.2025.1729279

Decoding antisemitism online: linguistic and multimodal challenges in the age of AI

Matthias J. Becker[1,2,3]*, Jordan Blatter[4]* and Oksana Stanevich[3,5]*

[1]Center for the Study of Antisemitism, New York University, New York, NY, United States, [2]Woolf Institute, University of Cambridge, Cambridge, United Kingdom, [3]AddressHate, New York City, NY, United States, [4]Blue Square Alliance Against Hate, Foxborough, MA, United States, [5]Social Sciences, Tel Aviv University, Tel Aviv-Yafo, Israel

Introduction: This article investigates the linguistic and computational challenges of detecting antisemitism in digital communication, integrating discourse-analytical and artificial intelligence (AI) perspectives. It conceptualizes antisemitic discourse as a continuum ranging from explicit incitement to implicit, coded expressions whose interpretation depends on contextual, cultural, and pragmatic knowledge.
Methods: The study draws on empirical case studies from the *Decoding Antisemitism* project, analyzing YouTube reactions to two events: the Hamas terror attack of 7 October 2023 and the antisemitic double murder in Washington, D.C., in May 2025. Qualitative discourse analysis is combined with computational considerations related to annotation practices and model design for automated detection.
Results: The analysis shows that antisemitic discourse has become normalized in mainstream digital spaces. Reactions to 7 October were characterized by open glorification of violence, whereas responses to the Washington case centered on denial, irony, and the inversion of victimhood. Together, these cases illustrate both the normalization and diversification of antisemitic communication online.
Discussion: Building on these findings, the article discusses methodological and computational implications for antisemitism detection. It highlights challenges such as semantic ambiguity, pragmatic drift, multimodal signaling, and data scarcity, and evaluates emerging computational approaches, including transformer-based fine-tuning, retrieval-augmented systems, and context-engineered large language models (LLMs). The study concludes that effectively confronting digital antisemitism requires sustained collaboration between linguists, data scientists, and policymakers to develop context-sensitive, transparent, and ethically grounded AI systems capable of reliable interpretive reasoning.

KEYWORDS

antisemitism, hate speech, discourse analysis, digital communication, social media, multimodality, artificial intelligence, large language models

1 Part I—Antisemitism in the Digitally Restructured Public Sphere

The rise of social media has fundamentally reshaped the public sphere. Unlike the era of traditional journalism—dominated by top-down flows of information and professional gatekeepers—the interactive web of today is characterized by horizontal, bottom-up communication. This transformation has enabled unprecedented participation and immediacy,

but it has also allowed harmful ideologies, such as antisemitism, to bypass established filters and spread into the societal mainstream, whether through dynamics of escalation or through gradual normalization.

Social media platforms accelerate these dynamics through a dual logic of perceptual distortion and algorithmic amplification. They first create the impression that one's own views are widely shared—an illusion reinforced by echo chambers and filter bubbles. This perceived consensus lowers thresholds of inhibition, encouraging users to test and transgress the boundaries of socially acceptable speech. At the same time, platform algorithms act as new, opaque top-down filters that privilege polarizing and emotionally charged content, amplifying precisely those utterances that generate outrage and attention. This dual logic—of bottom-up transgression and top-down amplification—creates favorable political-cultural opportunity structures for antisemitism and other forms of hate speech, facilitating their migration from fringe subcultures into mainstream publics and normalizing extremist framings as legitimate opinion (Becker and Rensmann, 2023).

These dynamics are not unique to antisemitism: they also drive the visibility of racist, misogynistic, and conspiratorial communication online (Becker et al., 2023; Fielitz and Thurston, 2019). Yet antisemitism provides a particularly instructive lens because of its discursive versatility—its ability to adapt to new social and technological contexts while retaining recognizable semantic cores (Bergmann and Erb, 1986; Wodak, 2015). Moreover, antisemitism often operates as a connective ideology, linking otherwise distinct forms of hatred and illiberal resentment (Rensmann, 2017; Jikeli et al., 2023). For this reason, the following analysis concentrates on antisemitic discourse as a paradigmatic example of how latent prejudice becomes normalized through digital communication.

Antisemitic discourse manifests in a variety of ways, from overt expressions of hatred to coded or ambiguous insinuations that require cultural knowledge to comprehend. To analyze this spectrum systematically, we distinguish four analytically significant categories of antisemitic discourse.

Our analytical framework follows the working definition of antisemitism proposed by International Holocaust Remembrance Alliance (2016) as its conceptual foundation. However, given the complexity of digital discourse, this definition requires substantial operationalization for empirical and computational research. Building on the *Decoding Antisemitism Lexicon* (Becker et al., 2024), we translate the IHRA's criteria into linguistically testable indicators that account for co-text, context, and multimodal expression. This enables a systematic yet nuanced distinction between explicit, implicit, and ambiguous forms of antisemitic communication.

Building on this operational foundation, we distinguish four analytically significant categories of antisemitic discourse that capture how such meaning materializes in digital communication.

1 *Explicit antisemitism* refers to openly hostile statements that leave little room for alternative interpretation, such as "Jews control the world" or "Hitler was right," as well as direct speech acts like calls for violence against Jews. In digital contexts, this also encompasses celebratory reactions to attacks on Jewish civilians, as observed in comment threads following the events of 7 October 2023 or the Washington shooting in May 2025.

2 *Co-text-dependent antisemitism (micro-level: within-thread reference)* comprises statements whose antisemitic meaning emerges only from their immediate discursive environment or from cross-references within a comment thread. For example, the utterance "They control the world" becomes antisemitic only when the omitted subject ("the Jews") can be inferred from a preceding statement.

3 *Implicit/context-dependent antisemitism* (macro-level: cultural, historical, discursive) includes statements that require broader background or world knowledge to decode their antisemitic meaning. Examples can include:

- Abbreviations such as *"6MWE"* ("6 Million Wasn't Enough"), which reference the Holocaust and connote endorsement of further killings or the complete annihilation of Jewish people (Anti-Defamation League, 2020).
- Utterances in another language, for instance, *"kvetching intensifies"*—a partly Yiddish expression implying that Jews are constantly complaining, weaponized here for mockery without the need to name the Jewish out-group explicitly (Becker, 2025).
- Allusions, such as the word "shower" in "Someone should give George Soros a shower," which, by invoking gas chambers, represents a coded death wish directed at a well-known Jewish individual (Becker and Troschke, 2023).
- Open allusions, for example "The Gaza war reminds us Germans of our past"—implying that Israel is committing a Nazi-like genocide or a war of extermination comparable to that of the Nazis. The supposed openness of the reference to "our past" is undermined by the high salience of Nazi atrocities in the collective memory (Becker, 2021).

4 *Gray-zone cases*—ambiguous utterances in which both antisemitic and non-antisemitic interpretations are possible. In our analysis, such statements are not labeled as antisemitic. Whenever a statement is sufficiently ambiguous to allow a plausible non-antisemitic interpretation alongside an antisemitic one, it is classified as non-antisemitic / negative in accordance with the conservative annotation approach. Beyond lexical meaning, co-text and context are therefore decisive.

- References to "the lobby" or "global elites" may structurally evoke conspiracy tropes of hidden Jewish power. Yet—if clear indicators are absent in the immediate or broader context—they may equally express a more general form of anti-elitism, as frequently observed in anti-capitalist discourse during the COVID-19 pandemic.
- Statements such as "Israel just slaughtered children" can, depending on framing, be read as a modern expression of the traditional antisemitic accusation that Jews habitually murder children. However, despite the harsh wording, statements that situate the accusation within a specific recent event may also be interpreted as expressions of political outrage over a concrete incident.
- Criticism of George Soros can—when accompanied by imagery of the "evil banker"—reproduce classical antisemitic stereotypes. Yet, as in the previous example, the criterion of factual reference may argue against this reading, allowing the statement to be understood instead as sharp criticism of Soros's actual activities as an investor—a pattern not uncommon in anti-capitalist rhetoric.

The proposed fourfold typology makes it possible to map the linguistic and pragmatic spectrum along which antisemitic discourse

unfolds in digital environments. In practice, however, distinguishing between these categories often proves difficult. The examples discussed illustrate the need for precise attention to wording, immediate linguistic context, and relevant world knowledge when assessing the gray zones between antisemitic and non-antisemitic communication. In particular, the boundaries between implicit antisemitism and these gray zones are fluid, as meaning frequently depends on situational framing, intertextual references, and the interpretive expectations of the audience. Overall, this fourfold typology—explicit, co-text-dependent, and implicit antisemitism, along with the problem of gray-zone cases—captures the complexity with which antisemitic discourse circulates in digital publics. The interrelations among these forms—how one may evolve into another or under what conditions specific constellations emerge—remain open questions for ongoing empirical research. Future fieldwork and large-scale LLM-assisted analyses will allow for a more systematic exploration of these dynamic interdependencies.

While conceptually distinct, recent computational work—such as Weinberg et al. (2025)—has nonetheless made important progress in empirically mapping the relationship between explicit and implicit antisemitic content within online communities. Their study of QAnon subreddits demonstrates how explicit antisemitic references provide an interpretive framework through which implicit terms acquire coded meaning for in-group audiences. Yet, despite its methodological sophistication, this approach remains largely lexical and network-based: it identifies co-occurrence patterns between words, but not how antisemitic meaning is pragmatically constructed across sentences or through discourse-level inference. Against this background, the fourfold typology proposed here invites a complementary shift from word-based correlation to context-sensitive interpretation. Implicit antisemitism cannot be fully captured by counting terms such as *Soros*, *globalist*, or *cabal* in proximity to explicit slurs; it emerges through the argumentative relations and presuppositional logic that tie these expressions into coherent narratives. Expressions such as *elites*, *lobby*, or *cabal* can, depending on their discursive environment, function either as elements of antisemitic projection or as components of broader anti-elitist or anti-globalization rhetoric. Similarly, references to *Soros* with negative attributions may, but do not necessarily, carry antisemitic undertones. Future research will therefore need to integrate community-level insights from network analysis with linguistic models capable of representing sentence-level and pragmatic dependencies.

The challenges of implicit or coded antisemitic communication, as well as gray-zone phenomena, are, of course, not confined to digital communication. They were already described by Bergmann and Erb (1986) as "communication latency": antisemitic meaning persists in latent, coded, or camouflaged forms that remain socially intelligible while retaining deniability. Such latent forms often rely on irony, ellipsis, and intertextual cues rather than openly hostile statements, making antisemitic meaning context-dependent and discursively mediated (Becker and Troschke, 2023).

These difficulties point to a more fundamental theoretical and methodological problem: antisemitic communication often operates through indirection, coding, and cultural resonance rather than through explicit hostility. This circumstance makes quantitative assessment particularly difficult, as it cannot rely solely on the surface level of insults, familiar labels, or slogans. Yet it is precisely the combination of context-sensitive qualitative analysis and representative statistical measurement that would offer insights into how frequently antisemitism actually occurs online.

Another key factor is that digital platforms—through their speed, anonymity, and algorithmic amplification—magnify both explicit and coded forms of antisemitism in unprecedented ways. The latter, in particular, are difficult to identify yet highly effective in their cumulative impact. Becker and Rensmann (2023) describe this as a "politics of transgression": implicit expressions that, through their apparent ambiguity, test and expand the boundaries of the sayable, normalizing themselves through repetition and algorithmic reinforcement while maintaining plausible deniability. The normalization of antisemitism, therefore, does not proceed through slurs or overt threats of violence but through the play with ambiguous codes—codes that can be disavowed situationally yet still activate associative chains that gradually establish compatibility with a specific enemy image within the hate ideology of antisemitism.

The aftermath of 7 October 2023 illustrates the transgression of discursive boundaries through the shift from antisemitic projections expressed as stereotypes to open, unfiltered glorifications of violence against Jews—appearing for the first time with such intensity in mainstream comment spaces. By contrast, reactions to the attack on the museum in Washington, D.C., in May 2025 more often relied on deflection, denial, and mockery—forms of modern antisemitism that trivialize Jewish victimhood.

Antisemitism today functions as a dynamic reservoir of stereotypes and resentments—endlessly adaptable, internally contradictory, and responsive to shifting cultural codes and situational triggers. This malleability explains both its resilience and its analytical difficulty—for civil society, researchers, and AI systems alike. Recent computational research has begun to address this challenge from the perspective of large language models (Becker et al., 2023; Halevy et al., 2024; Jikeli et al., 2023; Halevy et al., 2024; Steffen et al., 2024). Patel et al. (2025) systematically evaluated state-of-the-art LLMs on antisemitism detection and confirmed what discourse analysis has long emphasized: the decisive fault lines lie not in explicit hate but in implicit patterns—and in the gray zones where political critique and antisemitic projection overlap.

Building on these theoretical foundations, the following case studies examine how the described dynamics manifest in real-world digital discourse. Both events—the Hamas-led attacks of 7 October 2023 and the Washington museum shooting of May 2025—serve as empirical test cases for the typology outlined above. They illustrate how explicit, co-text-dependent, and implicit forms of antisemitism evolve within moments of political crisis, and how the gray zones between criticism, denial, and hostility become discursively negotiated in comment spaces. By moving from theoretical reflection to empirical analysis, these studies demonstrate how antisemitic meaning is not a static category but a shifting, context-sensitive practice shaped by digital affordances, emotional contagion, and algorithmic amplification.

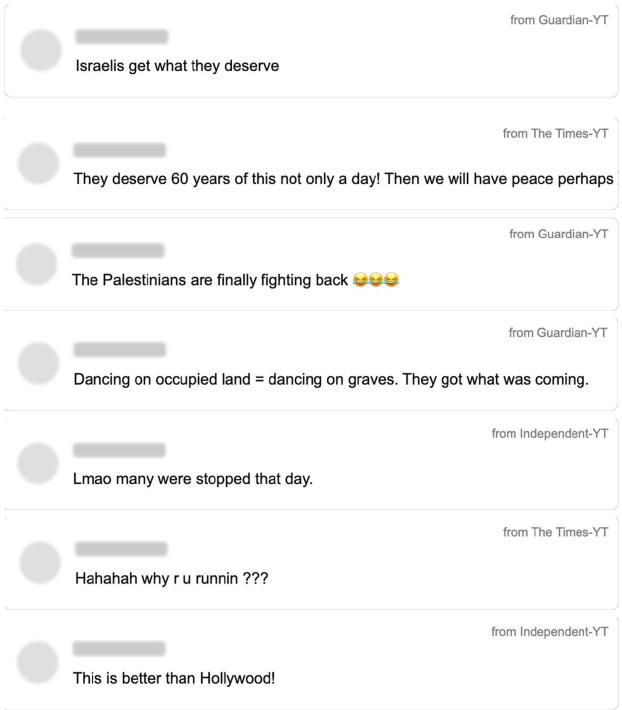
2 Case study I: The Digital Aftermath of October 7, 2023

The Hamas-led terror attacks of 7 October 2023 marked a rupture not only in the Israeli–Palestinian conflict but also in the

online discursive environment. Within hours, antisemitic rhetoric surged across mainstream platforms—particularly in the *Decoding Antisemitism* datasets, most visibly in the YouTube comment sections of major English-language outlets such as the *BBC*, *The Guardian*, *The Independent*, and *The Times*. The corpus of our first case study (Becker et al., 2023), comprising more than 11,000 user comments collected between 7 and 11 October 2023 and coded with MAXQDA, enables us to trace how antisemitic discourse shifted in real time across mainstream media comment sections.

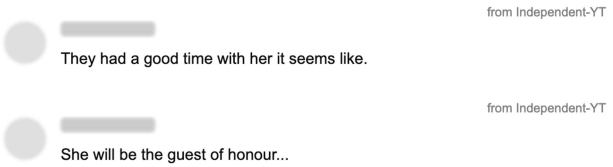
The analysis reveals a clear discursive rupture: antisemitism that had long circulated in coded or implicit registers now appeared openly and euphorically. Its expressions were less projective—that is, less focused on classic stereotypes or demonizing analogies such as Nazi comparisons, and thus involved fewer allegations directed at the Jewish out-group—than in previous escalation phases of the Middle East conflict. Instead, they predominantly centered on the commenters' own positioning, taking the form of celebratory statements that rejoiced in acts of violence without disguise.

This moment illustrates the categories of explicit and cotext-dependent antisemitism. Unlike in earlier escalation phases, many users abandoned rhetorical camouflage altogether. Comment sections featured unambiguous glorifications and justifications of violence, death wishes, and expressions of schadenfreude directed at Israeli–Jewish civilians:

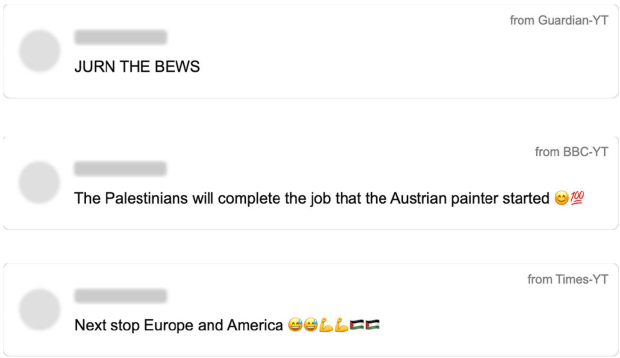


These examples reveal how antisemitic expression shifted from latent and coded forms to an open celebration of violence, marking a significant erosion of discursive boundaries. Such unfiltered expressions exemplify what we term a politics of transgression: by publicly and explicitly indulging in sadistic enjoyment, users actively pushed and redefined the boundaries of the speakable within mainstream spaces.

A further pattern was the convergence of antisemitism with misogyny. In comments on the Nova music festival massacre, female victims were mocked or sexualized:



Here, antisemitic cruelty merged with gendered humiliation, underscoring the intersectionality of online hate. At the same time, implicit antisemitism remained present:



These examples demonstrate how antisemitic meaning can persist in coded or allusive form even when explicit slurs are avoided. The comment "JURN THE BEWS" employs deliberate orthographic distortion to evade moderation while transparently echoing "burn the Jews," thus functioning as a barely disguised call for annihilation. The reference to "the Austrian painter" alludes to Hitler and serves as a coded endorsement of Nazi Germany's genocidal project. Finally, "Next stop Europe and America" reflects a conspiratorial worldview in which Islamist violence against Israel is reframed as the opening act of a broader struggle against the West. Within this discursive logic, Israel becomes the imagined spearhead of a corrupt, "Jewish-dominated" Western order—an enduring antisemitic trope that merges anti-Israel hostility with civilizational resentment.

Multimodal elements reinforced these dynamics. Even in text-dominant YouTube environments, emojis and symbols functioned as shorthand for alignment and celebration: paragliders (🪂), Palestinian flags (PS), and watermelons (🍉) served as proxies for solidarity with Hamas or approval of the massacre. The watermelon emoji, in particular, performs a double function: it not only recontextualizes the colors of the Palestinian flag but also implies that expressions of solidarity with Palestinians are subject to censorship and must therefore adopt coded forms—an insinuation that aligns with conspiratorial notions of Jewish control over the public sphere. These markers evaded keyword detection systems while remaining readily legible to in-group audiences.

Taken together, the October 7 corpus shows how antisemitism functions as a dynamic ideological reservoir: old motifs such as blood libel, Holocaust inversion, and Jewish world control reappeared, but alongside them, a striking normalization of blunt, celebratory hatred emerged. What distinguishes this case from earlier cycles is not only the volume of antisemitic discourse but its abandonment of

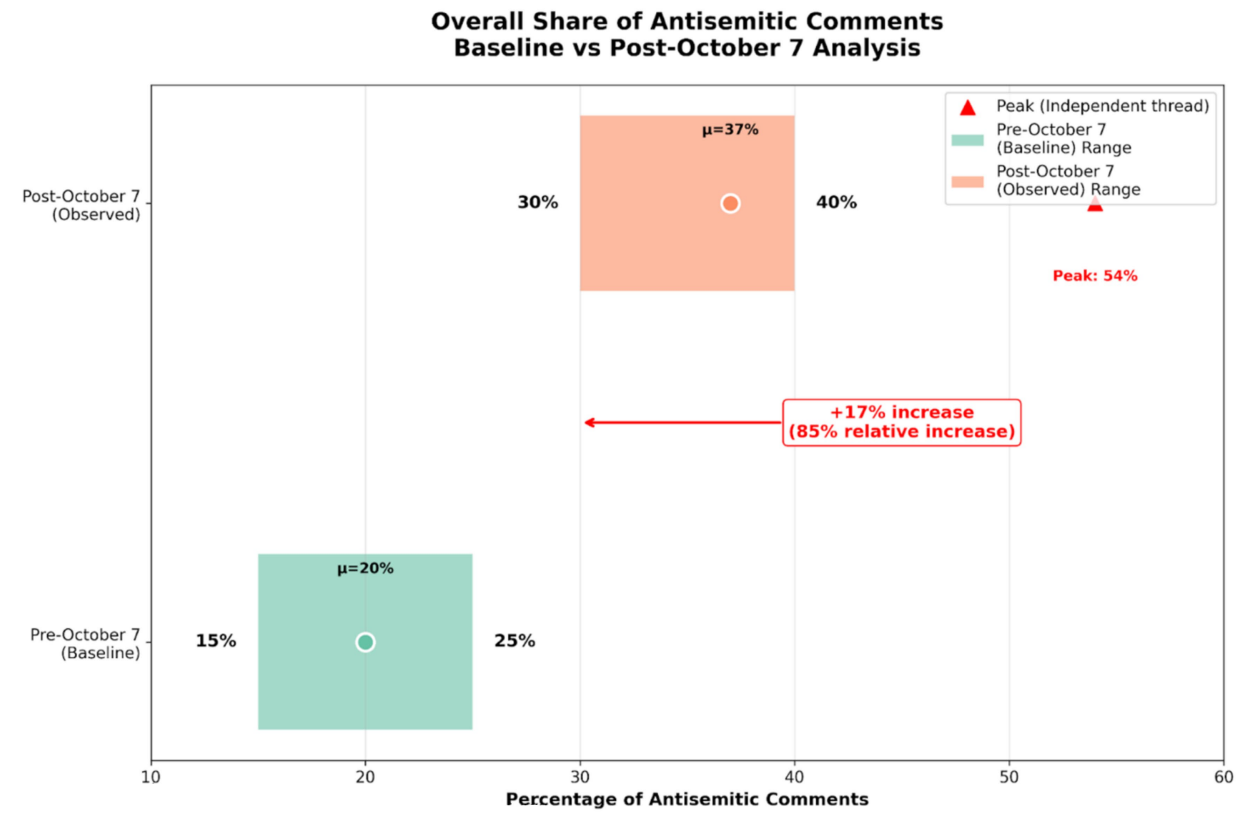


FIGURE 1
Overall share of antisemitic comments—UK YouTube (Oct 7–12, 2023). Measurement period: 7–12 October 2023. Focus: comments on reports covering the Hamas attacks and immediate aftermath. The typical pre–Oct 7 baseline was 15–25%; observed levels rose to 30–40% (dataset mean ≈36–38%; peak thread 54%). This implies a relative risk of about 1.85 × versus baseline (range ≈ 1.2–2.7×).

justification. Hatred was no longer masked as critique; it was asserted as spectacle (Figures 1, 2).

3 Case study II: Antisemitic Reactions to the Washington double Murder, May 2025

On 21 May 2025, a 30-year-old assailant entered the Capital Jewish Museum in Washington, D.C., shot two Jewish embassy staff members—Yaron Lischinsky and Sarah Milgrim—and shouted "Free, free Palestine." The attack was quickly identified as an antisemitic hate crime.

In the YouTube comment sections of eight major English-language news channels analyzed by Decoding Antisemitism—including the BBC, *The Guardian*, *Sky News*, *LiveNow from Fox*, *CTV News*, *Forbes*, and *Al Jazeera English*—the reactions took a markedly different form than those following 7 October 2023. Our dataset of 1,600 comments indicates that antisemitism here manifested not through open celebration of violence, but primarily through conspiracy narratives, the instrumentalization of antisemitism, and the moral justification of the attack. While 7 October represented a peak of explicit transgression, the Washington 2025 case illustrates how antisemitism increasingly shifts into gray zones and implicit registers.

This growing implicitness may be particularly pronounced in jurisdictions with stricter hate speech enforcement, where coded language and semantic ambiguity tend to replace more explicit forms of antisemitic expression. This shift may also relate to the fact that—unlike in the first case study—Jewish victimhood was here situated at a greater spatial and thematic distance from the Middle East conflict, making open celebration of the murder of a young couple less socially acceptable. In this case, the recognition of the crime was accompanied not by affirmation but by justification, relativization, and conspiratorial reasoning. Antisemitism was thus partially reframed semantically—not as an expression of hatred, but as a supposedly political or morally comprehensible reaction.

While open celebration of violence was rare, antisemitism often appeared through rhetorical displacement—by denying antisemitism as a relevant category of harm, reframing the attack as political blowback, or implying that Jewish victimhood was self-inflicted (Figure 3).

(1) *Conspiracy myths*

A first strategy consisted of the reproduction of classic conspiracy narratives, which alluded—either explicitly or implicitly—to Jewish or Israeli culpability. Some comments directly accused the Israeli intelligence service Mossad, while others used insinuation and rhetorical questioning to suggest that the attack was an "inside job."

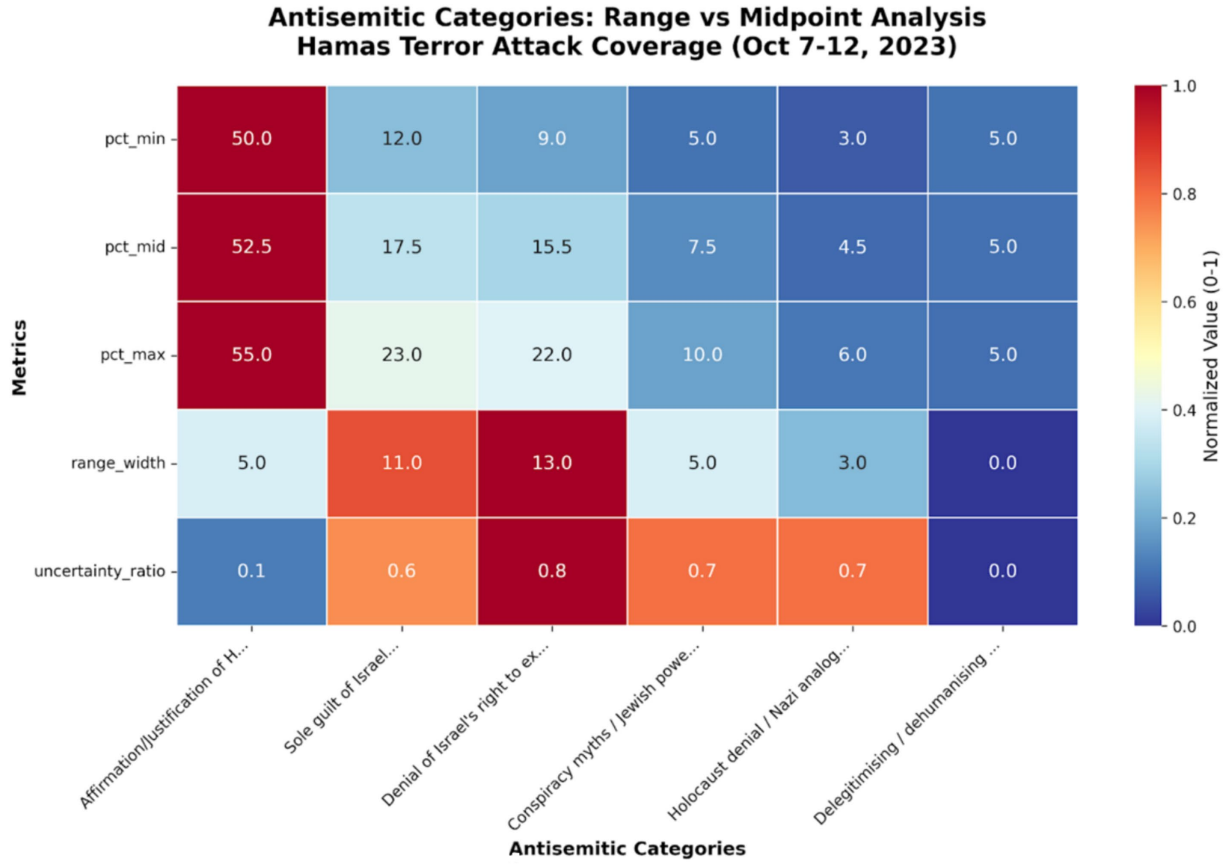
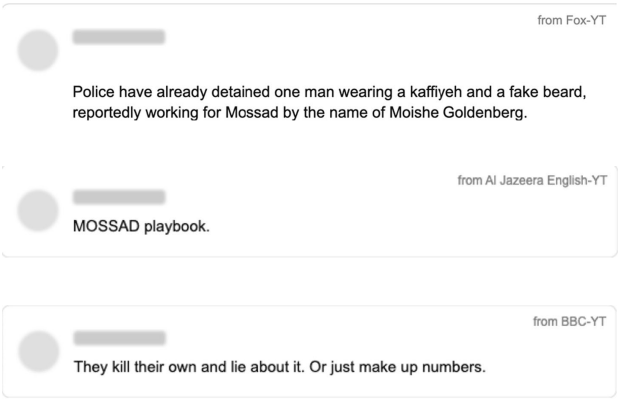
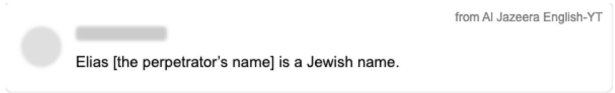


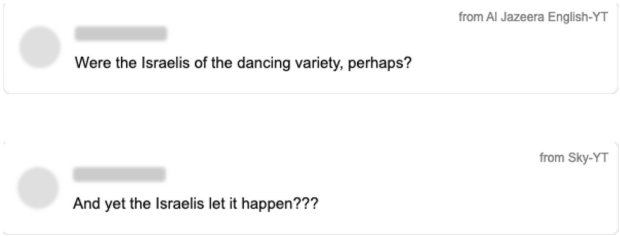
FIGURE 2
Categories differences heatmap. Normalized (0–1) heatmap comparing six metrics across antisemitic categories in UK YouTube coverage of the Oct 7–12, 2023 events. Darker cells indicate higher values. Rows show pct_min/mid/max, range_width (pct_max–pct_min), and uncertainty_ratio (range_width ÷ pct_mid). "Affirmation/Justification" is the largest category (mid ≈ 52.5%) with low relative uncertainty (~0.10). "Delegitimising tropes" shows perfect agreement (range 0). "Denial of Israel's right to exist" has the greatest uncertainty (range 13 pts; ratio ≈0.8). Source: UK YouTube, Oct 7–12, 2023. Values: Min/Mid/Max percentages, range width uncertainty ratio.

Explicit:



Implicit:





These statements shift antisemitic blame into a system of insinuation. The comment "Elias is a Jewish name" implies a reversal of perpetrator and victim roles by suggesting the Jewish identity of the attacker and thereby casting doubt on the antisemitic motivation of the crime. The subsequent rhetorical questions evoke the old myth of the "dancing Israelis" after the attacks of 11 September 2001, insinuating that Israel had foreknowledge of—or even participated in—the event. In this way, Jewish culpability is not asserted directly but insinuated through irony and cultural allusion—a typical example of latent antisemitic communication.

(2) *Instrumentalization of antisemitism*

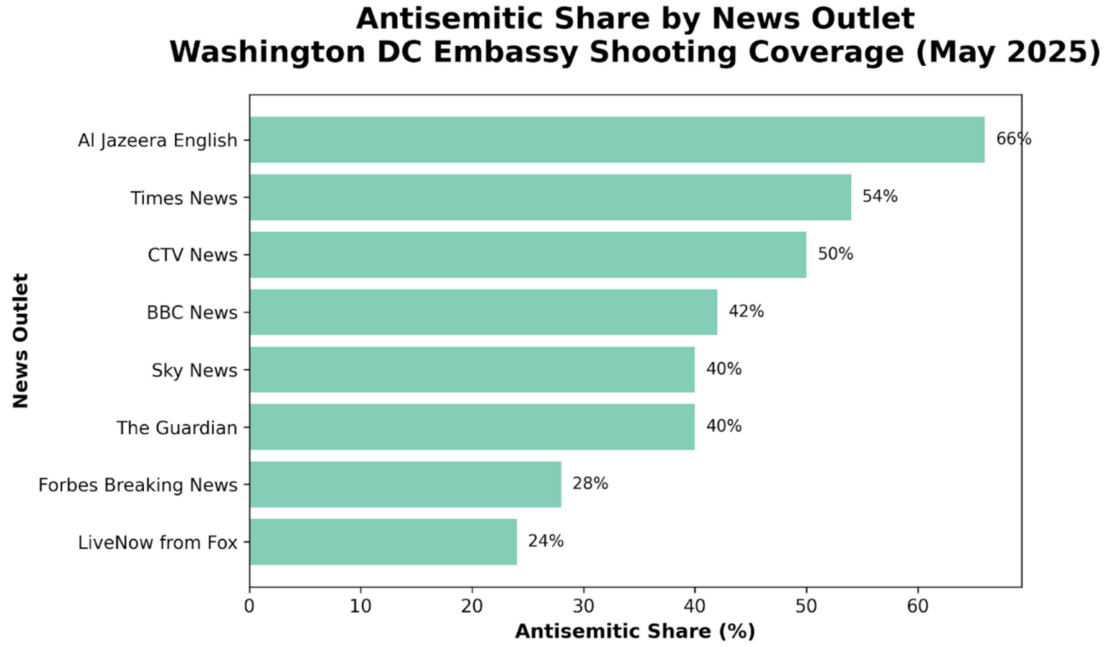
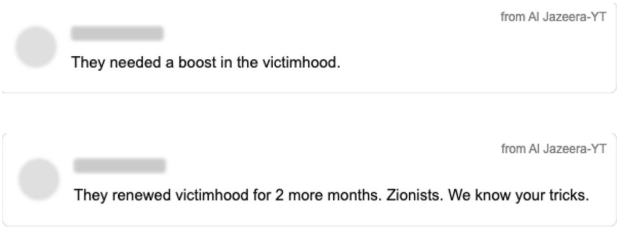


FIGURE 3
Antisemitic share by News Outlet—Washington DC Embassy Shooting (May 21–22, 2025). Horizontal bar chart ranking eight YouTube news channels by the percentage of antisemitic comments. Al Jazeera English shows the highest share (66%), followed by Times News (54%) and CTV News (50%); BBC News (42%), Sky News (40%), and The Guardian (40%) form a mid-tier, while Forbes Breaking News (28%) and LiveNow from Fox (24%) are lowest. Overall, the weighted average across outlets is ≈43%, indicating substantial platform-level exposure during the event coverage. Source: DA media datasets (YouTube, May 21–22, 2025).

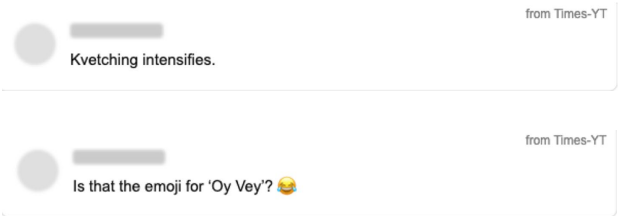
A second, increasingly visible strategy was the claim that Jews instrumentalize antisemitism for political purposes. This form does not deny that Jews are attacked but rather trivializes or mocks their victimization, portraying Jewish suffering as exaggerated, calculated, or manipulative.

Explicit or co-text-dependent examples:



Such comments draw on the antisemitic trope of a "cult of Jewish victimhood": Jewish remembrance and mourning are framed as strategic or excessive and thereby stripped of their moral legitimacy.

Implicit:



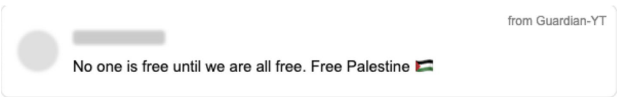
This logic closely aligns with Adorno's notion of secondary antisemitism—resentment not only toward Jews but toward the memory of their suffering itself. Here, the use of Yiddish expressions (*kvetch*, *oy vey*) serves as a cultural marker employed with ironic intent. These comments rely on shared background knowledge and function as coded signals of antisemitic Schadenfreude: Jewish grief is caricatured, its authenticity questioned.

Antisemitism thus operates here not through open hostility but through cynical irony—a form of digital denigration that punishes empathy and rewards mockery.

(3) *Affirmation and justification of violence*

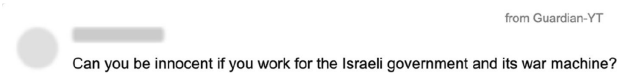
A third strategy consisted of the moral justification or indirect endorsement of the attack. It appeared in varying degrees—from explicit approval to co-text-dependent allusions and implicit symbolic references.

Explicit:



At first glance, this formulation may appear humanistic or solidaristic. Yet within the co-textual framework—namely, a targeted double murder at a Jewish institution—it functions as an affirmative signal, rhetorically reframing the crime as part of a "liberation struggle."

Implicit:



In this rhetorical question, no explicit approval is expressed, yet it implicitly suggests that the victims were not civilians but part of a system, and therefore shared culpability. Responsibility is collectivized, and Jewish life is morally relativized.

Multimodal reference:

The inverted red triangle (▼) here functions as a visual symbol used to mark allegedly "guilty" individuals. Its antisemitic connotation becomes clear only through background and (digital-)contextual knowledge. Despite lacking verbal aggression, it represents an implicit justification and personalization of guilt.

Multimodal elements appeared less frequently than in the 7 October dataset but served a more condensed semantic function: they signaled alignment and belonging without expressing open approval.

Across these strategies—ranging from conspiratorial insinuation to moral inversion and the instrumentalization of Jewish suffering—the underlying structure remains the same: violence is not denied but morally reframed—as understandable, justified, or self-inflicted. The discourse thus shifts from the recognition of an antisemitic crime toward the legitimation of antisemitic violence.

Taken together, the comparison between October 7 and Washington 2025 highlights the dynamic repertoire of antisemitic discourse in digital publics. In the first case, antisemitism appeared in overt and explicit forms: blunt celebrations of murder, open death wishes, and unfiltered demonization of Jews. In the second, it shifted into more implicit and gray-zone registers, where hostility was expressed through denial, whataboutism, and sarcastic trivialization of Jewish victimhood. This variation illustrates the full spectrum of antisemitic communication online—from overt hatred to camouflaged codes—and shows how antisemitism adapts discursively when acts of violence are less easily celebrated. Instead of open glorification, it operates by undermining Jewish victimhood—through conspiracy narratives, mockery, relativization, and moral reframing of violence. All these strategies share a rhetorical reversal of perpetrator and victim roles, through which Jewish suffering is relativized and antisemitic violence is portrayed as politically or morally defensible. The shift from self-affirming expressions of hate after October 7 to cynicism, sarcasm, and derision in the Washington case underlines why simple keyword-based approaches are insufficient: detecting antisemitism requires attention to context, world knowledge, and discursive patterns. It is precisely in these gray areas—where interpretation is contested—that theory-guided annotation, interdisciplinary analysis, and computational tools become indispensable (Figure 4).

4 Challenges for AI and Computational Modeling

The preceding case studies underline that antisemitism online is not a stable object easily captured through keyword lists or statistical sentiment patterns. It is instead a dynamic and adaptive discourse that thrives on ambiguity, irony, and projection. For computational

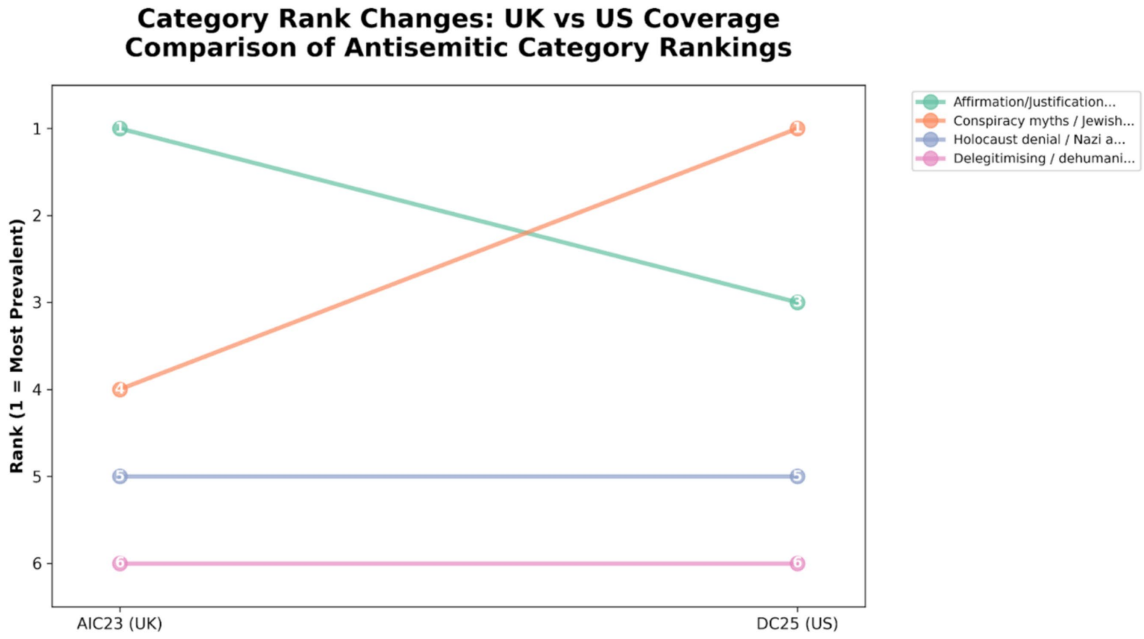


FIGURE 4
Category rank changes (UK AIC23 → US DC25). Affirmation/Justification declines (1 → 3) while Conspiracy myths rise (4 → 1), indicating a shift from overt celebration to implicit, coded reframing; Holocaust denial and the Delegitimization of Israel remain least prevalent. Source: DA media datasets. AC/23: UK YouTube, Oct 7–12, 2023| DC25: US YouTube, May 21–22, 2025.

approaches, this poses a distinctive challenge: antisemitism represents both an ancient repertoire of stereotypes and a continually mutating communicative practice. Unlike other hate categories that rely more heavily on slurs or direct epithets (e.g., racial pejoratives), antisemitism frequently disguises itself in the language of political critique, satire, or moral outrage. This elasticity compels annotation teams and AI developers to confront not only technical hurdles but also foundational definitional debates.

Compounding this difficulty, antisemitic language evolves on two levels. On the one hand, it deliberately adapts to evade moderation through coded expressions and dog whistles; on the other, it shifts subconsciously at a societal level as words acquire new connotations over time. Two mechanisms illustrate this problem. First, pragmatic re-evaluation: terms such as "Zionist" and "anti-Zionist" have not changed denotatively, yet their connotations and usage contexts have diversified. Depending on the framing, the statement "I am anti-Zionist" can denote policy-based opposition to specific Israeli actions or, alternatively, a categorical rejection of Jewish statehood—an interpretation aligning with antisemitic frameworks such as those articulated in the IHRA definition. Second, co-textual reinterpretation: phrases like "Free Palestine" retain a stable core meaning—support for Palestinian self-determination—but in certain placements (for example, posted under an article about an antisemitic hate crime) they can function as coded hostility toward Jews or Zionism.

Both dynamics blur the boundary between political speech and antisemitic projection, placing distinct demands on classification. Because meanings evolve and contexts shift, even advanced models struggle to determine whether a user's language is antisemitic, ironic, or politically motivated. A further complication arises from the structure of online discourse itself: explicitly antisemitic comments may be harder for AI to classify than original posts relying on dog whistles or cultural knowledge. Without access to the surrounding conversational thread, the meaning of a single comment can remain opaque—causing models to either over-detect benign speech or under-detect antisemitic insinuation.

4.1 Annotation as a knowledge-intensive task

Annotation is the foundation of supervised machine learning. For antisemitism detection, however, annotation cannot be reduced to a mechanical task of tagging keywords. It requires contextual knowledge of Jewish history, antisemitic tropes, and the shifting semantics of digital culture.

All annotation in our project is guided by the IHRA working definition of antisemitism, operationalized through the Decoding Antisemitism Lexicon (Becker and Fillies, 2024) to ensure consistent application across linguistic and computational analyses.

As Steffen et al. (2024) demonstrate, widely used moderation services such as Perspective API fail to capture subtle antisemitism: three quarters of antisemitic comments in their dataset scored below toxicity thresholds, while counter-speech was frequently misclassified due to keyword bias ("Jew," "Israel"). This underlines why annotation for antisemitism must go beyond keyword spotting and requires deep contextual knowledge. It also demands attentiveness to how the meanings of key terms evolve in public discourse, creating persistent disagreement over whether a statement constitutes legitimate political critique, carries antisemitic references or inferential cues, or oscillates between both interpretive frames.

In the pilot phase of *Decoding Antisemitism*, our team annotated more than 100,000 social media comments drawn from mainstream media platforms in English, German, and French. The annotation framework was structured in three layers of granularity:

1. *Binary classification*: antisemitic vs. non-antisemitic.

This step provided a baseline for training classifiers but proved insufficient on its own. Binary labels tend to collapse the diversity of antisemitic expression and are prone to false negatives (missing implicit forms) and false positives (misclassifying satire, counter-speech, or mere references to Jewishness).

2. *Stereotype categories*: mapping each antisemitic utterance onto a set of historically documented stereotypes (e.g., greed, power, deceit, child murder, blood libel, disloyalty, media control).

This approach helped capture not just the presence of antisemitism but its discursive function. For example, claims that Israel deliberately kills children are not only a political accusation but also echo the long-standing "blood libel" motif.

3. *Communicative forms*: identifying the rhetorical strategy through which antisemitism is articulated—analogy, insinuation, irony, rhetorical question, hyperbole, meme, or emoji.

This dimension proved crucial for computational modeling, as it highlights patterns that transcend specific words. An ironic phrase like *"Schindler's List? More like Schwindler's List"* requires world knowledge and an understanding of irony to be detected.

This layered approach provided the conceptual basis for fine-tuning models such as BERT. But it also revealed the limits of annotation itself.

4.2 The problem of gray-zone cases

A central difficulty in annotation is the presence of gray-zone utterances: statements that can plausibly be interpreted as either antisemitic or non-antisemitic depending on context. Consider three examples:

- "The lobby controls the world."
  If understood as generic anti-elitism, this could be non-antisemitic. If "the lobby" is a coded reference to AIPAC or Jewish influence, it clearly maps onto the conspiracy trope of Jewish control.

- "Israel just slaughtered children."
  This could be a reaction to specific war reporting, but when generalized or repeated as a narrative of Israeli essence, it draws on centuries-old blood libel motifs.

- "Soros is an evil banker."
  This might be harsh criticism of an individual financier, but combined with references to global elites, it activates stereotypes of Jewish manipulation.

- "I am anti-Zionist."
  Interpretation hinges on societal usage. If it denotes opposition to specific Israeli government policies, it may fall outside

antisemitic criteria. If it denotes a categorical rejection of Jewish statehood, it aligns with longstanding antisemitic frameworks.

- "Free Palestine."
  The core meaning is support for Palestinian self-determination. Yet cotext matters: under a post about an antisemitic hate crime, it reads as coded hostility toward Jews or Zionism; under news about Palestinian affairs, it may not.

Such cases highlight why annotation requires expertise and intercoder discussion. In our project, intercoder reliability often dropped precisely in these gray zones. Conservative coding schemes marked these utterances as "ambiguous," while more expansive schemes argued for classification based on context and world knowledge. This reflects Becker and Troschke (2023) finding that implicit antisemitism dominates online discourse and that coders must adopt a conservative interpretive approach: minimizing false positives while carefully documenting potential antisemitic readings.

This tension mirrors broader societal debates: what counts as legitimate criticism of Israel, and what constitutes antisemitism? Computational models trained on ambiguous labels will reproduce these uncertainties. Without careful definition and annotation, they risk either under-detection (missing antisemitic tropes) or over-detection (misclassifying political critique).

4.3 The evolution of annotation workflows

Annotation also had to evolve methodologically. At first, annotation was conducted manually in MAXQDA with comment-by-comment coding. This proved too slow for larger datasets. We subsequently adopted hybrid workflows:

- *Rule-based pre-filtering*, using keyword lists to surface potentially antisemitic comments for human review.
- *Iterative annotation*, where machine learning models suggested likely labels and human experts confirmed or corrected them.
- *Feedback loops*, where recurring disagreements were discussed in coder workshops, and annotation guidelines were updated.

Over time, these workflows revealed a fundamental point: annotation is not just a technical precondition for AI but itself a research activity. It produces knowledge about the ambiguity, frequency, and contextual dependencies of antisemitic discourse. Annotation guidelines became living documents, integrating historical scholarship, discourse analysis, and empirical coder experience.

4.4 Model design: from BERT fine-tuning to LLM prompting

The computational side of *Decoding Antisemitism* began with transformer-based models such as BERT and RoBERTa. These models offered a step change compared to older dictionary-based detection tools, which typically relied on keyword spotting. By capturing contextual embeddings, transformers allowed us to move beyond surface vocabulary and approximate the pragmatic layer of antisemitic discourse (Figure 5).

4.4.1 Fine-tuning BERT

In the pilot phase, we fine-tuned BERT models on our annotated datasets. The process followed standard supervised-learning steps:

1. *Preprocessing*: Comments were cleaned, tokenized, and aligned with annotation layers.
2. *Training splits*: Annotated corpora were divided into training, validation, and test sets, with careful stratification to avoid over-representing explicit antisemitism at the expense of implicit or gray-zone cases.
3. *Label design*: Models were trained at different levels of granularity—binary, stereotype categories, and rhetorical strategies.

Results were promising but uneven:

- *High accuracy* for explicit antisemitism, particularly when slurs or direct Holocaust references were present.
- *Moderate success* for implicit forms, such as analogies (*"Zionists are the new Nazis"*) or rhetorical questions (*"Why do they always control the media?"*).
- *Low accuracy* in gray-zone areas, where annotation uncertainty was reflected in model misclassifications.

Two lessons stood out:

- Transformer models perform best where human coders already agree. They replicate annotation quality rather than improve it.
- The weakest areas were precisely those most socially consequential: implicit, coded, and context-dependent forms of antisemitism.

Comparable experiments by Pustet and Mihaljević (2024) reached F1 $\approx$ 0.69 for antisemitic texts and 0.96 for non-antisemitic ones, underscoring the difficulty of capturing coded and implicit antisemitism even with fine-tuned BERT models.

4.4.2 The move toward LLMs

As *Decoding Antisemitism* progressed into its later phases (2023–2025), the research landscape shifted with the rise of Large Language Models (LLMs) such as GPT-3.5 and GPT-4. Unlike BERT, which is optimized for classification tasks, LLMs excel at generative reasoning and zero−/few-shot learning.

For antisemitism detection, this meant three new opportunities:

1. *Prompt engineering*: Instead of training classifiers from scratch, we could design prompts instructing LLMs to apply annotation guidelines. For example:
   "Classify the following comment as antisemitic, non-antisemitic, or ambiguous. Use the IHRA definition as reference and explain your reasoning."
   This allowed models to handle new, unseen comments without extensive retraining.

2. *Chain-of-thought prompting*: Asking the model to explain why a comment may or may not be antisemitic increased transparency and interpretability, which was crucial for trust in moderation workflows.
3. *Multilingual flexibility*: LLMs demonstrated surprising competence in cross-lingual transfer, enabling preliminary

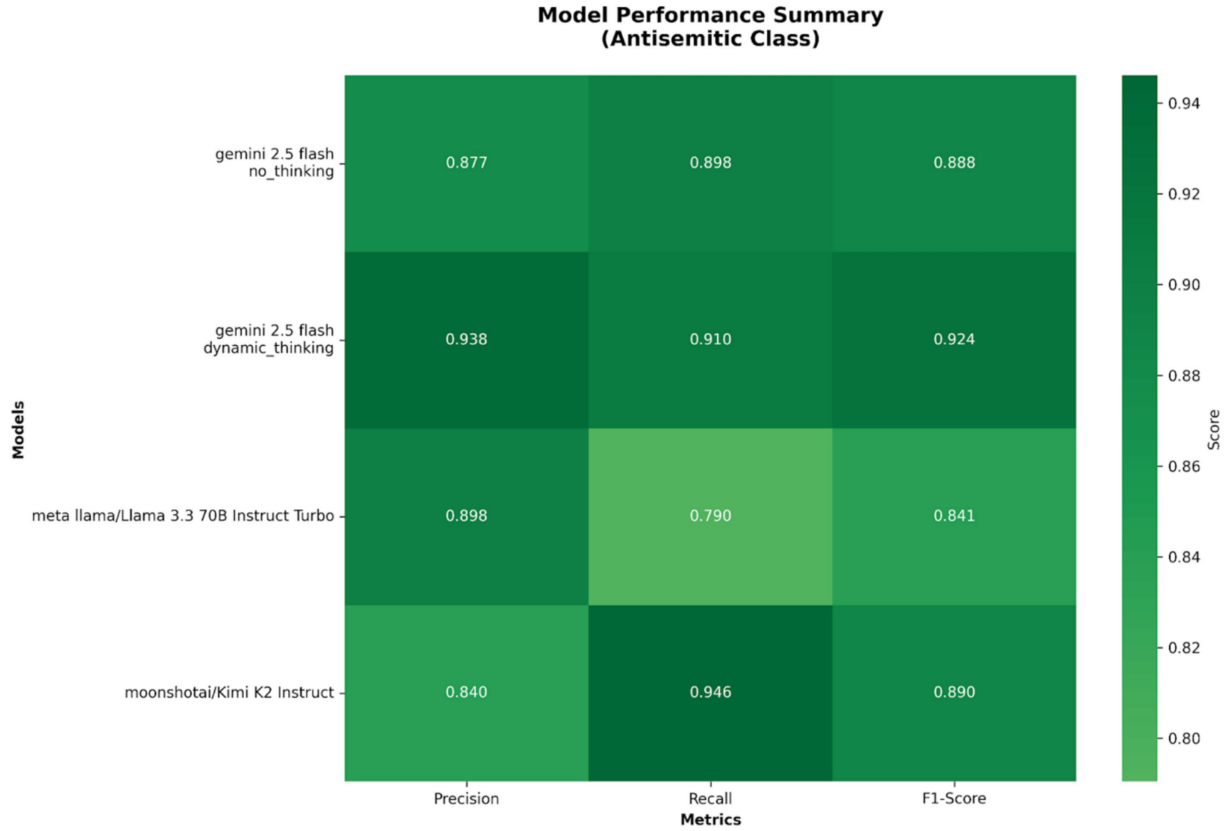


FIGURE 5
Model performance on antisemitic Class (n = 445 posts). Heatmap of precision, recall, and F1 for four models. Gemini-2.5-flash (dynamic_thinking) achieves the best balance (F1 ≈ 0.92; Prec 0.94; Rec 0.91). moonshot/Kimi K2 is recall-oriented (Rec ≈ 0.95) with lower precision (0.84), while llama-3.3-70B is precision-leaning (Prec≈0.90; Rec ≈ 0.79). These profiles indicate a trade-off between conservative (high-precision) and liberal (high-recall) detection strategies.

classification in German, French, and Spanish without new annotation datasets.

Still, challenges persisted:

- *Hallucination and overconfidence*: LLMs sometimes invented antisemitic meaning where none was present, especially in ambiguous utterances.
- *Bias replication*: Pre-trained LLMs mirrored dominant discourses in their training corpora, sometimes normalizing anti-Israel rhetoric while downplaying antisemitic undertones.
- *Scalability limits*: API-based LLMs required careful cost–benefit balancing when applied to millions of comments.

These observations align with Patel et al. (2025), who show that even advanced models such as GPT-4 and Claude struggle with coded antisemitism and context-dependent gray-zone cases. Their results reinforce our own findings: while LLMs offer flexibility and surprising multilingual competence, this comes at the cost of inconsistency and overconfidence, particularly when antisemitic tropes are embedded in political or moralizing rhetoric.

No peer-reviewed research has yet systematically evaluated newer LLMs on antisemitism detection. Preliminary benchmark experiments conducted at the Blue Square Alliance (BSA) Command Center suggest that, when guided by carefully structured prompts emphasizing both annotation principles and linguistically ambiguous discourse—particularly regarding Israel and Zionism—several recent models display substantial performance gains (Blatter, 2025). While exploratory and limited in scope, these findings indicate that with nuanced contextual framing, the latest LLMs show genuine potential for reliable large-scale antisemitism classification. On a benchmark dataset of approximately 450 cotext-independent posts (that is, posts that are not replies or comments on other posts, and that do not reference real-world events occurring after model training or obscure incidents outside general world knowledge)––comprising roughly 37% antisemitic and 63% non-antisemitic examples, some of which required world knowledge such as familiarity with antisemitic tropes or conspiracy references––Gemini-1.5-Pro, Gemini-2.5-Flash (in both standard and "dynamic thinking" configurations), Llama-3.3-70B-Instruct-Turbo, and MoonshotAI Kimi-K2-Intruct achieved weighted F1 scores above 0.88. These preliminary results suggest that, with careful prompt design and nuanced contextual framing, the latest generation of models demonstrates genuine potential for reliable large-scale classification of antisemitism.

In these experiments, all LLMs were guided by a single, shared task prompt. The prompt defined antisemitism as "hostility, prejudice, or discrimination toward Jews as Jews," explicitly including conspiracy theories, Holocaust denial, dehumanizing portrayals, and attacks on Jewish identity, institutions, or the legitimacy of the State of Israel. It

then enumerated recurring themes where misclassification is common, e.g., accusations of genocide, apartheid or settler-colonialism framings, praise for Hamas and "resistance," Holocaust and Nazi comparisons, delegitimization of Israel's existence, and references to Soros or Rothschilds, and supplied concrete examples for each. The prompt encoded a conservative ambiguity rule, instructing models to classify as *Not Antisemitic* whenever a post plausibly supported both antisemitic and non-antisemitic readings.

This design had two methodological advantages. First, it aligned the LLMs with BSA's existing gray-zone adjudication guidelines, reducing the frequency of over-attribution. Second, it helped models avoid naïve keyword-matching and instead reason about distinctions central to antisemitism studies––for example, the difference between categorical claims such as "Zionism is a colonialist project" versus time-bound criticism of military policy.

Despite its strengths, the prompt revealed several consistent gaps across models:

1 Missing harassment-style antisemitism
The prompt emphasizes ideological tropes but gives fewer explicit examples of simple harassment or slur-based antisemitism. As a result, several models misclassified short, demeaning expressions.

Examples include:

- *"You read like a trans-Jew."*
  - Gold label: *Antisemitic*
  - Misclassified by three of five top models

- *"Jews and Jeets 🪰"*

Gold label: *Antisemitic*

- Correctly flagged by only some models.

The models' failures here reflect a prompt design gap: while ideological antisemitism was well-covered, the prompt under-specified how to treat insults, slurs, and denigrating constructions.

2 Older or coded conspiracy tropes not explicitly listed

While the prompt covers "stereotypes of Jewish control over media/finance/politics," it does not mention well-known conspiracy codes such as:

- *"dancing Israelis"*
- The Great Replacement Theory

This omission contributed to systematic errors. For example:

- *"dancing israelis b&w pics at the end? Neat"*
  - o Gold label: Antisemitic
  - o Only Kimi-K2 classified it correctly; all other top models missed it.
- *"Please stop the white holocaust. I believe 6 million whites have been murdered by black and brown people."*
  - o Gold label: Antisemitic
  - o Missed by both Gemini-2.5 variants; caught by Llama-3.3-70B and Kimi-K2.

These cases show that without explicit prompt examples, several LLMs failed to connect certain phrases to historical antisemitic narratives.

3 Uneven application of the ambiguity rule
Although all models received the same instruction, compliance varied.

- *"Hi Jew, I'm Dad"*
  - o Gold label: *Not Antisemitic*
  - o Only Gemini-2.5-Flash (dynamic_thinking) correctly applied the benefit-of-the-doubt rule.

Conversely:

- A counterspeech post beginning with:
  "Anti-Zionist: 'I do not hate Jews. I just hate Zionists/Israel.' Me: 'Why?' Anti-Zionist … proceeds to say the most rabid Nazi shit …"
  - o Gold label: *Not Antisemitic*
  - o Only Gemini-2.5-Flash (dynamic_thinking) correctly understood the stance.

This demonstrates that the ambiguity clause is effective, but only when a model can reliably infer stance—a capability that varied substantially between models.

4 Difficulty identifying rhetoric that broadly associates Jewish/Israeli entities with violence
Some models showed inconsistent performance on examples where the rhetoric attributes violence or murderous intent to Jewish or Israeli groups as a whole, rather than criticizing specific actions or policies.

This explains treatment of cases such as:

- *"Pope Francis Mourned By Gaza's HYPOCRITICAL Murderers via @YouTube"*
  - o Gold label: *Antisemitic*
  - o Missed by some models.

- *"yes, aipac are the terrorists funneling millions…"*
  - o Gold label: *Antisemitic*
  - o Misclassified by Llama-3.3-70B.

4.4.3 Hybrid solutions: retrieval-augmented generation and context-engineering

To address these limitations, we began experimenting with RAG architectures. Here, the LLM is coupled with an external knowledge base—in our case, annotated corpora, the *Decoding Antisemitism Lexicon*, and stereotype taxonomies.

- When confronted with a comment like *"Soros pulls the strings in Brussels,"* the model retrieves entries on conspiracy tropes, Jewish financial stereotypes, and gray-zone indicators.
- The LLM then integrates this evidence into its reasoning, reducing hallucination and anchoring classifications in scholarly knowledge.

This hybrid architecture combines the scalability of LLMs with the domain expertise of antisemitism research. It also allows for transparent citation chains, where each classification is linked to prior empirical findings.

Looking ahead, one promising direction lies in extending this approach toward what might be called *context-engineered* or *complete-context* architectures. Such systems would aim to provide models with the same interpretive information that humans sometimes rely on when judging antisemitic expression. This includes both *cotextual* information––the immediate textual surrounding/environment of a post––and *contextual* information––the extralinguistic or referential background knowledge that shapes interpretation.

A future context-engineered model could integrate several complementary information types, each capturing a distinct facet of meaning:

- *Conversational cotext*, enabling models to assess how the existing lexical material is taken up within the thread.
- *Current and historical event context*, retrieved through time-filtered semantic search to situate posts responding to news, anniversaries, or historical references.
- *Named-entity context*, resolved through entity-linking databases to clarify who or what is being referenced.
- *Situational metadata*, including timestamps, platform, and engagement indicators, which help constrain relevance, chronology, scope.

Each of these cotextual and contextual layers could be retrieved through a combination of semantic search and structured lookup, then injected into model prompts in a standardized format. While still theoretical, this concept points toward models that reason with multiple evidence types rather than isolated text fragments. In this sense, RAG could evolve into a broader paradigm of *context-engineered inference*, where classification operates as a structured interpretive process approximating human expert reasoning.

Exploratory work in this direction is currently being discussed within the BSA Command Center, which plans to develop prototypes integrating external knowledge retrieval and context-layered reasoning.

4.4.4 Comparative evaluation

Our comparative evaluation showed a clear trajectory:

- Keyword spotting → high precision but low recall, blind to implicit forms
- *BERT fine-tuning* → improved context awareness but limited by annotation quality (F1 ≈ 0.70 for antisemitic texts, missing about one third of cases).
- *LLM prompting* → flexibility, multilingual reach, but prone to inconsistency.
- *RAG-enhanced LLMs* → best balance of scalability, accuracy, and interpretability, though still dependent on curated knowledge.
- *Context-engineered LLMs* → theoretically capable of near-expert accuracy when provided with both conversational context and relevant contextual information, replicating the interpretive environment available to human annotators.

Their comparative evaluation echoes our own findings. Fine-tuned BERT models reached F1 ≈ 0.70 for antisemitism, missing about one third of cases, while GPT-3.5 improved recall (F1 > 0.77) but at high computational and financial cost (Steffen et al., 2024). Patel et al. (2025) confirm this pattern: despite respectable zero- and few-shot performance, antisemitism remains markedly under-detected compared to other hate categories, with model errors clustering precisely in the gray zones where even expert coders disagree.

Beyond the top-performing models (Gemini-2.5-Flash variants, Llama-3.3-70B, and Kimi-K2), BSA also evaluated several mid-tier LLMs. Their macro F1 scores were:

- GPT-4-Turbo: 0.77
- GPT-4o: 0.81
- GPT-4o-mini: 0.81
- Gemini-1.5-Flash: 0.84

These models outperform early fine-tuned BERT architectures, but they lag behind the highest-performing models. Their patterns of error mirror many of the same issues seen in the larger models—particularly difficulties with coded conspiratorial shorthand, Holocaust numerology, and slur-based harassment—but occur with greater frequency, contributing to lower recall and reduced robustness in borderline cases.

Importantly, the gap between these mid-tier models and top-tier models appears driven less by access to world knowledge and more by difficulty with:

- following long-form domain-specific instructions, and
- correctly applying the ambiguity rule embedded in the shared prompt.

In short, the weaker models did not fail in qualitatively different ways—they failed in the same way but more often.

Model-specific strengths and limitations:

Gemini-2.5-Flash (dynamic_thinking) *(F1: 0.94)*

- Best overall balance of precision and recall.
  - o Only model to correctly classify:
  - o "Hi Jew, I'm Dad" as Not Antisemitic (ambiguity rule)
  - o the long counterspeech post as Not Antisemitic
- *Weaknesses*: Missed older trope cases such as "white holocaust" and "dancing Israelis."

Gemini-2.5-Flash (no_thinking) (F1: 0.91)

- More conservative than its dynamic counterpart.
- Correctly flagged slur-mixing posts like "Jews and Jeets 🤡."
- Frequently missed conspiracy-coded examples and Holocaust inversions.

Llama-3.3-70B-Instruct-Turbo (F1: 0.88)

- Strong on identifying:
- Holocaust inversion (e.g., "white holocaust")
- Ideological antisemitic logics
- *Weaknesses*: Over-flags ambiguous cases ("Hi Jew, I'm Dad"); Occasionally under-flags attacks on Jewish institutions (AIPAC example)

Moonshot Kimi-K2 (F1: 0.91)

- Most sensitive to coded antisemitic markers:
- Correct on "dancing israelis"
- Correct on "trans-Jew" insult
- Correct on "Jews and Jeets 🤦"
- *Weaknesses:* Tends to over-interpret harsh anti-Israel rhetoric as antisemitic; Lower precision on borderline political criticism (Figures 6–8).

These findings show that LLM performance is driven by two interacting factors:

- *Instruction adherence* (ability to follow domain-specific prompts)
- *World knowledge* (ability to recognize culturally encoded antisemitic meaning)

Models that excel at both—such as *Gemini-2.5-Flash (dynamic_thinking) (F1: 0.94)*—outperform others. Models that are strong in only one dimension show predictable error patterns:

- *Llama-3.3-70B (F1: 0.88)* is strong on ideological logic but weak on ambiguity rules.
- *Gemini-2.5-Flash (no_thinking) (F1: 0.91)* is strong on instruction adherence but weaker on coded trope recognition.
- *Kimi-K2 (F1: 0.91)* excels at recognizing coded antisemitic markers but struggles with nuanced political criticism boundaries.

Mid-tier LLMs *(F1: 0.77–0.84)* show the same error types as top models, but more frequently, particularly in:

- Applying ambiguity rules consistently
- Recognizing subtle coded language
- Balancing sensitivity with precision in borderline cases

4.5 Adaptability, evasion, and multimodal hate

A recurring insight from both qualitative and computational work is that antisemitism is not static. It adapts to its environment, exploits technological affordances, and deliberately evades detection. Online spaces accelerate this dynamic: the velocity of communication, the absence of accountability, and the reward structure of algorithms encourage users to experiment with new coded forms of expression.

4.5.1 Tactical evasion

Users frequently modify their language to bypass content moderation. Strategies include:

- *Orthographic manipulation*: *"Iisrael," "Zion@zi,"* or *"j00z"* — intentionally distorted spellings that evade keyword filters.
- *Compound neologisms*: Terms like *"Zionazis"* or *"holohoaxers"* blend slurs with historical references, packaging antisemitism in novel linguistic forms.
- *Rhetorical indirection*: Instead of direct accusations, speakers pose loaded questions (*"Why do they always own the banks?"*) or rely on implication (*"You know who really controls Hollywood"*).

These tactics demonstrate not only creativity but also a form of cat-and-mouse logic: as moderation systems improve, users find new linguistic detours. Annotation teams face similar difficulties, as

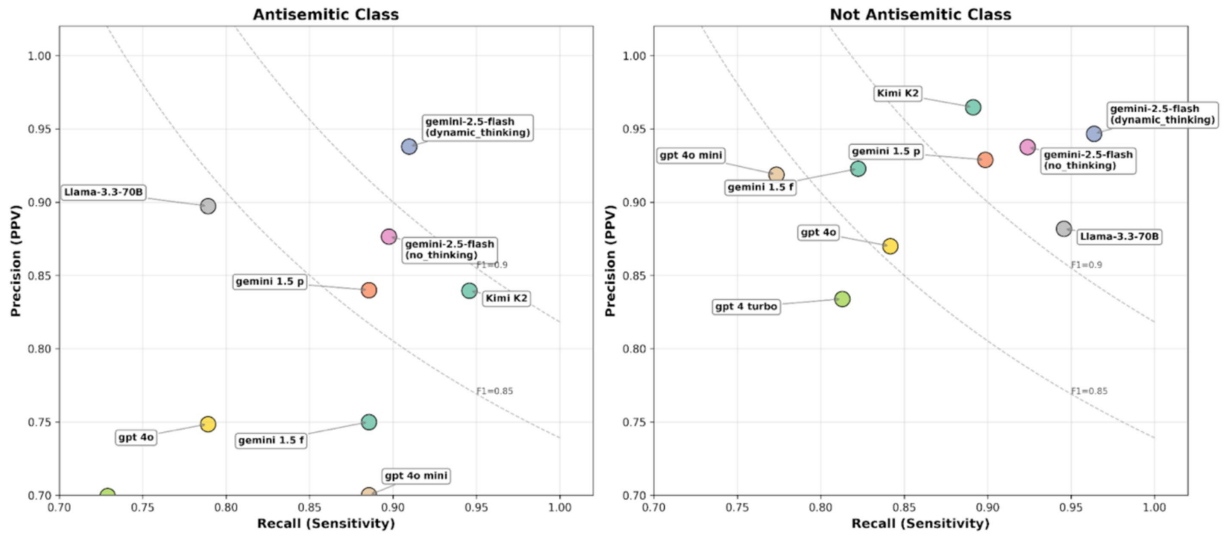


FIGURE 6
Precision–recall by class (n = 445 posts). Two scatter plots show model trade-offs for antisemitic (left) and non-antisemitic (right) content; recall on x-axis, precision on y-axis, dashed F1 iso-lines (0.85, 0.9) indicate performance benchmarks. Gemini-2.5-flash (dynamic_thinking) achieves the highest precision-recall balance on both panels, positioning in the top-right excellence zone. For antisemitic content: Kimi K2 maximizes recall but with lower precision (liberal strategy), while Llama-3.3-70B emphasizes precision over recall (conservative approach). For non-antisemitic content: Kimi K2 shifts to high precision with moderate recall, while Llama-3.3-70B maintains balanced performance. Model positioning reveals strategic trade-offs between conservative (high precision, lower recall) and liberal (high recall, lower precision) detection approaches.

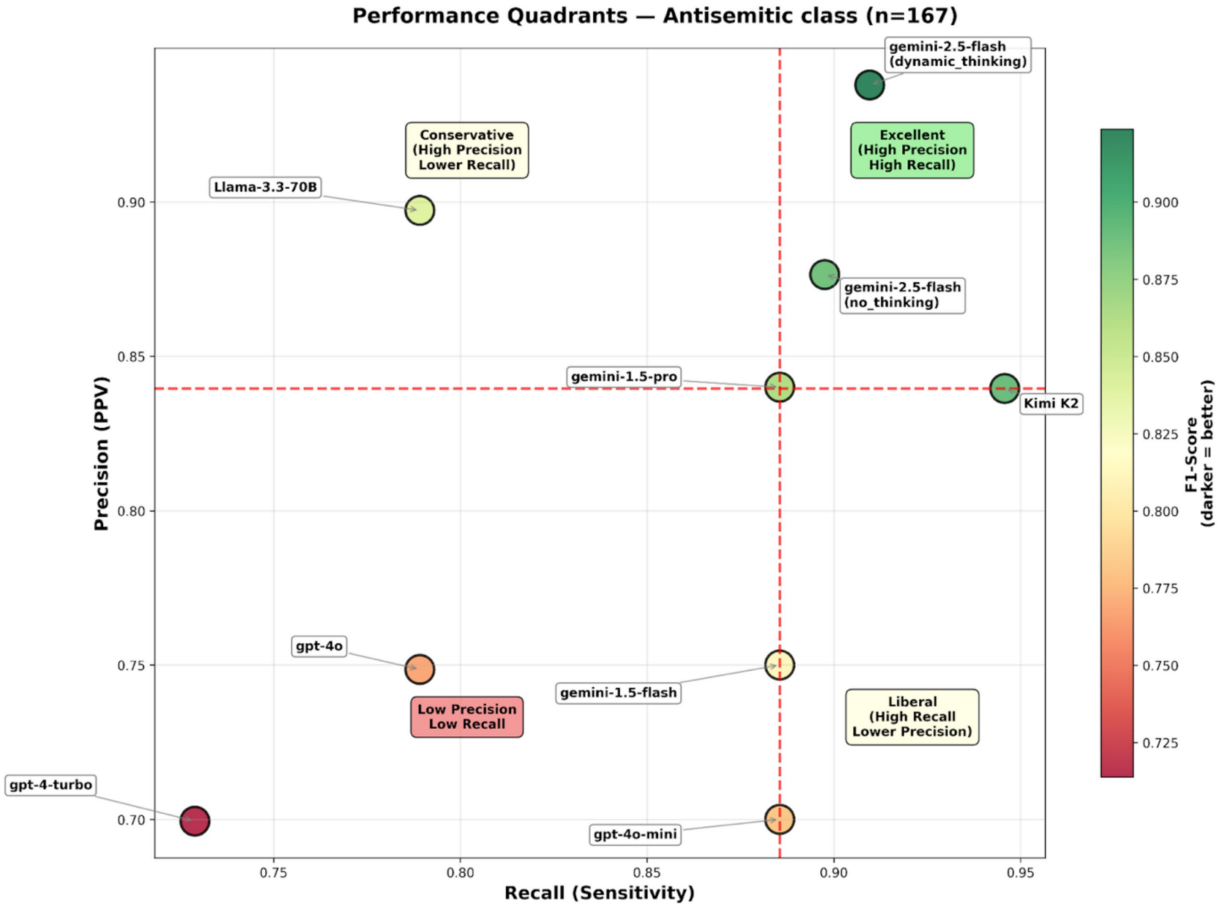


FIGURE 7
Performance quadrants—antisemitic class (n = 167). Scatter plot of precision (y) vs. recall (x); dashed medians split four strategy quadrants, point color encodes F1 (darker = better). gemini-2.5-flash (dynamic_thinking) lands in the Excellent zone (high precision and recall). Llama-3.3-70B is Conservative (high precision, lower recall). Kimi K2 is Liberal (high recall, lower precision). Gemini-2.5-flash (no_thinking) falls below both medians, indicating tuning need. Use Conservative when false positives are costly; Liberal when misses are costlier.

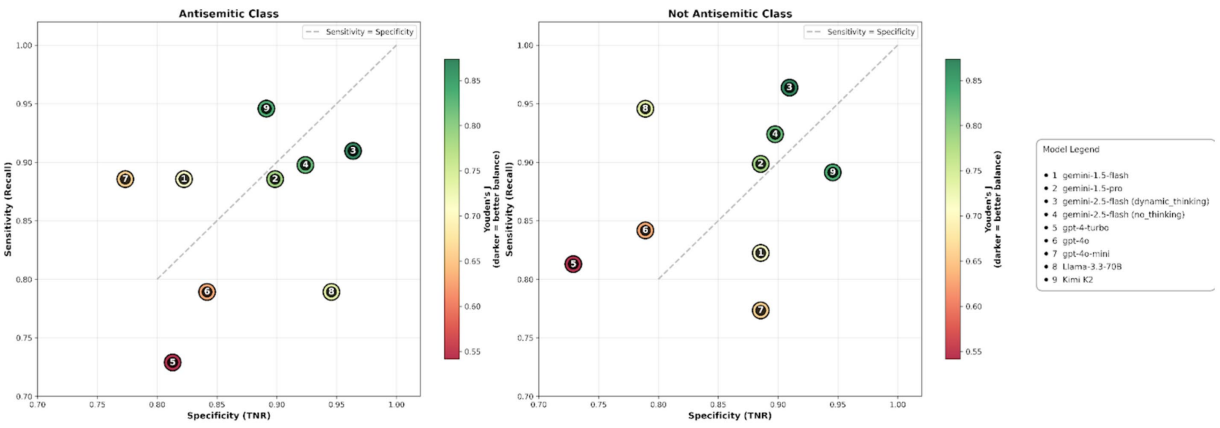


FIGURE 8
Sensitivity vs. specificity by class. Paired panels plot Sensitivity (y) against Specificity (x) with point color showing Youden's J (darker = better balance). Antisemitic class: Kimi K2 maximizes sensitivity at high specificity; gemini dynamic_thinking offers the best overall balance; Llama-3.3-70B is specificity-oriented with lower sensitivity. Non-antisemitic class: Gemini dynamic_thinking attains the strongest balance, while Kimi K2 is more conservative (high specificity, lower sensitivity).

evasion strategies blur the boundary between intentional antisemitism and eccentric spelling.

4.5.2 Implicit and latent codes

As highlighted earlier, antisemitism thrives in gray areas. Latent expressions often remain socially intelligible while avoiding direct reference. Examples include:

- "*The lobby*" → may denote a generic political lobby, or specifically AIPAC/Jewish influence.
- "*Globalist elites*" → in some contexts a critique of neoliberal capitalism; in others a coded reference to Jews.
- "*Child-killers*" → tied to specific war imagery, but also resonant with the blood libel stereotype.
- *Soros references* → can oscillate between targeted criticism and activation of conspiratorial tropes.

This flexibility is precisely what makes antisemitism resilient: it can camouflage itself as legitimate critique, tapping into mainstream discourses of anti-elitism, anti-imperialism, or human rights. For AI, the difficulty lies in distinguishing when context tips the balance from political speech to antisemitic projection.

4.5.3 The multimodal challenge

Textual analysis alone is insufficient in today's digital environments. Antisemitic meaning increasingly appears in multimodal assemblages, where images, memes, emojis, and videos carry as much weight as words.

- *Memes*: Popular meme formats (e.g., *Distracted Boyfriend*, *Drake Hotline Bling*) are repurposed to insert antisemitic analogies—portraying Jews or Israel as deceptive lovers, greedy bosses, or manipulative elites.
- *Emojis*: Watermelons (🍉), paragliders (🪂), and Palestinian flags (PS) function as shorthand for solidarity with Hamas or celebration of violence, particularly after October 7. In other contexts, emojis serve as ironic ridicule (🤦‍♂️😂 *oy vey*).
- *Profile images*: Swastikas, Hitler portraits, or cartoon caricatures appear as avatars, signaling antisemitic alignment without requiring textual content.
- *Video and sound*: Remix culture allows antisemitic clips, chants, or Hitler speeches to circulate in disguised form, often framed as "satire" or "historical education."

For computational systems, multimodality introduces a higher order of complexity: models must align visual, textual, and symbolic cues, while also integrating cultural literacy. A crying-laughing emoji after news of murdered Israelis is intelligible as antisemitic mockery only when the temporal and discursive context is considered.

4.5.4 Shifting normative boundaries

One of the most disturbing dynamics observed since October 7 is the normalization of open antisemitism. Where implicit or coded strategies once dominated, many users now bypass camouflage altogether:

- Calls to violence ("*Death to all Zionists*") circulate unfiltered.
- Glorification of terror ("*October 7 was a victory worth celebrating*") is expressed without irony.
- Mockery of victims ("*Cry harder, settlers*") no longer seeks plausible deniability.

This turn toward blunt self-positioning reflects what we have elsewhere described as a politics of transgression: the deliberate crossing of taboos to redefine the boundaries of permissible speech. Algorithms amplify such content because it provokes engagement, while community norms adjust through repeated exposure.

For annotation and AI alike, this creates a paradox:

- *Explicit antisemitism* is easier to detect computationally, but it is now so widespread that moderation systems are overwhelmed.
- *Implicit forms* remain subtle and evasive, requiring contextual interpretation.
- *Multimodal signals* often escape both categories, circulating below detection thresholds while still reinforcing antisemitic narratives.

4.6 Structural limitations and data scarcity

Even the most advanced modeling approaches face structural constraints. These limitations stem less from technical capacity than from the availability, quality, and representativeness of training data. Antisemitism as a discursive phenomenon is not only elusive but also under-documented in computational resources.

4.6.1 Lack of high-quality datasets

Unlike racism, misogyny, or homophobia, antisemitism is rarely represented in large-scale labeled corpora. Existing hate speech datasets often collapse it into a generic "hate" category, obscuring its specific logics and tropes. This is problematic because:

- Antisemitism does not always rely on slurs or explicit hate words.
- Its implicit and coded forms demand fine-grained annotation informed by historical and cultural knowledge.
- Cross-linguistic and cross-platform variations (e.g., German Holocaust trivialization, U.S. "Zionist lobby" frames, Arabic dehumanization tropes) cannot be captured through English-only corpora.

Without tailored datasets, even the best models will misclassify or overlook antisemitic discourse.

4.6.2 Imbalance and bias in annotation

Annotation is resource-intensive. In *Decoding Antisemitism*, teams coded over 300,000 comments, but this remains small compared to the billions of posts circulating online. Moreover, annotation is prone to bias and imbalance:

- *Overrepresentation* of explicit antisemitism may lead models to underperform on subtle cases.
- *Underrepresentation* of implicit cases risks reinforcing the invisibility of gray-zone antisemitism.
- *Cultural bias* arises when annotators come from only one linguistic or national background, limiting their recognition of culturally specific codes (e.g., Yiddish expressions, Islamic references, German Nazi analogies).

Intercoder disagreements are especially pronounced in borderline cases. Some annotators classify "Israel commits genocide" as political critique; others view it as antisemitic Holocaust inversion. Such divergences highlight the need for clear, theory-based coding manuals and adjudication by experts. As Becker and Troschke (2023) emphasize, only by pooling shared world knowledge and applying conservative attribution guidelines can coders maintain reliability when dealing with implicitness and ambiguity.

4.6.3 The problem of rarity

Antisemitism is pervasive, but it is rare in statistical terms. In large datasets, antisemitic comments may constitute only 1–2% of all entries. For machine learning models, this imbalance leads to:

- *False negatives* (antisemitic content classified as neutral).
- *False positives* (neutral content classified as antisemitic because of keyword overlap).
- *Difficulty in training models* to generalize, as rare events are easily overshadowed by more frequent patterns.

Data scarcity is further exacerbated by platform restrictions: researchers often face obstacles in collecting comments at scale due to API limitations, legal constraints, or ethical concerns. They confirm this imbalance empirically: in their English-language subcorpus, only 10% of comments were antisemitic, mirroring our observations that annotation is disproportionately costly relative to statistical prevalence (Steffen et al., 2024). In their dataset, nearly 10 non-antisemitic texts appeared for every antisemitic one, and even with augmentation strategies such as back-translation or word replacement, model performance did not substantially improve (Pustet and Mihaljević, 2024). This reinforces our observation that annotation costs are disproportionately high relative to statistical prevalence.

4.6.4 Need for iterative, interdisciplinary collaboration

Given these constraints, no single discipline can solve the problem. What is needed is an iterative research cycle:

1. *Qualitative analysis* → identifies new antisemitic codes and discursive patterns.
2. *Annotation* → integrates these insights into structured, reliable labels.
3. *Model training* → builds classifiers capable of detecting known patterns.
4. *Error analysis* → reveals blind spots, which feed back into qualitative exploration.

This cycle requires ongoing collaboration between linguists, historians, data scientists, and AI engineers. It also demands infrastructure: centralized databases, interoperable annotation schemes, and funding for sustained dataset curation.

4.7 Policy and ethical considerations

The methodological and technical challenges of detecting antisemitism online are inseparable from broader policy and ethical questions. At stake is not only the effectiveness of detection but also the balance between combating hate speech and preserving free expression in democratic societies.

4.7.1 Risks of over-blocking and under-blocking

Automated moderation systems face a persistent dilemma:

- *Over-blocking*: When content is removed too aggressively, legitimate political critique or even neutral references to Jewishness risk being censored. As Discourse Report 5 (Chapelan et al., 2023) shows, the Perspective API systematically inflated toxicity scores for comments containing words like "Jew" or "Israel"—regardless of stance. This "false positive bias" has been confirmed in other studies (Dixon et al., 2018; Hutchinson et al., 2020; Röttger et al., 2021).
- *Under-blocking*: When antisemitic content remains online, it risks normalization, emboldening perpetrators, and causing emotional harm to Jewish users.

The balance is delicate. Consider the phrase *"Israel is committing genocide."* Depending on context, it may represent:

- a (contested) political critique of Israeli policy, or
- an antisemitic inversion that equates Jews with Nazis, a classical trope of Holocaust relativization.

Models that lack contextual awareness risk misclassifying both ways, either suppressing critical speech or permitting hate to circulate unchecked.

4.7.2 Platform accountability

Current approaches to moderation often shift responsibility onto individual users, who must report violations, or onto opaque algorithms. This status quo is insufficient. The EU's Digital Services Act (DSA) represents a first attempt to address these challenges by mandating greater transparency in content moderation and algorithmic processes. Yet the algorithmic amplification of polarizing and hateful content means platforms are not neutral hosts but active curators of discourse.

- Recommendation engines systematically boost content that triggers engagement, regardless of harm.
- Moderation tends to focus on slurs or direct threats while ignoring implicit or multimodal forms of antisemitism.
- Enforcement is inconsistent across languages and regions, leaving vulnerable communities unevenly protected.

Without collaboration with academic experts and access to high-quality datasets, even regulatory frameworks like the DSA will remain limited in impact. Discourse Report 5 of the Decoding Antisemitism project demonstrates this tension: neutral or even educational comments containing words like "Jew" or "Israel" were frequently misclassified as toxic due to keyword bias, while at the same time coded antisemitic expressions (emojis, irony, neologisms) often slipped through undetected (Dixon et al., 2018; Hutchinson et al., 2020; Röttger et al., 2021; Chapelan et al., 2023). Such cases illustrate why platform accountability must extend beyond compliance checklists to include context-sensitive moderation and independent auditing.

4.7.3 The democratic stakes

Unchecked antisemitism online has consequences beyond Jewish communities. As our case studies show, it frequently converges with anti-democratic discourses, misogyny, racism, and conspiracism. Antisemitism thus functions as a gateway resentment, mobilizing broader illiberal currents and undermining democratic trust.

Conversely, overzealous censorship risks reproducing the very authoritarian dynamics it seeks to counteract. Public legitimacy depends on moderation systems that are transparent, proportionate, and accountable. To achieve this, platforms must:

1. *Invest in context-sensitive moderation* informed by cultural and historical expertise.
2. *Open their data* to independent researchers for auditing and monitoring.
3. *Develop feedback loops* between human moderators, academic experts, and automated systems.

4.7.4 Ethical responsibility of researchers

Finally, researchers themselves face ethical obligations. Building detection systems means grappling with:

- *Privacy concerns* when handling user data.
- *Potential misuse* of detection models by authoritarian regimes to suppress dissent under the pretext of fighting hate speech.
- *Impact on communities*: Jewish communities expect protection from online antisemitism, but poorly designed interventions risk re-traumatization or further marginalization.

The guiding principle must therefore be democratic resilience: safeguarding open debate while curbing the spread of hate that corrodes the very foundations of pluralistic societies.

5 Limitations

5.1 Data scope and sampling

The datasets are narrowly focused, which limits the generalizability of the findings. The 2023 (23AIC) dataset is restricted exclusively to UK YouTube channels, while the 2025 (25 DC) analysis focuses on a specific set of eight mainstream English-language YouTube outlets. Neither study includes other platforms, such as Facebook, nor do they analyze non-mainstream channels. Both datasets are event-driven, capturing comments during acute crisis periods—the immediate aftermath of the October 7 attacks and a 48-h window following the 2025 Washington shooting. The high rates of antisemitism observed (e.g., 30–40% in 2023 and 43% in 2025) may reflect crisis-driven discourse rather than typical baseline online behavior. The 25 DC study's sampling was also fixed at 200 comments per channel, which may not be fully representative of the total volume or distribution of engagement on each video.

5.2 Conservative treatment of ambiguity (gray-zone cases)

As noted in the sections on gray-zone utterances, we adopt a deliberately conservative annotation rule: whenever a statement plausibly supports both an antisemitic and non-antisemitic reading, we label it non-antisemitic. This approach prevents over-labeling, but it also means results can underestimate how much antisemitism there really is and make subtle cases harder to detect. Downstream models trained on these labels will inherit that bias, tending to miss borderline cases, especially if not fully context-aware. To mitigate this, future systems could add a parallel "possible antisemitic reading" flag.

5.3 Cross-jurisdictional ambiguity in "gray-zone" interpretation

Finally, the distinction between political critique and antisemitic projection varies considerably across legal and cultural contexts. What counts as protected speech in the United States may, under EU or German jurisprudence, qualify as group defamation or hate speech. This divergence complicates both annotation and model training: a statement deemed borderline in one jurisdiction may be categorized differently in another. While our present study applies a linguistically grounded framework independent of specific legal systems, future research should systematically examine how national regulations and cultural norms shape the operational boundaries of antisemitic discourse detection.

5.4 Evolving semantics and missing conversational context for AI models

As highlighted in the discussion of shifting meanings, terms like "Zionist," "Free Palestine," or references to "the lobby" exhibit pragmatic drift across events and communities. In addition, many items lack full co-text (thread structure) and context (event metadata, world-knowledge hooks). Under these conditions, both keyword systems and context-aware models can struggle to disambiguate political critique from antisemitic projection, risking over- or under-detection. Our experimental evaluations therefore reflect upper bounds conditioned on curated inputs; real-world performance will degrade when thread co-text, temporal anchors, or entity disambiguation are absent.

5.5 Multimodality constraints

Our multimodal treatment focuses on emojis and simply symbols. Images, memes, video, audio, and profile artifacts were not systematically annotated or modeled, limiting conclusions about cross-modal signaling.

5.6 Model evaluation caveats

Newest LLM results depend on specific prompts, hyperparameters, and curated, co-text-independent subsets. Preliminary LLM findings are exploratory and not peer-reviewed. Real-time costs, rate limits, and model drift constrain scalability and reproducibility.

6 Conclusion

The analysis presented here has highlighted the linguistic, multimodal, computational, and political challenges of confronting antisemitism in digital spaces. Our empirical case studies—the aftermath of October 7 (2023) and the Washington, D.C., shooting of May 2025—demonstrate how antisemitism manifests across a spectrum from explicit incitement to implicit tropes and gray-zone expressions. This range underscores why antisemitism cannot be reduced to slurs or explicit hate terms. Instead, it must be understood as a discursive phenomenon: historically layered, context-sensitive, and adaptable to new communicative environments.

The linguistic challenge lies in recognizing implicitness, irony, and camouflage. This requires a conservative interpretive approach to avoid false positives while still capturing the prevalence of implicit antisemitism—a form that is not peripheral but structurally dominant in online discourse of the political mainstream (Becker and Troschke, 2023). The multimodal challenge extends this difficulty: antisemitism now circulates not only through words but also through memes, emojis, and other visual-symbolic codes. The computational challenge is twofold. While machine learning and LLM-based models offer new tools for detection, they require fine-tuned, high-quality datasets grounded in expert annotation. Yet data scarcity remains a structural obstacle, with antisemitism both statistically rare and underrepresented in labeled corpora. Finally, definitional clarity is indispensable. Without a shared framework distinguishing legitimate political critique from antisemitic delegitimization, neither human coders nor AI systems can operate consistently or legitimately.

These findings have direct implications for policy and democratic practice. Over-blocking risks silencing critical voices; under-blocking normalizes hate. Both outcomes erode the integrity of the digital public sphere. Platforms must therefore assume greater accountability—not only for content removal but also for the algorithmic amplification of harmful narratives. Equally, researchers carry ethical responsibilities to ensure that detection tools are technically robust and resistant to misuse.

Looking ahead, progress depends on iterative, interdisciplinary collaboration. Linguists, historians, data scientists, and policymakers must work together in a continuous cycle of analysis, annotation, modeling, and critical reflection. Only such collaboration can capture the evolving dynamics of antisemitism and develop tools that balance accuracy, contextual sensitivity, and democratic legitimacy.

Antisemitism is not a relic of the past nor a marginal phenomenon. It is a structurally embedded, adaptive discourse that exploits the affordances of the interactive web to reassert itself in mainstream publics. It thrives on communication latency and a politics of transgression, leveraging the political-cultural opportunity structures of the digitally restructured public sphere (Becker and Rensmann, 2023). Confronting it requires not only technical innovation but also renewed commitments to accountability, transparency, and pluralism. The task ahead is formidable but essential if digital societies are to remain open, inclusive, and resilient. Recent evaluations (Steffen et al., 2024; Patel et al., 2025) converge on the same point: neither toxicity APIs nor current transformer architectures suffice for the nuanced detection of antisemitism. Both studies underscore the need for context-sensitive, interdisciplinary approaches. Our findings extend this evidence by combining discourse-analytic case studies with computational trials, pointing toward hybrid architectures that anchor LLM-based reasoning in expert-curated corpora and theory-guided annotation.

This study is not exhaustive. Its datasets remain limited in size and language scope, and further research should extend these analyses across platforms and cultural contexts. Nevertheless, the findings illuminate broader dynamics that are crucial for both scholarship and democratic practice.

Data availability statement

The datasets analyzed in this study are stored at the local research repository of the Center for Research on Antisemitism (ZfA) at Technische UniversitÃ¤t Berlin and at the Blue Square Alliance Against Hate (Massachusetts, USA). Due to platform terms of service and data-protection regulations, the datasets are not publicly available but can be accessed upon reasonable request for academic, non-commercial purposes. Requests to access these datasets should be directed to mjb15@posteo.de, JordanB@bsqa.org.

Author contributions

MB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Writing – original draft, Writing – review & editing. JB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. OS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Anti-Defamation League (2020). Proud boys' bigotry is on full display. ADL. Available online at: https://www.adl.org/resources/article/proud-boys-bigotry-full-display (Accessed 17 October 2025).

Becker, M. J. (2021). Antisemitism in reader comments: analogies for reckoning with the past. London: Palgrave Macmillan.

Becker, M. J. (2025). Kvetching intensifies: Antisemitic discourse online after the Washington Embassy shooting. Substack, 23 May 2025. Available online at: https://substack.com/home/post/p-164232280 (Accessed 17 October 2025).

Becker, M. J., Ascone, L., Bolton, M., Chapelan, A., Haupeltshofer, P., Krugel, A., et al. (2023). Celebrating terror: antisemitism online after the Hamas attacks on Israel. preliminary results I. Berlin: Technische Universität Berlin, Center for Research on Antisemitism.

Becker, M. J., and Fillies, J. (2024). "KI im Trainingslager: Wie Künstliche Intelligenz gegen antisemitische Codes und die Normalisierung von Hassrede im Netz eingesetzt werden kann" in Code & Vorurteil. Über Künstliche Intelligenz, Rassismus und Antisemitismus. eds. D. Schnabel, E. Berendsen, L. Fischer and M.-S. Adeoso (Berlin: Verbrecher Verlag).

Becker, M. J., and Rensmann, L. (in press). The dynamics of digital antisemitism: Anti-Jewish narratives in the aftermath of October 7th. In: T. Pittinsky (ed.), Antisemitism online: an ancient hatred in the modern world. (2023), Oxford: Oxford University Press.

Becker, M. J., and Troschke, H. (2023). "Decoding implicit hate speech: the example of antisemitism" in Challenges and perspectives of hate speech research. eds. C. Strippel, S. Paasch-Colberg, M. Emmer and J. Trebbe, vol. 12 (Berlin, Germany: Digital Communication Research), 335–352.

Becker, M. J., Troschke, H., Bolton, M., and Chapelan, A. (2024). Decoding antisemitism: a guide to identifying antisemitism online London Palgrave Macmillan/Springer Nature. Available online at: https://link.springer.com/book/9783031492372

Bergmann, W., and Erb, R. (1986). Kommunikationslatenz, Moral und öffentliche Meinung. Theoretische Überlegungen zum Antisemitismus in der Bundesrepublik Deutschland. *Kölner Z. Soziol. Sozialpsychol.* 38, 223–246.

Blatter, J., Blue Square Alliance Against Hate (2025). Testing next-generation AI to better detect antisemitism. Blue Square Alliance, Command Center Insights. Available online at: https://www.bluesquarealliance.org/command-center-insights/testing-next-generation-ai-to-better-detect-antisemitism (Accessed October 17, 2025).

Chapelan, A., Ascone, L., Becker, M. J., Bolton, M., Haupeltshofer, P., Krasni, J., et al. (2023). Decoding antisemitism: an AI-driven study on hate speech and imagery online: Berlin, Center for Research on Antisemitism.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, New York, NY, USA: Association for Computing Machinery (ACM). 67–73.

Fielitz, M., and Thurston, N. (2019). Post-digital cultures of the far right: online actions and offline consequences in Europe and the US. Bielefeld: transcript.

Halevy, K. H., Mendelsohn, J., Younger, N., Rossman-Benjamin, T., Park, C. Y., Tsvetkov, Y., et al. (2024). On the importance of nuanced taxonomies for LLM-based understanding of harmful events: A case study on antisemitism. Eighth Widening NLP Workshop (WiNLP 2024) Phase II. Available online at: https://openreview.net/forum?id=8H67u1GlS6 (Accessed 17 October 2025).

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Unintended machine learning biases as social barriers for persons with disabilities. New York, NY, USA: ACM Access/SIGACCESS.

International Holocaust Remembrance Alliance (2016). Working definition of antisemitism. Available online at: https://holocaustremembrance.com/resources/working-definition-antisemitism (Accessed 17 October 2025).

Jikeli, G., Karali, S., Miehling, D., and Soemer, K. (2023). Antisemitic messages? A guide to high-quality annotation and a labeled dataset of tweets. *arXiv preprint arXiv*:2304.14599. doi: 10.48550/arXiv.2304.14599

Patel, J., Mehta, H., and Blackburn, J. (2025). Evaluating large language models for detecting antisemitism. *arXiv preprint arXiv*:2509.18293 (Accepted to EMNLP 2025 Main Conference.). doi: 10.48550/arXiv.2509.18293

Pustet, M., and Mihaljević, H. (2024). Automated detection of antisemitic texts: Is context all we need? In: M. Beckeret al. Decoding antisemitism: an AI-driven study on hate speech and imagery online. 6. Berlin: Technische Universität Berlin, Center for Research on Antisemitism. Available online at: https://decoding-antisemitism.eu/publications/sixth-discourse-report (Accessed 17 October 2025).

Rensmann, L. (2017). The politics of unreason: the Frankfurt school and the origins of modern antisemitism. Albany: State University of New York Press.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. B. (2021). HateCheck: functional tests for hate speech detection models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & 11th IJCNLP, (Volume 1 Association for Computational Linguistics, Stroudsburg, PA, USA: Long Papers). 41–58.

Steffen, E., Pustet, M., and Mihaljević, H. (2024). "Algorithms against antisemitism? Towards the automated detection of antisemitic content online" in Antisemitism in online communication: transdisciplinary approaches to hate speech in the twenty-first century. eds. M. J. Becker, L. Ascone, K. Placzynta and C. Vincent (London: Open Book Publishers).

Weinberg, D. B., Levy, M. D., Edwards, A., Kopstein, J. S., Frey, D., Antonaros, P., et al. (2025). Hidden in plain sight: antisemitic content in QAnon subreddits. *PLoS One* 20:e0318988. doi: 10.1371/journal.pone.0318988

Wodak, R. (2015). The politics of fear: what right-wing populist discourses mean. London: SAGE Publications.