You Frame It: How Conceptual Representations Shape LLM Detection and Reasoning about Antisemitism

Anonymous ACL submission

Abstract

LLMs enable the integration of external conceptual resources at inference time, creating new opportunities for detecting ideologically and historically complex phenomena such as antisemitism. We investigate how different forms of conceptual grounding affect antisemitism detection and explanation behavior across four state-of-the-art LLMs. Using two expert-annotated datasets, we compare definitional, fine-grained taxonomic, example-augmented, and large-context representations of antisemitism. We find that fine-grained taxonomic representations substantially improve recall, while simultaneously reducing precision. Surprisingly, supplying substantially larger conceptual resources yields no additional quantitative benefit. Post-Holocaust antisemitism poses the most persistent challenge across models and configurations. Analysis of explanations further reveals systematic limitations including overproduction of conceptual references, reliance on lexical cues, overconfidence, and difficulties with subtle or justificatory forms of antisemitism. Our findings highlight both the potential and the remaining limitations of conceptually grounded LLMs for antisemitism detection and reasoning.[1]

1 Introduction

Despite advances in hate speech detection using Large Language Models (LLMs) (Gilardi et al., 2023; Li et al., 2024), recent studies indicate that antisemitic content detection remains particularly challenging (Steffen et al., 2024; Patel et al., 2025). Models struggle particularly with implicit, coded, or trope-based rhetoric (Mihaljević and Steffen, 2022; Mendelsohn et al., 2023), and may reproduce or amplify anti-Jewish biases (Anti-Defamation League, 2025).

Antisemitism poses a particular challenge for automated detection due to its conceptual and rhetorical complexity.[2] It operates through a heterogeneous, historically evolving repertoire of stereotypes, conspiracy theories, and narratives which are not always explicitly hateful, e.g., claims of Jewish financial power and control (Becker et al., 2024b). Even expert annotators show substantial disagreement (Steffen et al., 2023), indicating that LLMs require specific contextual knowledge to capture these ideological patterns.

The knowledge encoded in LLMs is opaque, fixed at training time, and potentially shaped by the same biases the models are expected to detect. Prompting offers the possibility to inject external conceptual representations at inference time, for example through definitions, taxonomies, or examples. While external knowledge infusion has been explored in adjacent areas (Yang et al., 2025; Melis et al., 2025), little is known about how different forms of conceptual grounding affect detection and classification of different forms of antisemitism.

In this work, we systematically evaluate the effect of external conceptual representations on antisemitism detection and classification across four state-of-the-art LLMs: Gemini-2.5-Flash, Claude Sonnet 4.6, GPT-5.4, and LLaMA-3.3-70b-instruct. We compare four prompting configurations: a baseline without external knowledge; the *IHRA* Working Definition of Antisemitism (hereafter: *IHRA*) (International Holocaust Remembrance Alliance, 2016); and a fine-grained taxonomy derived from the *Decoding Antisemitism Lexicon* (hereafter: *Lexicon*) (Becker et al., 2024b) with and without examples. In addition, we evaluate Gemini using the complete 550-page *Lexicon*.

Our study makes four main contributions: (1) We provide the first systematic multi-model evaluation

[1] Code is publicly available at: https://github.com/anon-researcher-hash/antisemitism-llm-knowledge-base

[2] A detailed introduction to the concept of antisemitism and its history is provided in Appendix A.1.

of external conceptual grounding for antisemitism detection and classification. (2) We show that compact but fine-grained conceptual representations substantially improve recall, while long contextual resources yield no additional benefit. (3) We demonstrate that different forms of antisemitism vary strongly in detectability across models and prompting configurations, with Israel-related and post-Holocaust antisemitism posing largest challenges. (4) We analyze explanation behavior as an indirect reasoning signal, revealing systematic overproduction of conceptual references, shortcut reliance on lexical cues, and mismatches between internal uncertainty and final predictions.

Our findings contribute both to the underexplored area of computational antisemitism research and to the broader design of more reliable LLM-based moderation and monitoring systems. Antisemitism constitutes a particularly demanding test case due to its ideological complexity, historical depth, and frequent implicitness, making it well suited for studying the limits of current LLM reasoning and external knowledge integration.

2 Related Work

2.1 LLMs for Detecting Antisemitism

LLMs achieve strong performance on a range of harmful-content detection tasks, often matching or exceeding traditional classifiers and crowd-based annotation approaches (Gilardi et al., 2023; Li et al., 2024; Roy et al., 2023). At the same time, prior work consistently reports limitations for implicit, coded, and context-dependent forms of hate speech (Kumar et al., 2024; Huang et al., 2023; Hartmann et al., 2025; ElSherief et al., 2021) and antisemitism (Mendelsohn et al., 2023), while exhibiting oversensitivity toward identity-related terms, controversial topics and ignoring intent (Zhang et al., 2024; Hartmann et al., 2025; Roy et al., 2023). Benchmarking studies additionally indicate large performance discrepancies for antisemitic content compared to other target groups (Balachandran et al., 2024).

Prompt-based antisemitism detection currently achieves substantially lower performance than typically reported for hate speech, offensive language or conspiracy theory detection (Guo et al., 2024; Roy et al., 2023). On Twitter data sampled using thematically explicit keywords, prompting open LLMs yields F_1 scores up to 0.69 (Patel et al., 2025). On more discourse-diverse datasets, performance is substantially lower, with reported F_1 scores around 0.55 even for GPT-4 (Pustet and Mihaljević, 2024; Steffen et al., 2024). These findings suggest that antisemitism remains a particularly difficult target for LLM-based moderation, especially in settings without strong lexical cues.

2.2 Conceptual Grounding and Prompting

Recent work has explored integrating external conceptual resources at inference time (Yang et al., 2025). Providing definitional, policy-based, or conceptual context can improve detection of harmful content (Atreja et al., 2025; Roy et al., 2023) and antisemitism (Becker et al., 2024a; Patel et al., 2025). Even the conceptualization and operationalization of hate speech substantially affect prompting behavior (Melis et al., 2025), with varying effects across models. In the context of antisemitism, *IHRA* constitutes the dominant definitional framework in both policy and computational research (Patel et al., 2025). However, there is still little research on how alternative or substantially more fine-grained conceptual representations affect LLM detection and reasoning behavior.

Findings regarding the inclusion of examples in prompts remain mixed. Some studies report improvements through few-shot prompting (Ghorbanpour et al., 2025), while others show strong sensitivity to prompt wording, model family, and example selection (Chae and Davidson, 2026; Reynolds and McDonell, 2021). Existing work therefore suggests that prompting performance depends not only on the amount of contextual information provided, but also on how concepts are structured and represented.

Recent long-context LLMs further enable the inclusion of substantially larger conceptual resources such as books or policy documents at inference time. However, it remains unclear whether increasing the amount of contextual material beyond compact conceptual representations further improves classification and reasoning performance.

2.3 LLM Reasoning Analysis

LLM-generated explanations are increasingly explored as a mechanism for improving transparency and human oversight in moderation systems (Ma and Kou, 2023; Kolla et al., 2024). Prior work suggests that prompting models to explain their decisions can improve classification performance itself (Agarwal et al., 2025; Atreja et al., 2025). Findings on the practical usefulness of LLM ex-

planations in moderation contexts remain mixed. Huang et al. (2023) report that ChatGPT-generated explanations for implicit hate speech are rated as clearer than human-written ones, but not necessarily more informative, while Di Bonaventura et al. (2024) find only weak alignment between LLM explanations and human expectations. Related work has further explored grounding explanations in explicit policy references. Kumar et al. (2024), for example, prompt LLMs to identify relevant community guidelines alongside classification decisions. Their manual error analysis further suggests that a substantial share of apparent false positives reflects annotation ambiguity, whereas false negatives more often stem from genuine reasoning failures. In this context, it should be noted that evaluating explanation quality and faithfulness remains challenging (Di Bonaventura et al., 2024; Zhong and Baghel, 2024), with studies frequently relying on qualitative examinations of explanations (Mendelsohn et al., 2023; Goyal et al., 2025).

Our work builds on this literature in several ways. First, we systematically compare different forms of external conceptual grounding, ranging from compact definitions to fine-grained taxonomies and large-context resources. Second, we focus on antisemitism as a particularly challenging and historically grounded form of harmful content. Third, we analyze explanations as indirect signals of conceptual reasoning by mapping generated references onto antisemitic concepts.

3 Data and Methods

3.1 Datasets

We use two expert-annotated datasets on online antisemitism that differ in sampling strategy and annotation granularity, and have served as primary sources of labeled examples in prior computational work (Becker et al., 2024a; Patel et al., 2025).

The **Bloomington** dataset (Jikeli et al., 2021) consists of English-language Twitter (now X) posts collected via the keywords *jews*, *israel*, *kikes*, and *zionazi**. Annotations follow *IHRA*, consisting of a general definitional paragraph and eleven illustrative categories. Posts labeled as antisemitic are further assigned up to two most salient *IHRA* sections (Jikeli et al., 2024).

The **Decoding** dataset (Becker et al., 2024b) contains English-language comments from mainstream social media platforms and news outlets, sampled around public events expected to elicit antisemitic discourse. The annotation is based on a fine-grained annotation scheme derived from, but extending, *IHRA*, comprising 46 codes that capture specific stereotypes, narratives, and rhetorical patterns of contemporary antisemitism. The number of sub-codes per post labeled as antisemitic is not limited. Table 1 summarizes the main characteristics of the two datasets.

Table 1: Comparison of the datasets used in this study.

Characteristic	Bloomington	Decoding
Sources	Twitter	Social media & news outlets
Time span	01/19–04/23	12/20–11/23
Sampling	Keyword-based	Discourse-driven
Size	11,311	42,799
% Antisemitic	1,953 (17.3%)	3,375 (7.9%)
Sub-labels	12 *IHRA* sections	46 codes
Sub-labels/post	1 or 2	No upper bound
Sublabel skew	Top 5: ∼90%	Top 5: ∼53%

The annotation scheme of the Decoding dataset structures a lexicon-format reference work consisting of 550 pages and 42 chapters (Becker et al., 2024b). Each chapter corresponds to a specific antisemitic concept and provides (i) a conceptual description, (ii) typical linguistic and narrative realizations, and (iii) contemporary examples. In this work, we use the set of Lexicon chapters as a structured taxonomy of antisemitic concepts.

We map both annotation frameworks (*IHRA* sections and *Lexicon* chapters) to a set of higher-level **content groups** that capture recurring dimensions of antisemitic discourse. This abstraction enables comparative and interpretable analyses across datasets and prompting configurations. Detailed descriptions, label distributions, and mappings are provided in Appendix A.2.

Both datasets were preprocessed using a unified pipeline; details are provided in Appendix A.3

3.2 Model Choice

We evaluate Gemini-2.5-Flash (Google DeepMind, 2025), Claude 4.6 Sonnet (Anthropic, 2026), GPT-5.4 (OpenAI, 2026), and LLaMA-3.3.70b-instruct (Llama Team, AI @ Meta, 2024), covering both proprietary and open-weight model families commonly examined in recent work on antisemitism detection (Becker et al., 2024a; Anti-Defamation League, 2025; Patel et al., 2025). Gemini-2.5-Flash additionally supports substantially larger input lengths and document-style inputs (e. g., PDF files). This enables evaluation of a full-lexicon con-

dition that cannot be replicated across all models due to context limitations.

3.3 Prompting Configurations and Evaluation

We use instruction-based prompting. Models are instructed to decide whether a post contains antisemitic speech while explicitly considering speaker intent; quoting or reporting antisemitic content for purposes of criticism or condemnation is not to be labeled as antisemitic. Each prompt specifies an expert role, a task description, and optionally includes structured conceptual guidance (see Appendix A.4 for full prompts).

We evaluate four prompt configurations that differ in the granularity and explicitness of the provided conceptual representation (Table 2). IHRA provides a coarse-grained definition with embedded examples, whereas STRUCT and STRUCT+EX use a fine-grained taxonomy corresponding to the *Lexicon* chapters For STRUCT+EX, examples were selected directly from the expert-authored *Lexicon*, using the first available examples for implicit antisemitic, explicit antisemitic, and non-antisemitic content.

Table 2: Prompt configurations used in the experiments.

Configuration	Structure	Examples
BASE	None	No
IHRA	Coarse	Embedded
STRUCT	Fine-grained	No
STRUCT+EX	Fine-grained	Yes

Models return a classification decision and, except for BASE, a brief explanation referencing relevant sections from the provided representation. To account for ambiguity and context sensitivity, models may respond with *Yes*, *No*, or *Unsure* (Goyal et al., 2025). Referenced sections are extracted in post-processing and mapped onto the content groups (Table 3 in Appendix A.2.3.).

We include all antisemitic posts and randomly sample 1,000 non-antisemitic posts per dataset. This yields stable estimates while keeping the computational cost manageable given the class-imbalance and the large number of model–prompt combinations evaluated. All metrics are computed on binary labels (*Yes* vs. non-*Yes*), with precision and F_1 estimated via bootstrap resampling and inverse-probability weighting.

Results are reported on the balanced (i.e., equally weighted) union of both datasets and, where relevant, on the individual datasets. Statistical significance is assessed using χ^2 tests with Holm correction; effect sizes are reported using Cohen's h.

4 Results

4.1 Effect of Knowledge Base

Figure 1 shows the F_1 scores of all models and configurations on the dataset union. Detailed results for all metrics and datasets are reported in Table 4 in Appendix A.5.

The three commercial models achieve similar peak F_1 scores (0.61–0.64), whereas LLaMA performs substantially worse (0.54). All models improve their F_1 scores when provided with external conceptual representations. Except for LLaMA, all models achieve their highest F_1 scores under STRUCT or STRUCT+EX. While LLaMA performs best under IHRA, its results for STRUCT and STRUCT+EX are comparable.

Optimal settings differ between precision and recall. STRUCT and STRUCT+EX consistently yield the highest recall, often with substantial gains; for GPT and LLaMA, recall nearly doubles relative to BASE. However, these gains generally come at the cost of lower precision, with GPT constituting the only exception. GPT is also the only model whose recall further improves through the inclusion of explicit examples. Moreover, GPT's precision remains consistently higher than recall across all configurations, while the remaining models achieve higher recall than precision in at least their best-performing setting. GPT also obtains the overall lowest recall.

These findings underscore the importance of conceptual structure. STRUCT, which provides only a fine-grained list of antisemitic concepts without explanations or examples, generally improves recall and F_1 more strongly than IHRA, despite the latter offering substantially more descriptive detail and embedded examples. This suggests that granularity and conceptual coverage may matter more than descriptive richness.

Overall, performance trends are broadly consistent across datasets despite moderate differences in difficulty. Recall is consistently higher on Bloomington than on Decoding across all models and configurations, indicating that Decoding constitutes the more challenging dataset. However, external conceptual representations, particularly STRUCT and STRUCT+EX, yield substantially larger improvements on Decoding than on Bloomington. Gemini largely closes the recall gap between datasets

under these settings. Precision differences between datasets are less consistent. Consequently, F_1 scores remain uniformly higher on Bloomington across configurations.

Finally, all models use the abstention option *Unsure* (that was mapped to *No* for metric computation) only rarely (typically <=6%), particularly on Bloomington data and positive samples from Decoding. However, there are notable exceptions for negative Decoding samples. Gemini and GPT abstain considerably more often (6–24 % and 5–19 % of cases, respectively), but interestingly, external knowledge representations reduce abstention for Gemini but increase it for GPT.

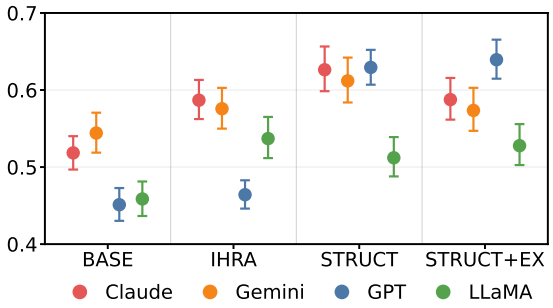

Figure 1: F_1 scores with confidence intervals for each model and configuration on the balanced dataset union.

4.2 Detection Performance Across Forms of Antisemitism

We focus exclusively on posts labeled as antisemitic by human annotators to examine whether models differ in their ability to detect particular forms of antisemitism. Figure 2 shows the deviation of each content-group recall from the mean recall of the respective model and configuration; exact values and significance tests are reported in Table 5 in Appendix A.5.

Across all models and configurations, aggressive antisemitic speech acts are detected most reliably. As reflected by the consistently positive deviations in Figure 2, these forms appear comparatively easy to identify, presumably because they often contain explicit derogatory language resembling overt hate speech. By contrast, Israel-related and post-Holocaust antisemitism constitute the most difficult content groups. In the BASE setting, Israel-related antisemitism typically exhibits the lowest recall, whereas post-Holocaust antisemitism remains difficult across all configurations.

Israel-related antisemitism exhibits the strongest dependence on prompting configuration. While recall is consistently low in BASE, STRUCT substantially improves performance across all models, at least doubling recall relative to BASE. STRUCT+EX yields similar or slightly smaller gains for most models, except for GPT, where explicit examples further increase recall.

The impact of IHRA is considerably less consistent. For GPT, IHRA preserves or decreases recall across all content groups. Claude and LLaMA generally benefit from IHRA, whereas Gemini exhibits mixed effects depending on the content group. Overall, STRUCT and STRUCT+EX produce the most robust improvements across models and forms of antisemitism. Most gains are statistically significant and frequently associated with large effect sizes.

The previously observed difference in overall dataset difficulty is also reflected at the content-group level. Recall on Decoding remains lower than on Bloomington in most comparisons, particularly in BASE. At the same time, structured conceptual representations improve performance substantially more on Decoding. Under STRUCT and STRUCT+EX, several models largely close the gap between datasets for specific content groups; in some cases, Decoding even matches or exceeds Bloomington, particularly for Israel-related antisemitism.

Finally, external conceptual representations reduce performance disparities between content groups. In most cases, the standard deviation of content-group recalls decreases under STRUCT or STRUCT+EX, indicating more balanced detection performance across forms of antisemitism. These gains are primarily driven by strong improvements for previously underperforming categories, especially Israel-related antisemitism. Gemini exhibits the lowest overall variance across content groups.

4.3 Grounded Explanations and Interpretability

We next analyze how models reference conceptual categories in their explanations. For IHRA, we extract references to *IHRA* sections; for STRUCT and STRUCT+EX, we extract references to *Lexicon* chapters and evaluate the resulting predictions as a multi-label classification task.

4.3.1 Number of References in Explanations

Across all models and datasets, generated explanations contain substantially more conceptual refer-

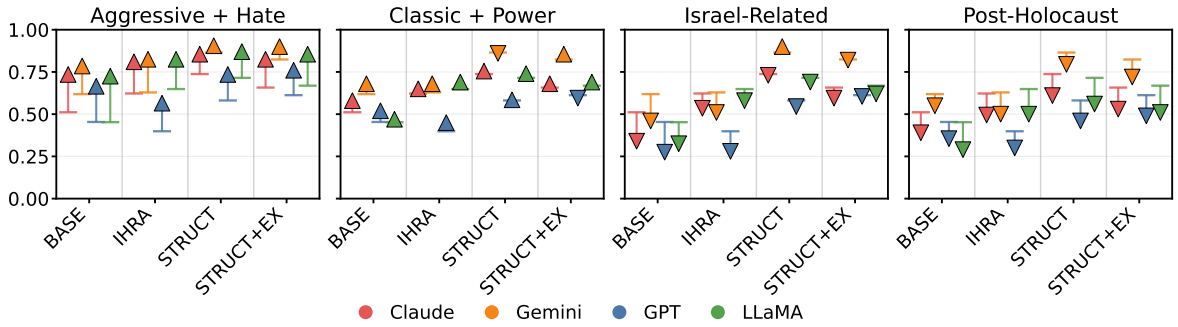

Figure 2: Recall for each content group. Arrows indicate gain/loss relative to average recall.

ences than human annotations. This effect is particularly informative for Decoding, where annotators could assign an unrestricted number of labels, yielding approximately 1.5 labels per antisemitic post on average. In comparison, STRUCT produces between 2.5 and 3.5 referenced chapters, roughly twice as many as human annotators. The effect is substantially weaker for IHRA, where explanations typically contain between 1.4 and 2.1 referenced sections per post. The number of generated references therefore scales strongly with the granularity of the conceptual representation, suggesting that fine-grained taxonomies encourage models to decompose antisemitic content into a larger number of simultaneously activated concepts.

At the same time, STRUCT+EX consistently reduces the number of generated references relative to STRUCT across all models and datasets, indicating that explicit examples encourage more selective category assignment. Model-specific differences are also pronounced: Claude produces the largest number of references, followed by GPT and LLaMA, whereas Gemini remains substantially more selective (Table 6 in Appendix A.5).

4.3.2 Distribution of Referenced Concepts

Figure 3 shows the deviation between the distribution of concepts referenced in model explanations and the corresponding distribution in human annotations.

Across models and configurations, aggressive speech acts are consistently over-represented in model explanations, whereas Israel-related concepts are systematically under-referenced. These tendencies closely mirror the recall results from the previous section: aggressive antisemitic content was detected comparatively reliably across models, while Israel-related antisemitism exhibited substantially lower recall, especially under BASE and IHRA. Post-Holocaust antisemitism exhibits a contrasting pattern. Despite comparatively low recall across models, post-Holocaust concepts are frequently over-referenced under STRUCT and, to a lesser extent, STRUCT+EX.

4.3.3 Classification Performance via References

Across all models, configurations and datasets, Israel-related antisemitism is classified most accurately, with F_1 scores between 0.79 and 0.94. This is notable given that Israel-related content was substantially more difficult to detect as antisemitic, especially under BASE and IHRA. By contrast, aggressive and hateful speech acts, while comparatively easy to detect as antisemitic, are classified considerably less precisely, with F_1 scores between 0.20 and 0.56. Post-Holocaust antisemitism remains the most challenging category overall, both for detection and reference-based classification, with consistently low performance across models and configurations. Detailed precision, recall and F_1 scores by content group, dataset, model and configuration are reported in Table 7 in Appendix A.5.

These results should be interpreted in light of the systematic overproduction of references: precision is systematically reduced because models cite substantially more sections or chapters per post than human annotators.

4.4 Large-Context Provision

So far, our results indicate that compact fine-grained conceptual representations substantially improve antisemitism detection and classification. This raises the question whether providing a substantially richer resource comprising theoretical background, historical context, and extensive examples would further improve model behavior. To investigate this, we leverage the large-context capa-

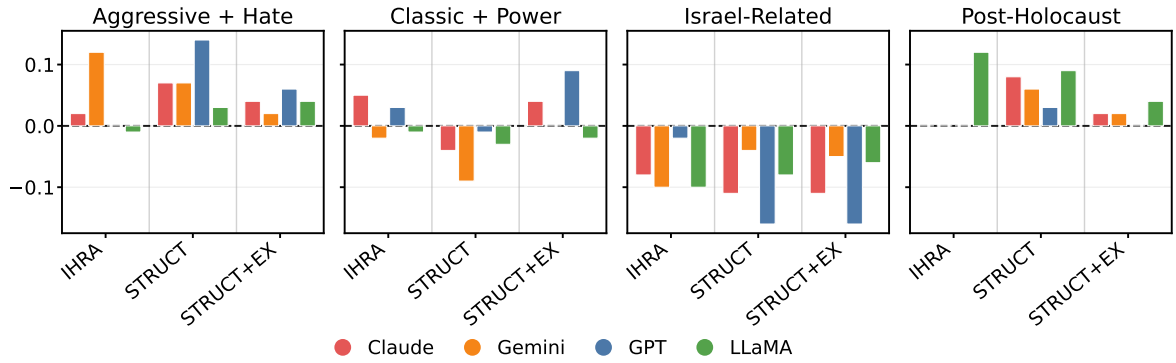

Figure 3: Deviation of model explanation distributions from human annotator distributions across content groups and configurations on the balanced dataset union.

bilities of Gemini by providing the complete *Lexicon* instead of the compact taxonomy alone.

Contrary to expectations, the *Lexicon* configuration does not improve quantitative performance.[3] Detection and classification metrics remain comparable to STRUCT and STRUCT+EX despite the substantially larger amount of contextual information available to the model.

4.4.1 Qualitative Error Analysis and Manual Explanation Inspection

We applied inductive category formation following Mayring (2014) to analyze all model errors, that is explanations for all false positives and negatives. We additionally inspected 450 generated explanations of correct predictions.

Many false positives involve posts that quote, criticize, or sarcastically reference antisemitic statements. In several cases, the model even correctly identifies sarcasm or condemnation within its explanation. Explanations of false positives also frequently indicate strong reliance on salient lexical cues such as *kikes* or *Holocaust*. Another group of false positives is better explained by differences between annotation frameworks than by model failure. For example, Holocaust distortion is explicitly operationalized in the *Lexicon* but not in the *IHRA*, which serves as the annotation basis for Bloomington. We further observe cases of model overconfidence, as the respective post showed a certain level of missing context. In some cases, the model even explicitly acknowledged respective uncertainty in the explanation.

False negatives predominantly involve subtle, indirect, or justificatory forms of antisemitism. These include Holocaust denial framed through legalistic reasoning (e.g., arguing that deaths due to disease in concentration camps do not constitute murder due to alleged lack of intent), indirect endorsements of violence against the Jewish state, symbolic delegitimization such as quotation marks around "Israel," and stereotypical allegations against Jews framed as accusations against Israel as a Jewish collective. Other missed cases include posts in which antisemitic narratives are framed as personal or political opinion. When posts rely on misinformation, historical distortions, or conspiratorial assumptions, explanations frequently accept these premises at face value rather than contextualizing or challenging them. This is less pronounced for explicit Holocaust denial, but remains visible for more subtle or historically less established antisemitic narratives.

Overall, the model generally identifies relevant antisemitic concepts and often handles implicit linguistic strategies such as intentional misspellings (e.g., "Juice" for "Jews") or dog whistles. However, when multiple tropes occur simultaneously, explanations typically focus on the most salient cue. In the presence of explicit slurs, explanations concentrate almost exclusively on lexical markers even when other antisemitic tropes are more central to the post. References to multiple chapters often reflect elaborations of a single detected pattern rather than recognition of conceptually distinct tropes.

5 Discussion and Future Work

Our results demonstrate that external conceptual representations substantially improve antisemitism detection and classification, confirming (Becker et al., 2024a, p. 56) and contrasting prior findings on hate speech (Melis et al., 2025; Roy et al., 2023). While *IHRA* generally improves perfor-

[3] Results for LEXICON are reported in Appendix A.5.

mance, the gains remain comparatively modest, similar to Patel et al. (2025), suggesting that it primarily reinforces knowledge already encoded in the models rather than substantially extending their conceptual coverage. In contrast, the fine-grained representations consistently yield the largest recall improvements across datasets and models. Particularly striking is the fact that providing a compact list of antisemitic concepts without further elaboration often performs best. Even supplying the complete *Lexicon* does not improve performance. Together, these findings suggest that LLMs benefit more from **explicit conceptual activation cues** than from extensive definitional elaboration.

At the same time, the improved recall using a fine-grained taxonomy comes at the cost of reduced precision. Neither curated examples nor the substantially richer *Lexicon* mitigated this **precision-recall trade-off**. This indicates that additional contextual detail alone is insufficient to improve conceptual selectivity once the relevant concepts have already been activated.

Our findings further reveal systematic differences between forms of antisemitism. Across all models and prompting configurations, aggressive speech is detected most reliably, aligning with prior observations that explicit hate is easier for LLMs to recognize (Ocampo et al., 2023; ElSherief et al., 2021). By contrast, Israel-related and post-Holocaust antisemitism are substantially more challenging, confirming existing research (Anti-Defamation League, 2025). However, the recall for **Israel-related antisemitism** doubles for all models, with the fine-granular taxonomy, while the improvement is modest with *IHRA*. This is interesting given ongoing objections that *IHRA* leads to a general interpretation of criticism of Israel as antisemitic. At the same time, Israel-related antisemitism is the most correctly classified content group, across models and configurations, once detected. A plausible explanation is the presence of lexical cues such as *Israel*.

Post-Holocaust antisemitism remains consistently difficult both for detection and classification. At the same time, the fine-grained taxonomy substantially increases references to post-Holocaust concepts in explanations despite comparatively low recall for the corresponding posts. The overly broad conceptual activation is constrained when explicit examples are provided. These findings point toward difficulties in calibrating when historically contextualized narratives should be interpreted as antisemitic. Interestingly, LLaMA exhibits similar overproduction even under *IHRA*, potentially reflecting model-specific differences in how antisemitism concepts are internally structured. One possible explanation is that several forms of post-Holocaust antisemitism, such as reversal of victims and perpetrators or resentment toward Holocaust remembrance, are particularly rooted in German post-war discourse and may therefore be less strongly represented in predominantly English-language training data. Future work should investigate these patterns more directly, including cross-lingual analyses and targeted evaluation on historically contextualized antisemitic narratives.

A qualitative analysis of Lexicon-prompted Gemini model errors confirmed previous findings on prompting-based detection and classification of hate speech, discussed in Related Work. We observed ongoing difficulties with **sarcasm, quotation, and condemnation** despite explicit prompt instructions to account for speaker intent. In line with (Tang et al., 2023), we additionally observed **laziness and reliance on lexical cues**, suggesting surface-level conceptual matching rather than robust contextual reasoning. When multiple antisemitic patterns co-occur, explanations typically elaborate only the most salient or easily detectable one. Another type of error is **model overconfidence**, reflected in the underuse of the abstention option. Across most model and configuration settings, the abstention label (*Unsure*) was selected in only a small percentage of cases, consistent with prior work (Madhusudhan et al., 2025; Wen et al., 2024). At the same time, our qualitative analysis of the explanations revealed that models often expressed uncertainty while still producing a binary classification. The variation in abstention behavior across specific conditions and datasets, as well as the discrepancy between explicit abstention and uncertainty expressed in explanations, represents an important direction for future research.

Generally Gemini explanations with Lexicon showed plausible reasoning, allowing indirect classification of antisemitic forms and providing insight into which concepts models attend to. However, we observe ignorance of antisemitism in posts relying on **misleading or historically distorted premises**. In such cases, explanations often accept problematic assumptions at face value rather than challenging them. This reveals a broader limitation of LLMs for social science research: the absence of contextual correction or epistemic stance-taking.

6 Limitations

To enable comparability across datasets, we mapped Decoding Antisemitism annotation codes to both *Lexicon* chapters and *IHRA* sections. While this mapping was conducted carefully, some ambiguity persists at the level of individual sections. Also, a few Decoding codes fell under two *IHRA* sections simultaneously, which meant an inflation of the number of referenced codes. Aggregating at the level of content groups mitigates this issue but necessarily reduces analytical granularity. Consequently, our analyses operate at the content-group level rather than the level of individual *Lexicon* chapters or *IHRA* sections, limiting insights into finer-grained distinctions between specific antisemitic tropes.

A second limitation concerns the operationalization of external conceptual resources. Questions of how fine-grained and up-to-date conceptual knowledge can be maintained and integrated efficiently remain largely unresolved. Possible directions include retrieval-augmented generation (RAG) and more structured representations such as knowledge graphs (Yang et al., 2025). However, such approaches remain underdeveloped for antisemitism research, with rather small Wikipedia-based knowledge bases utilized so far (Halevy, 2023). Future work should therefore investigate more systematic forms of structured conceptual grounding and their interaction with contextual reasoning.

Another limitation concerns the provenance of the *Lexicon* used in the additional large-context experiment (Section 4.4). Many illustrative examples originate from the Decoding dataset used in our experiments, potentially advantaging Lexicon-based prompting. While these examples have also appeared in published reports and may therefore form part of the models' broader pretraining knowledge, this overlap cannot be fully disentangled. Importantly, however, Lexicon-based prompting also achieves strong recall on the Bloomington dataset, suggesting that its benefits are not confined to the source corpus. We additionally observe no systematic overrepresentation of early or late *Lexicon* chapters in generated references, indicating no obvious positional attention bias in the long-context setting. However, because the table of contents was part of the provided *Lexicon*, future work should investigate attention allocation in long conceptual documents more directly.

Finally, while evaluating four models permits initial assessment of which findings generalize across systems and which remain model-specific, broader model coverage would naturally yield richer comparative insights. Our selection of four predominantly commercial models reflects a deliberate balance between contrastive breadth and interpretive manageability. Extending the analysis to more diverse architectures, scales, and open-weight systems remains an important direction for future work.

References

Theodor W. Adorno. 1950. *Prejudice in the Interview Material*, pages 605–653. The Authoritarian Personality. Harper & Row, New York.

Theodor W. Adorno. 1999. *Minima Moralia: Reflections from Damaged Life*. Verso, London.

Siddhant Agarwal, Adya Dhuler, Polly Ruhnke, Melvin Speisman, Md Shad Akhtar, and Shweta Yadav. 2025. MAMA-Memeia! Multi-Aspect Multi-Agent Collaboration for Depressive Symptoms Identification in Memes. *arXiv preprint*. ArXiv:2512.25015 [cs].

Anthropic. 2026. Claude Sonnet 4.6 system card. Technical report, Anthropic.

Anti-Defamation League. 2025. Generating hate: Anti-Jewish and anti-Israel bias in leading large language models. Technical report, ADL Center for Technology and Society.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2025. What's in a prompt?: A large-scale experiment to assess the impact of prompt design on the compliance and accuracy of llm-generated text annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):122–145.

Vidhisha Balachandran, Jingya Chen, Neel Joshi, Besmira Nushi, Hamid Palangi, Eduardo Salinas, Vibhav Vineet, James Woffinden-Luey, and Safoora Yousefi. 2024. Eureka: Evaluating and understanding large foundation models. *arXiv preprint*. ArXiv:2409.10566 [cs].

Matthias J. Becker, Laura Ascone, Matthew Bolton, Veronika Bundzíková, Alexis Chapelan, Pia Haupeltshofer, Alexa Krugel, Iael Kurjan, Helena Mihaljević, Stefan Munnes, Karolina Placzynta, Milena Pustet, Mohamed Salhi, Markus Scheiber, and Victor Tschiskale. 2024a. Decoding antisemitism: An AI-driven study on hate speech and imagery online. discourse report 6. TU Berlin.

Matthias J. Becker, Hagen Troschke, Matthew Bolton, and Alexis Chaeplan. 2024b. *Decoding Antisemitism: A Guide to Identifying Antisemitism Online*. Springer Nature Switzerland.

Youngjin Chae and Thomas Davidson. 2026. Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*, 55(2):501–567.

Milo Comerford and Lea Gerster. 2021. The rise of antisemitism online during the pandemic: A study of French and German content. Report, Institute for Strategic Dialogue (ISD), Luxembourg. Commissioned by the European Commission, Directorate-General Justice and Consumers.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara Mcgillivray. 2024. Is Explanation All You Need? An Expert Survey on LLM-generated Explanations for Abusive Language Detection. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy. CEUR Workshop Proceedings.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Finkelstein, Pamela Paresky, Alex Goldenberg, Savvas Zannettou, Lee Jussim, Denver Riggleman, John Farmer, Paul Goldenberg, Jack Donohue, and Malav H. Modi. 2020. Antisemitic Disinformation: A Study of the Online Dissemination of Anti-Jewish Conspiracy Theories. Technical report, The Network Contagion Research Institute.

Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. 2025. Can prompting LLMs unlock hate speech detection across languages? a zero-shot and few-shot study. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 413–425, Vienna, Austria. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint*. ArXiv:2507.06261 [cs].

Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. 2025. MoMoE: Mixture of moderation experts framework for AI-assisted online governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12645–12660, Suzhou, China. ACL.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An Investigation of Large Language Models for Real-World Hate Speech Detection. *arXiv preprint*. ArXiv:2401.03346 [cs].

Karina Halevy. 2023. A group-specific approach to nlp for hate speech detection. *arXiv preprint*. ArXiv:2304.11223 [cs].

David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in moderation: How commercial content moderation apis over- and under-moderate group-targeted hate speech and linguistic variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. ACM.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, page 294–297. ACM.

International Holocaust Remembrance Alliance. 2016. The working definition of antisemitism. `https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antisemitism`. Accessed: 2026-01-05.

Gunther Jikeli, Deepika Awasthi, David Axelrod, Daniel Miehling, Pauravi Wagh, and Weejoeng Joeng. 2021. Detecting anti-jewish messages on social media. building an annotated corpus that can serve as a preliminary gold standard. In *Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media*.

Gunther Jikeli, David Axelrod, Rhonda K Fischer, Elham Forouzesh, Weejeong Jeong, Daniel Miehling, and Katharina Soemer. 2024. Differences between antisemitic and non-antisemitic english language tweets. *Computational and mathematical organization theory*, 30(3):232–266.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. ACM.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:865–878.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Renkai Ma and Yubo Kou. 2023. "defaulting to boilerplate answers, they didn't engage in a genuine conversation": Dimensions of transparency design in creator moderation. In *Proceedings of the ACM on Human-Computer Interaction*, volume 7, New York, NY, USA. ACM.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345, Abu Dhabi, UAE. Association for Computational Linguistics.

Kenneth L. Marcus. 2015. *The Definition of Anti-Semitism*. Oxford University Press, Oxford.

Philipp Mayring. 2014. *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*. Klagenfurt.

Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521, Vienna, Austria. Association for Computational Linguistics.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. ACL.

Helena Mihaljević and Elisabeth Steffen. 2022. How toxic is antisemitism? potentials and limitations of automated toxicity scoring for antisemitic online content. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022).*, pages 1–12.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2026. GPT-5.4 thinking system card. Technical report, OpenAI.

Jay Patel, Hrudayangam Mehta, and Jeremy Blackburn. 2025. Evaluating large language models for detecting antisemitism. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35356–35385, Suzhou, China. ACL.

Dina Porat. 2020. The working definition of antisemitism. a 2018 perception. In Armin Lange, Kerstin Mayerhofer, Dina Porat, and Lawrence H. Schiffman, editors, *Comprehending and Confronting Antisemitism. A Multi-Faceted Approach*, volume 1 of *An End to Antisemitism*, pages 475–488. De Gruyter, Berlin/Boston.

Milena Pustet and Helena Mihaljević. 2024. Automated detection of antisemitic texts: is context all we need? In Matthias J. Becker, Laura Ascone, Matthew Bolton, Veronika Bundzíková, Alexis Chapelan, Pia Haupeltshofer, Alexa Krugel, Iael Kurjan, Helena Mihaljević, Stefan Munnes, Karolina Placzynta, Milena Pustet, Mohamed Salhi, Markus Scheiber, and Victor Tschiskale, editors, *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online. Discourse Report 6*, pages 52–57.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the ACL: EMNLP 2023*, pages 6116–6128, Singapore. ACL.

Monika Schwarz-Friesel. 2019. *Judenhass im Internet: Antisemitismus als kulturelle Konstante und kollektives Gefühl*. Hentrich und Hentrich, Leipzig.

Monika Schwarz-Friesel. 2020. "Antisemitism 2.0" — The Spreading of Jew-hatred on the World Wide Web. In *Comprehending and Confronting Antisemitism. A Multi-Faceted Approach*, volume 1 of *An End to Antisemitism*, pages 311–338. De Gruyter, Berlin/Boston.

Elisabeth Steffen, Helena Mihaljević, Milena Pustet, Nyco Bischoff, Maria Do Mar Castro Varela, Yener Bayramoglu, and Bahar Oghalai. 2023. Codes, patterns and shapes of contemporary online antisemitism and conspiracy narratives – an annotation guide and labeled german-language dataset in the context of COVID-19. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1082–1092.

Elisabeth Steffen, Milena Pustet, and Helena Mihaljević. 2024. *8. Algorithms Against Antisemitism?: Towards The Automated Detection of Antisemitic Content Online*, page 205–236. Antisemitism in Online Communication. Open Book Publishers.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. Association for Computational Linguistics.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024. Characterizing LLM abstention behavior in science QA with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3437–3450, Miami, Florida, USA. Association for Computational Linguistics.

Robert S. Wistrich. 1994. *Antisemitism: The Longest Hatred*. Schocken Books, New York.

Wenli Yang, Lilian Some, Michael Bain, and Byeong Kang. 2025. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 318:113503.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal Understanding of Memes with Fair Explanations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2007–2017, Seattle, WA, USA. IEEE.

A Appendix

A.1 Antisemitism

Adorno describes antisemitism as "the rumor about the Jews" (Adorno, 1999), stressing that antisemitism is not explainable by what Jews do, but what the antisemite believes them to do. Reflecting its persistence as a "cultural constant" (Schwarz-Friesel, 2019) over more than 2,000 years and its capacity to adapt to changing historical, social, and political contexts, antisemitism is often referred to as "the longest hatred" (Wistrich, 1994). The ideology manifests through a heterogeneous set of stereotypes and narratives targeting Jews both as individuals and as a collective. Many of these stereotypes have deep historical roots, yet they are repeatedly rearticulated in new forms, particularly in times of social or political crisis.

A distinctive feature of antisemitism is its inherently contradictory structure. Jews are frequently portrayed simultaneously as inferior and superior, weak and powerful, victimized and manipulative. This paradoxical imagery enables antisemitic discourse to accommodate a wide range of conspiracy narratives, in which Jews are imagined as acting covertly or exercising hidden influence. As Porat notes, antisemitism is particularly difficult to grasp "since antipathy to Jews involves a deep-seated emotional dimension as well as a conglomerate of historic religious, political, and economic elements" (Porat, 2020, 476).

Historically, Christian anti-Judaism represents the oldest articulation of antisemitism (Wistrich, 1994). The term itself was made popular by Marr in the late 19th century, expressing the emergence of racial antisemitism which was no longer (primarily) religiously motivated, seeing Jews as a race rather than a religious community (Marcus, 2015). Under National Socialism, antisemitism reached its most extreme and industrialized form in the Holocaust. After 1945, antisemitic discourse did not disappear but instead adapted to new normative environments, giving rise to forms commonly described as post-Holocaust antisemitism. These include Holocaust denial and distortion, inversion of victim–perpetrator roles, and Israel-related antisemitism that targets the Jewish state as a Jewish collective (Marcus, 2015).

The longevity of antisemitism has resulted in a range of linguistic and narrative patterns. Contemporary antisemitism is not limited to explicit Jew hatred but is frequently articulated in encoded and implicit ways. Prior research shows that antisemitism, offline and online, can be expressed in a variety of coded forms, including dog whistles, allusions, metaphors, wordplay, and other semiotic markers (Schwarz-Friesel, 2020). At the same time, antisemitism may also be expressed through conspiracy narratives that attribute hidden power, malevolent intent, or global control to abstract entities such as 'elites' without mentioning Jews directly (Finkelstein et al., 2020; Comerford and Gerster, 2021).

Common forms of antisemitism are defined in the IHRA working definition of antisemitism (International Holocaust Remembrance Alliance, 2016), which consists of a general paragraph accompanied by eleven examples of common contemporary manifestations. The definition is the result of decade-long debates involving multiple stakeholders such as scholars and policy makers together with representatives of the Jewish community, and is nowadays adopted by the majority of Jewish organizations as well as 35 of the IHRA member states. The indeed interesting history of articulating the necessity for a definition and the process for formulating one can be found in Marcus (2015); Porat (2020).

A.2 Dataset details

While both datasets are grounded in the *IHRA* definition, they differ along several key dimensions: Bloomington relies on keyword-based sampling and a coarse-grained, predominantly single-label annotation scheme, whereas Decoding uses discourse-driven sampling and a fine-grained,

multi-label annotation scheme.

A.2.1 Bloomington

The Bloomington dataset[4] consists of English-language posts from Twitter (now X) (Jikeli et al., 2021) containing one of the keywords *jews*, *israel*, *kikes*[5], and *zionazi**. Posts were annotated by two field experts following *IHRA* in its twelve-section formulation. For posts labeled as antisemitic, annotators selected the most dominant *IHRA* section as a rationale. Disagreements regarding antisemitism labels were resolved through discussion, whereas disagreements on section assignment were retained, allowing antisemitic posts to be associated with at most two section labels (Jikeli et al., 2024).

The dataset contains 11,311 posts, of which 17.3% are labeled as antisemitic. Prevalence varies substantially across keyword-based subsamples: posts containing *zionazi* are predominantly antisemitic ($88.3\%, n = 529$), followed by *kikes* ($34.3\%, n = 283$); in contrast, posts containing *jews* ($16.1\%, n = 6{,}395$) or *israel* ($11.4\%, n = 4{,}134$) are mostly non-antisemitic.

A.2.2 Decoding

The Decoding dataset[6] comprises comment sections of a wide range of mainstream news outlets and social media platforms, with threads selected around public events and topics expected to trigger antisemitic discourse (Becker et al., 2024b, p. 544). For the present study, we use the English-language subset of the corpus, comprising 42,799 comments, most of which originate from Facebook (52.8%), followed by YouTube (21.1%), mainstream news outlets such as The Guardian and the BBC (18%), and Twitter (8.1%).

Considering *IHRA* to be insufficiently precise for systematic corpus analysis of contemporary antisemitism (Becker et al., 2024b, p. 11), the definition was extended and refined into an annotation scheme comprising 46 codes for antisemitic content. The resulting scheme specifies particular forms of post-Holocaust antisemitism such as the tabooing of criticism (Becker et al., 2024b, p. 309); differentiates stereotypes subsumed under broader *IHRA* categories, including lie and deceit (Becker et al., 2024b, p. 93) and vengefulness (Becker et al., 2024b, p. 107); and explicitly includes Israel-related narratives such as generalized apartheid analogies and calls for boycott, divestment, and sanctions (BDS). In addition, distinctions between antisemitism targeting Jewish individuals and the Jewish state are reorganized compared to *IHRA*. The annotators were allowed to assign an unlimited number of these codes to antisemitic comments.

Overall, 7.9% of comments are labeled as antisemitic, substantially fewer than in the Bloomington dataset. This difference likely reflects the predominance of moderated mainstream platforms as well as the discourse-driven sampling strategy, which does not rely on explicit keywords related to Jewish life or Israel.

A.2.3 Annotation Frameworks and Joint Representation

To enable joint analysis, we developed a mapping that aligns the fine-grained codes of Decoding with both *Lexicon* chapters and *IHRA* sections. When a single Decoding code corresponds to multiple *IHRA* sections, all relevant sections are retained; such instances therefore contribute to multiple *IHRA* categories.

We further abstract both annotation schemes into a shared set of higher-level *content groups* that capture recurring dimensions of antisemitic discourse and enable robust comparison across datasets and knowledge bases. Table 3 summarizes the four content groups, together with their definitions and the corresponding *IHRA* sections and *Lexicon* chapters.

While Israel-related antisemitism can be understood historically as a subset of post-Holocaust antisemitism (see Appendix A.1), we treat it as a separate content group for analytical reasons: In both knowledge bases, Israel-related antisemitism is elaborated in multiple sections or chapters; empirically it constitutes a substantial proportion of antisemitic content in both datasets.

A.3 Preprocessing

Both datasets were preprocessed using the same pipeline. We removed URLs, removed explicit markers indicating reply structure and stripped HTML markup while preserving the underlying textual content. User handles occurring at the beginning of a post were removed, while those appear-

[4]The binary labeled dataset is available on Zenodo: https://zenodo.org/records/14448399; CC BY 4.0 license. The section labels were made available upon request.

[5]The slur *kikes* has a long history in antisemitic discourse. In the *Studies in Prejudice*, Adorno et al. describe its use in constructing "the standard division of Jews into two groups, the good ones and the bad ones, a division frequently expressed in terms of the 'white' Jews and the 'kikes'" (Adorno, 1950).

[6]The dataset has been provided by the PI of the Decoding Antisemitism research project upon request.

Table 3: Joint content groups used for analysis, including definitions and their alignment with knowledge bases.

Content group	Description	Lexicon chapters	IHRA sections
Aggressive+Hate	Explicit expressions of hostility toward Jews, including hate speech, aggressive language, and calls for violence.	37–42	0, 1
Classic+Power	Stereotypical allegations against Jews, particularly narratives of power, manipulation, conspiracy, or malevolent influence.	2–15	2, 3, 6
Israel-Related	Antisemitic expressions targeting Israel as the Jewish state, including collective attributions, delegitimization, and analogies with Nazism.	28–36	7, 8, 9, 10
Post-Holocaust	Forms of antisemitism emerging after 1945, including denial, relativization, inversion, and secondary antisemitic narratives.	16–27	4, 5, 11

ing within the text were retained, as they may function as meaningful references to individuals or organizations (e.g., @netanyahu, @WashingtonPost). In almost all cases, models produced grounded explanations for a *Yes* classification, meaning they provided section and chapter references. Only for three posts, LLaMA did not mention a section number, but referenced the section by quoting it. We manually inspected these explanations and recoded them to the section number the model was quoting.

A.4 Prompts

The modular prompt template shared by all conditions, is shown in the box below. The `system_prompt`, shared task instructions, and JSON output format are identical across all conditions.

Knowledge component variants:
NO KB: None;
IHRA: IHRA Working Definition sections 0–11;
STRUCT: taxonomy labels 2–42;
STRUCT+EX: taxonomy labels 2–42 + examples;
LEXICON: Decoding Antisemitism Lexicon (full document as PDF).

system_prompt

You are an expert in classifying social media posts with regard to antisemitic content.

Shared task instructions

Decide whether the post below contains antisemitic speech. Thereby consider the **intent** of the message. In particular, **quoting or reporting** antisemitism for the purpose of criticism or condemnation should NOT be labeled as antisemitism. Answer preferably using 'Yes' or 'No'. If the decision is not possible, you can use 'Unsure'.

Output format (JSON conditions)

```
{
  "index": <original index of the text>,
  "classification": "<Yes | No | Unsure>",
  "explanation": "<brief explanation with
  reference to relevant taxonomy categories
  by number and title>"
}
```

Only output a JSON array of objects, no extra text.

A.5 Analysis details

Table 4: Model performance across all models on Bloomington (B) and Decoding (D) and their balanced union (U). Best performing configuration per model, metric and dataset is highlighted in bold. Significance is tested with BASE as reference group. Statistically significant differences are indicated by ↑ or ↓; large effects (Cohen's $|h| \geq 0.5$) are marked with *.

Model	**Cfg**	**Precision**			**Recall**			**F1**		
		B	D	U	B	D	U	B	D	U
Claude	BASE	**.65**	**.67**	**.66**	.55	.33	.44	.60	.44	.52
	IHRA	.59 ↓	.63	.61	.68 ↑	.48 ↑	.58 ↑	.63 ↑	.54 ↑	.59
	STRUCT	.56 ↓	.54 ↓	.55	**.79** ↑*	**.67** ↑*	**.73** ↑*	**.65** ↑	**.60** ↑	**.63** ↑
	STRUCT+EX	.58 ↓	.53 ↓	.56	.71 ↑	.54 ↑	.63 ↑	.64 ↑	.54 ↑	.59
Gemini	BASE	.57	.52	.54	.64	.46	.55	.60	.49	.54
	IHRA	**.61** ↑	**.55**	**.58**	.68 ↑	.48 ↑	.58 ↑	**.64** ↑	.51	.58
	STRUCT	.49 ↓	.45	.47	**.88** ↑*	**.87** ↑*	**.87** ↑*	.63	**.59** ↑	**.61**
	STRUCT+EX	.49 ↓	.40 ↓	.44 ↓	.83 ↑	.80 ↑*	.82 ↑*	.61	.53	.57
	LEXICON	.47 ↓	.45	.46	.84 ↑	.85 ↑*	.85 ↑*	.60	**.59** ↑	.59
GPT	BASE	.57	.62	.60	.51	.26	.38	.54	.36	.45
	IHRA	**.74** ↑	**.84**	**.79**	.49	.21 ↓	.35	.59 ↑	.34	.46
	STRUCT	.70 ↑	.78	.74	**.64** ↑	.48 ↑	.56 ↑	**.67** ↑	.59 ↑	.63 ↑
	STRUCT+EX	.68 ↑	.70	.69	**.64** ↑	**.55** ↑*	**.60** ↑	.66 ↑	**.61** ↑*	**.64** ↑
LLaMA	BASE	**.66**	**.48**	**.57**	.48	.29	.38	.56	.36	.46
	IHRA	.51 ↓	.44	.48	.67 ↑	.55 ↑*	.61 ↑	**.58**	**.49** ↑	**.54**
	STRUCT	.48 ↓	.33 ↓	.41 ↓	**.73** ↑*	**.66** ↑*	**.70** ↑*	**.58**	.44 ↑	.51
	STRUCT+EX	.50 ↓	.41	.45	.68 ↑	.59 ↑*	.64 ↑*	.57	.48 ↑	.53

Table 5: Recall of antisemitism detection, disaggregated by model, configuration, content group and dataset (B = Bloomington, D = Decoding, U = balanced union). Significance is tested with BASE as reference group. Statistically significant differences are indicated by ↑ or ↓; large effects (Cohen's $|h| \geq 0.5$) are marked with *.

Model	**Cfg**	**Aggressive+Hate**			**Classic+Power**			**Israel-related**			**Post-Holocaust**		
		B	D	U	B	D	U	B	D	U	B	D	U
Claude	BASE	.82	.65	.74	.78	.38	.58	.41	.27	.34	.49	.29	.39
	IHRA	.86	.76	.81	.83	.47 ↑	.65 ↑	.58 ↑	.49 ↑	.54 ↑	.62	.37 ↑	.50 ↑
	STRUCT	.87	.84	.86 ↑	.86 ↑	.65 ↑*	.76 ↑	.75 ↑*	.71 ↑*	.73 ↑*	.67	.55 ↑*	.61 ↑
	STRUCT+EX	.85	.80	.83	.83	.53 ↑	.68	.64 ↑	.55 ↑*	.60 ↑*	.59	.47 ↑	.53 ↑
Gemini	BASE	.82	.75	.79	.85	.51	.68	.51	.41	.46	.71	.39	.55
	IHRA	.87	.78	.83	.86	.50	.68	.56 ↑	.51 ↑	.51 ↑	.64	.36	.50
	STRUCT	.92	.89	.91 ↑	.90 ↑	.82 ↑*	.86 ↑	.87 ↑*	.93 ↑*	.90 ↑*	.83	.76 ↑*	.80 ↑*
	STRUCT+EX	.91	.89	.90 ↑	.91 ↑	.80 ↑*	.86 ↑	.79 ↑*	.85 ↑*	.82 ↑*	.79	.65 ↑*	.72 ↑
	LEXICON	.90	.88	.89 ↑	.89	.85 ↑*	.87 ↑	.79 ↑*	.87 ↑*	.85 ↑*	.79	.77 ↑*	.81 ↑*
GPT	BASE	.80	.53	.67	.73	.31	.52	.36	.19	.28	.47	.24	.36
	IHRA	.73	.40	.57	.67	.23 ↓	.45	.37	.19	.28	.45	.15 ↓	.30
	STRUCT	.81	.66	.74	.75	.42 ↑	.59 ↑	.56 ↑	.53 ↑*	.55 ↑*	.53	.39 ↑	.46 ↑
	STRUCT+EX	.78	.74	.76	.72	.47 ↑	.60	.60 ↑	.61 ↑*	.61 ↑*	.52	.46 ↑	.49
LLaMA	BASE	.74	.71	.73	.62	.32	.47	.38	.27	.33	.40	.18	.29
	IHRA	.86 ↑	.79	.83 ↑	.80 ↑	.58 ↑*	.69 ↑	.59 ↑	.57 ↑*	.58 ↑*	.64 ↑	.36 ↑	.50 ↑
	STRUCT	.87 ↑	.87	.87 ↑	.80 ↑	.68 ↑*	.74 ↑*	.69 ↑*	.69 ↑*	.69 ↑*	.66 ↑*	.46 ↑*	.56 ↑*
	STRUCT+EX	.87 ↑	.84	.86 ↑	.77 ↑	.61 ↑*	.69 ↑	.62 ↑	.62 ↑*	.62 ↑*	.62 ↑	.40 ↑	.51 ↑

Table 6: Average number of referenced sections (IHRA) or chapters (STRUCT, STRUCT_EX) per post by model vs. human annotators (*in parentheses*). Overall, annotators assigned 1.3 sections for B (= Bloomington) and and 1.48 chapters for D (= Decoding). Numbers in parentheses vary due to model and configuration variance in detection.

Model	**Cfg**	**B**	**D**
Claude	IHRA	2.00 *(1.32)*	2.12 *(1.60)*
	STRUCT	3.30 *(1.33)*	3.50 *(1.58)*
	STRUCT+EX	3.03 *(1.33)*	3.22 *(1.60)*
Gemini	IHRA	1.65 *(1.32)*	1.62 *(1.60)*
	STRUCT	2.52 *(1.32)*	2.60 *(1.51)*
	STRUCT+EX	2.10 *(1.32)*	2.20 *(1.54)*
	LEXICON	2.35 *(1.32)*	2.62 *(1.53)*
GPT	IHRA	1.41 *(1.33)*	1.43 *(1.66)*
	STRUCT	3.14 *(1.35)*	3.07 *(1.58)*
	STRUCT+EX	2.81 *(1.34)*	2.73 *(1.61)*
LLaMA	IHRA	1.75 *(1.32)*	2.07 *(1.52)*
	STRUCT	2.99 *(1.34)*	3.24 *(1.50)*
	STRUCT+EX	2.49 *(1.36)*	2.62 *(1.52)*

Table 7: Distribution of referenced content groups in model explanations compared to human annotations. Rows correspond to precision, recall, and F1.

Model	**Cfg**	**Aggressive +Hate**		**Classic +Power**		**Israel -related**		**Post -Holocaust**	
		B	D	B	D	B	D	B	D
Claude	IHRA	.38	.19	.52	.65	.89	.77	.36	.32
		.44	.74	.94	.84	.93	.93	.62	.39
		.41	.30	.67	.73	.91	.84	.46	.35
	STRUCT	.35	.18	.52	.67	.91	.76	.16	.35
		.76	.87	.91	.72	.93	.97	.80	.78
		.48	.30	.66	.69	.92	.85	.27	.48
	STRUCT+EX	.38	.27	.50	.65	.93	.82	.22	.55
		.64	.84	.93	.90	.91	.95	.70	.74
		.48	.41	.65	.75	.92	.88	.33	.63
Gemini	IHRA	.29	.12	.62	.71	.93	.82	.38	.24
		.53	.67	.86	.70	.87	.79	.52	.27
		.37	.20	.72	.70	.90	.80	.44	.25
	STRUCT	.34	.15	.60	.73	.95	.74	.21	.44
		.64	.68	.80	.52	.93	.96	.78	.66
		.44	.25	.69	.61	.94	.84	.33	.53
	STRUCT+EX	.47	.23	.56	.74	.95	.80	.29	.59
		.49	.63	.81	.74	.86	.92	.69	.63
		.48	.34	.66	.74	.90	.86	.41	.61
	LEXICON	.47	.25	.48	.67	.93	.79	.22	.57
		.44	.84	.85	.85	.87	.93	.63	.74
		.45	.39	.61	.75	.90	.85	.33	.64
GPT	IHRA	.64	.35	.70	.77	.94	.83	.44	.35
		.50	.64	.82	.75	.90	.79	.44	.34
		.56	.45	.76	.76	.92	.81	.44	.34
	STRUCT	.29	.12	.51	.51	.96	.83	.21	.45
		.86	.86	.85	.70	.91	.92	.75	.73
		.43	.21	.64	.59	.93	.87	.33	.56
	STRUCT+EX	.35	.22	.43	.50	.97	.88	.23	.63
		.76	.79	.91	.85	.89	.91	.64	.72
		.48	.34	.58	.63	.93	.89	.34	.67
LLaMA	IHRA	.51	.25	.55	.62	.92	.74	.18	.14
		.35	.51	.84	.71	.89	.85	.60	.46
		.42	.34	.66	.66	.90	.79	.28	.21
	STRUCT	.33	.17	.51	.64	.92	.70	.14	.31
		.35	.64	.89	.68	.91	.95	.74	.64
		.34	.27	.65	.66	.91	.81	.24	.42
	STRUCT+EX	.39	.20	.54	.68	.93	.73	.17	.42
		.45	.68	.85	.68	.86	.94	.58	.61
		.42	.31	.66	.68	.89	.82	.26	.50