Stress-Testing LLMs for Hateful Event Detection: A Case Study on Antisemitism

Karina Halevy,[1] **Julia Mendelsohn,**[2] **Chan Young Park,**[1] **Naomi Younger,**[3]
Tammi Rossman-Benjamin,[3] **Yulia Tsvetkov,**[4], **Maarten Sap**[1]
[1]Carnegie Mellon University,
[2]University of Michigan, [3]AMCHA Initiative, [4]University of Washington
khalevy [at] andrew [dot] cmu [dot] edu

Abstract

Surfacing incidents of hate and violence from news is essential for understanding and addressing their broader societal impacts. Large language models (LLMs) can help elucidate these harms, but the task of identifying harmful events—distinct from detecting direct hate speech—remains unexplored. This paper explores the capability of LLMs for discovery and fine-grained classification of reports of hateful events, focusing on antisemitism. We stress-test fine-grained hateful event classification on two datasets with expert-labeled instances of various categories of antisemitism. LLMs are far from perfect at understanding antisemitic events, with baseline F1 as low as 38.76% on the best LLM. However, providing precise definitions from our taxonomy and in-context examples steers GPT-4o and Llama-3.2 to perform slightly better on tagging antisemitic event descriptions (+4–5% wF1). We also test our classifier's generalizability on recent news reports:despite low fine-grained performance, LLMs can help surface some events. In sum, we introduce the novel task of hateful event detection, revealing important gaps in LLMs' reasoning capabilities and their ability to discern nuanced manifestations of harm. These findings have broad implications for NLP research and society, underscoring the need for better LLM alignment strategies and policy efforts to define hate precisely.

1 Introduction

Detecting reports of hate incidents from the news is crucial for monitoring societal trends (Pontiki et al., 2020) and harms toward marginalized communities.[1] To be useful for practitioners, labels should convey the types of hate reported, including information such as which group was targeted and what hateful stereotype was spread. This information creates a more fine-grained picture of hate trends (Figure 1), which could inform education campaigns or targeted interventions against hate.[2][3]

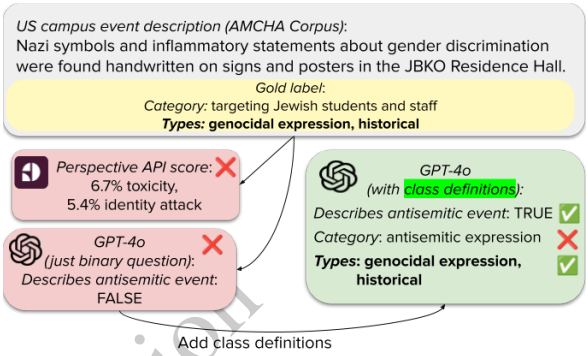

Figure 1: We introduce the hate event detection task. Given raw text reporting a hate event, the Perspective API (a hate speech detector) assigns low toxicity probabilities. GPT-4o, given only the description and its corresponding date/university, does not predict antisemitism described in the text either. However, the adding terms definitions helps GPT-4o predict the gold label types.

"Harm" is a subjective concept that annotators operationalize differently (Breitfeller et al., 2019; Sap et al., 2022; Alkomah and Ma, 2022; Kansok-Dusche et al., 2023; Yin and Zubiaga, 2021; Fleisig et al., 2023), especially because they can disagree when labeling coarse concepts such as "harmful," "toxic," or "antisemitic" given abstract, implicit, loaded, or contested translation guidelines (Kim et al., 2023; ElSherief et al., 2021; Pavlovic and Poesio, 2024; Richardson, 2021). Multi-level, fine-grained classification could make classifiers and their errors more explainable, and their corresponding gold labels would be less prone to such subjectivity-induced disagreement. Thus, we need finer-grained taxonomies to name concrete types of harm (Bibal et al., 2025).

This work stress-tests LLMs' ability to per-

[1]https://un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm

[2]https://www.fcas.org/

[3]https://www.noplaceforhate.org/

form nuanced classification for descriptions of antisemitic events. The case of antisemitism is fit for this investigation because of its frequently debated definitions and conflicting interpretations of what counts as antisemitic (Klug, 2023; Harrison and Klaff, 2021; Feldman and Volovici, 2023; Herf, 2021; Penslar, 2022; Nexus, 2023; Jerusalem, 2021). Studying antisemitism is also important due to increased hate crimes against Jewish people[4] as well as the general harmful consequences that online hate can have both online and offline (e.g., harassment, mental distress, hate crimes, Räsänen et al., 2016; UN, 2018; Byman, 2021).

To study coarse- and fine-grained classification of antisemitic events, we experiment with two datasets. First, we extract and release to the research community the AMCHA Corpus—a unique challenge set of textual descriptions of antisemitic events from U.S. university campuses, collected continuously by the AMCHA Initiative from 2015 to the present, annotated with 2 coarse-grained categories and 9 fine-grained types. We also consider the antisemitism-related subset of the Anti-Defamation League (ADL) Hate, Extremism, Antisemitism, and Terrorism (H.E.A.T. Map) dataset.[5]

We empirically evaluate two widely-used LLMs (GPT-4o and Llama-3.2-3B-Instruct) on fine-grained hateful event detection. We also assess the effects of adding in-context examples, modifying underlying assumptions, and extensively defining parts of taxonomies. Both models perform poorly on fine-grained classification when given only the event description; this is likely due to lack of knowledge of both historical and recent antisemitic events. Supporting this hypothesis, adding term definitions and in-context examples slightly helps classifiers disambiguate potentially unfamiliar action-oriented or historical knowledge-laden types of hate. We observe that GPT-4o is more steerable towards improved fine-grained classification, while Llama 3.2-3B-Instruct is more steerable toward improved coarse-grained classification.

We showcase the generalizability and usefulness of fine-grained classification models for detecting hateful events in real-time. We scrape the past two years of news articles from five U.S. universities with frequent antisemitic event reports and use GPT-4o to estimate prevalence of on-campus incidents of antisemitism based on AMCHA's taxonomy. Our generalized classifier can achieve high precision and is thus useful in generating a lower bound of antisemitic incident reports from news.

Overall, our findings suggest that *specificity* in annotation guidelines and label definitions is important for the *generalizability* of hateful event detection across datasets, domains, and groups. We further advocate for developing more precise, informative definitions and taxonomies of concepts like antisemitism and toxicity to enhance both computational and societal understanding.

2 Task Definition

We formally define the task of hateful event detection as follows: given an input event description extracted from a news article, along with the event date and location (collectively, input d), model M must classify the coarse-grained category c of the event, the set of n fine-grained types $t = t_1, \ldots, t_n$, and optionally, the binary hate label l. For humans performing this task, especially for real-time hateful event detection, this typically requires deep historical knowledge of the relevant types of hate, a strong understanding of the taxonomies and label space, and global context related to recent events.

The challenge with fine-grained taxonomy-based event classification is that labeling whether and which type of harmful event is described in text is non-trivial. It is distinct from detecting hate speech (a well-studied phenomenon; Jahan and Oussalah, 2023; Ramos et al., 2024). First, detecting an event involving hate speech, rather than the speech itself, requires additionally detecting evidential markers that characterize the text as an event report (e.g., phrases like "This person said [x]," where x is the direct hate speech). This language is typically found in domains such as news media that employ a more formal editorial style than user-run social media platforms. Second, some hateful actions (e.g., physical assault, posting or tearing down signage, boycotts) have no component of explicit speech, so a hate speech classifier may not be trained on data that discusses such events.

3 Background: Antisemitism

Antisemitism, an old yet constantly evolving, pervasive form of hatred or prejudice, often fails to fit neatly into modern notions of race, racism,

[4] https://www.fbi.gov/news/press-releases/fbi-releases-2022-crime-in-the-nation-statistics

[5] https://www.adl.org/resources/tools-to-track-hate/heat-map

religion, and religious discrimination.[6] According to the International Holocaust Remembrance Association (IHRA), antisemitism is "a certain perception of Jews, which may be expressed as hatred toward Jews." This perception includes "rhetorical and physical manifestations of antisemitism [that] are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities."[7]

There is an urgent need to address antisemitism, especially given recent surges in the U.S. since October 2023.[8] These incidents must be labeled and addressed in a fine-grained and dynamic yet largely automatic way—their sheer volume is too much for humans to process and be exposed to. However, antisemitism detection research has mostly focused on isolated rhetorical snippets rather than descriptions of events with larger contexts.

4 Datasets & Taxonomies

The AMCHA Corpus We scrape and release the AMCHA Corpus, a growing database of English-language entries [9] of contextualized descriptions of antisemitic events that have occurred on higher education campuses, annotated for coarse- and fine-grained categories of antisemitism. The dataset is collected by the AMCHA Initiative through a continuous monitoring, screening, and consensus-coding procedure done by experts (See Appendix C for more details and data statistics).

We use a version of the AMCHA Taxonomy of antisemitism organized into two coarse-grained categories—**Targeting Jewish Students and Staff** (**Targeting**) and **Antisemitic Expression** (**Expression**)—and nine fine-grained types: seven under **Targeting** and two under **Expression**.

The ADL H.E.A.T. Map & Taxonomy The ADL H.E.A.T. Map is a continuously-collected dataset of hateful and extremist incident descriptions across the U.S., focused on white supremacy, anti-LGBTQ+ hate, and antisemitism. We download and filter the data collected as of December 25, 2024 to 4,522 incidents labeled with the "Antisemitism" tag. The fine-grained categories for antisemitic incidents are **Harassment**, **Assault**, and **Vandalism**. We source category definitions from the ADL's 2023 Audit of Antisemitic Incidents.[10]

5 Methods: Classification Setups

5.1 Experiments on Benchmarking LLMs

We design experiments to assess how off-the-shelf LLMs perform on fine-grained antisemitic event classification. We use the closed-source `gpt-4o` and the open-source `llama3.2-3b-instruct`.

Prompt augmentations. In our baseline setup (NoCtx), we simply formulate d into a prompt that asks M for l, c, and (if applicable) t. In further stress-testing experiments, we explore the utility of providing additional inputs. One additional input is definitions (Def, as with the Def and Assumed-Def setups), in which we supply Wikipedia's general definition of antisemitism, the definitions of each candidate for c and (if applicable) t from the corresponding taxonomy. We also provide in-context examples (ICE, as in the Assumed-ICE setup), in which we prepend one randomly selected entry corresponding to each potential value of t (or c, in the case of the ADL H.E.A.T. Map) from the corpus to create a few-shot learning setting. Finally, we ask M to assume $l =$ antisemitic (Assumed, as in the Assumed, Assumed-Def, and Assumed-ICE setups), thus eliminating the binary task and only prompting the model for coarse- (c) and fine-grained (t) labels. Appendix A provides prompt templates.

Evaluation setup. As appropriate, we report binary detection rate, accuracy, precision, recall, F1, and a weighted modification of F1 (WF1).

For experimental setups that do not assume the event is antisemitic (i.e., NoCtx, Def), we first compute the binary detection rate of antisemitic events, defined as the percentage of entries where the model predicts that the text describes an antisemitic event. We report this rate on the overall datasets and for each category and type. The rate for a category/type is computed among the entries whose gold label contains that category/type.

For all experiments, we compute accuracy, precision, recall, F1 per category and type as well as the means of each score across categories and types. Additionally, since not all types and categories have equal frequency, we compute WF1,

[6] https://encyclopedia.ushmm.org/content/en/article/antisemitism-in-history-racial-antisemitism-18751945

[7] https://holocaustremembrance.com/resources/working-definition-antisemitism

[8] https://time.com/6763293/antisemitism/

[9] We use the collected dataset as of October 10, 2024, with 4410 entries after our filtering process.

[10] https://www.adl.org/resources/report/audit-antisemitic-incidents-2023

Text	Gold	NoCtx	Assumed-ICE	Def
Chalking stating "Ye was Right," which referenced antisemitic comments made by the rapper <PERSON>, and "It's Not Cool to Shill for Israel" was found on Bruin Walk.	**Expression**; *Historical*	**Expression**; *Denigration*	**Expression**; *Denigration*	**Expression**; *Denigration*, *Historical*
According to the ADL, a University at Buffalo student made online threats against an on-campus march organized by the school's Jewish Student Union.	**Targeting**; *Bullying*, *Suppression*	**Targeting**; *Bullying*, *Suppression*	**Targeting**; *Bullying*	**Targeting**; *Bullying*
Swastika graffiti was found on a fence post.	**Targeting**; *Genocidal*, *Historical*	**Expression**; *Denigration*, *Historical*	**Expression**; *Genocidal*, *Historical*	**Expression**; *Genocidal*, *Historical*

Table 1: Examples of classification errors on entries from the AMCHA Corpus.

an F1 score weighted by the frequencies of categories or types within the gold labels of the corpus. Precise score definitions are in Appendix D.

5.2 Assessing Validity of Classifier Setups

To examine whether our classifiers did not just default to classifying all event descriptions as antisemitic, we generate a control set of positive news stories about Jewish people and communities with inspiration from the methodology of Hartvigsen et al. (2022). Full details on control set generation are given in Appendix B. We find some non-trivial false positive rates for Llama-3.2, but we do not find any false positives for GPT-4o, which we then deploy in our generalizability experiments.

6 Results

We examine performance along a few axes: binary/coarse-grained vs. fine-grained classification, action vs. rhetoric-oriented categories and types, and historical knowledge-laden vs. more explicit categories and types. This analysis helps us pinpoint which types of events LLMs can better classify and which need more prompt augmentation.

6.1 Binary & Coarse-Grained Categories

Coarse-grained categories are the immediate next level of tags below the binary antisemitism classification. The categories are **Expression** and **Targeting** for the AMCHA corpus and *Assault*, *Vandalism*, and *Harassment* for the H.E.A.T. Map.

AMCHA Neither model achieves high WF1 scores across the board on the NoCtx baseline (Figure 2; mean 35.8% across categories for GPT-4o, 32.1% for Llama-3.2). Both models have high detection rates across categories (95.4% for GPT-4o, 99.8% for Llama-3.2), though that could be due to Llama's high false positive rate (Appendix E).

For GPT-4o, both adding the binary assumption and definitions increase WF1 (37.7% on Def, 38.58% on Assumed, 40.1% on Assumed-Def), as does adding in-context examples (41.3% on Assumed-ICE). Definitions have the opposite effect on Llama-3.2, decreasing category-mean WF1 by 17.3% between NoCtx and Def and by 4.5% between Assumed and Assumed-Def. In-context examples also hurt performance, decreasing WF1 from 72.7% in Assumed to 61.7% in Assumed-ICE. However, the combination of in-context examples and adding the binary assumption has a 32.1% improvement over the NoCtx WF1.

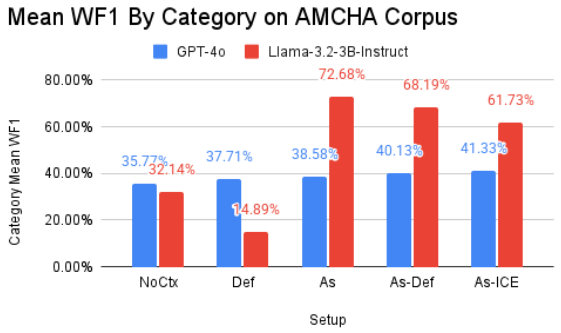

Figure 2: Mean WF1 scores by coarse-grained category on the AMCHA Corpus.

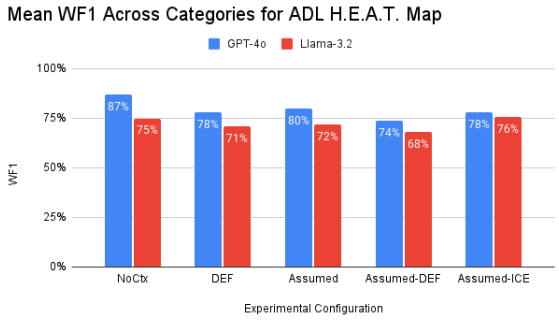

Figure 3: Mean WF1 scores by coarse-grained category on the ADL H.E.A.T. Map.

ADL H.E.A.T. Map. Examining category-mean WF1s for the H.E.A.T. Map (Figure 3), we observe that the optimal setup is NoCtx (86.8%), while the worst is Assumed-Def (73.8%), suggesting that adding definitions and removing the binary question hurt GPT-4o's performance, even more so than adding in-context examples. Assumed-ICE performs the best of all configurations for Llama-3.2 (75.5% WF1), with NoCtx being the second best at 74.9%. As with GPT-4o, the worst configuration is Assumed-Def (68%). For H.E.A.T. Map data, the binary assumption of antisemitism and including definitions both hurt the baseline. Examining binary detection rates, NoCtx (97.6% overall) > Def (90.8%) for GPT-4o, but Def (98.45%) > NoCtx (98.4%) for Llama-3.2.

In sum, **binary detection rates are quite high for both datasets, but category-level distinction is more challenging for both models**, especially for the AMCHA Corpus. With respect to prompt augmentations, **definitions hurt baseline performance, but in-context examples along with adding the binary assumption of antisemitism have opposite effects on AMCHA vs. ADL**.

6.2 Fine-Grained Types

GPT-4o is more aligned to AMCHA's gold labels than Llama-3.2 for fine-grained labels (e.g., 38.8% mean WF1 across types for GPT-4o vs. 27.4% for Llama-3.2 on NoCtx). For GPT-4o, adding definitions (+4.2%), the binary assumption (+0.7%), and both together (4.68%) help WF1 scores. Adding in-context examples helps as well (43.6% mean WF1 across types for Assumed-ICE).

For Llama-3.2, definitions help (33.3% WF1 on Def, 30.6% on Assumed-Def), adding the binary assumption does not (26.9% on Assumed), and in-context examples help slightly (28.6% on Assumed-ICE). In subsequent sections, we examine the effects of different configurations broken down by different clusters of fine-grained types.

6.3 Action vs. Rhetoric

We further analyze how prompts affect coarse- and fine-grained classification, by differentiating between action-oriented (primarily involving physical actions) vs. rhetoric-oriented (primarily involving verbal expressions of hate) categories or types.

AMCHA. *Terrorism*, *Genocidal*, *Historical*, *Denigration*, and *Bullying* are primarily rhetoric-oriented, and *Assault*, *Discrimination*, *Suppression*, and *Destruction* are action-oriented.

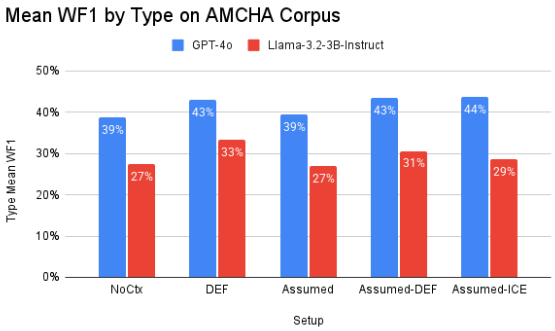

Figure 4: Mean WF1 scores by fine-grained type on the AMCHA Corpus.

Mean WF1s across these clusters show comparable results on GPT-4o for NoCtx (40.9% on rhetoric, 39.8% on action) and Assumed (41.1% on rhetoric, 40.9% on action). The gap expands when we add definitions, giving 46.2% for rhetoric and 39.2% for action in Def and 47.4% for rhetoric and 39.1% for action in Assumed-Def. For Assumed-ICE, we see a larger gap than with no context but smaller than with definitions: 45.5% for rhetoric and 42.8% for action.

These results suggest **that GPT-4o consistently classifies rhetoric-oriented types better than action-oriented types, and definitions help significantly for rhetoric but not for action, while in-context examples help moderately for both**. For Llama-3.2, a similar pattern mostly holds, with the exception that Assumed and Assumed-ICE (both have the binary assumption but no term definitions) have higher WF1 scores for action than rhetoric. This suggests that the performance gap may close or reverse with the binary assumption.

ADL H.E.A.T. Map. For the H.E.A.T. Map, *Harassment* is primarily rhetoric-based, while *Assault* and *Vandalism* are primarily action-based.

There is no clear pattern of higher or lower performance based on whether a category primarily involves action or rhetoric. Notably, *Assault* is the most stable across configurations, perhaps because assault can be verbal or physical. For instance, GPT-4o gives *Vandalism* (90.7%) > *Assault* (88.3%) > *Harassment* (84.6%) for WF1s on NoCtx. With definitions in Def, the performance of *Harassment* and *Vandalism* decrease (-3.7% and -18.8%, respectively), while the performance of *Assault* remains quite stable (88.8%). The performance of *Assault* also remains stable in Assumed

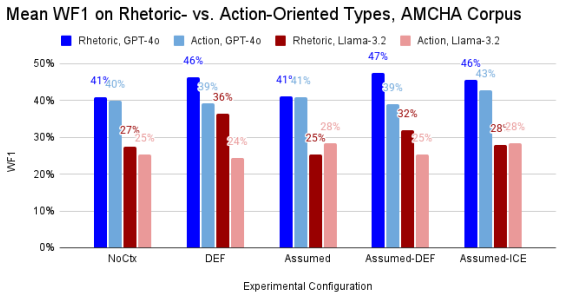

Figure 5: Mean WF1 scores on rhetoric- vs. action-oriented fine-grained types in the AMCHA Corpus. GPT-4o consistently classifies rhetoric-oriented types better than action-oriented types, and definitions help significantly for rhetoric but not for action, while in-context examples help moderately for both clusters.

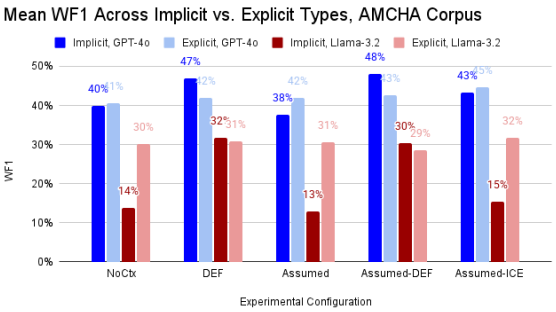

Figure 6: Mean WF1 scores on rhetoric- vs. action-oriented fine-grained types in the AMCHA Corpus. GPT-4o is good at binarily detecting implicit types but not at disambiguating them at a more fine-grained level without definitions or in-context examples.

and ASSUMED-DEF, and it changes the least (-4.1%, vs. -5.9% for *Harassment* and -14.2% for *Vandalism*) for ASSUMED-ICE.

Alternatively, looking at category-level accuracy, we see consistently that *Assault* > *Vandalism* > *Harassment*, showing the surprising result that **both models perform better at action-oriented categories than rhetoric-oriented ADL categories**.

However, Llama-3.2 shows *Harassment* > *Vandalism* > *Assault* in WF1 for all configurations, **supporting the hypothesis that Llama-3.2 is more well-versed in classifying rhetoric than action**.

6.4 Historical Knowledge

In the AMCHA Corpus, *Historical* and *Genocidal* (henceforth "implicit types") require significant historical knowledge to discern, as the tropes involved in both types usually require historical context to understand. Thus, we analyze the mean performance of those two types compared to the mean of the other 7 types (henceforth "explicit types") to assess the effect of implicit historical knowledge.[11]

For GPT-4o, the binary detection rate of both clusters is similar across the board (all above 94%), with the implicit types being detected as antisemitic slightly more often than the explicit types. However, despite being detected, the implicit types are usually misclassified within the taxonomy more often: in NOCTX, we have a 39.9% WF1 for implicit types and a 40.6% WF1 for explicit types, while ASSUMED gives 37.6% (implicit) and 42.0% (explicit), suggesting that the bottleneck is making the distinction among the fine-grained types conditional on the knowledge that the event is antisemitic in a binary sense. Comparing the two clusters, we see slightly better (though still similar) performance for explicit types, and adding the binary assumption widens the performance gap.

Adding definitions alleviates this problem moderately: DEF gives 47% (implicit) vs. 42% (explicit), and ASSUMED-DEF gives 48% vs 42.5%. Adding in-context examples helps, but more for explicit types (44.6%) than for implicit ones (43.3%). This suggests that GPT-4o is good at binarily detecting implicit types but not fine-grained disambiguation without definitions or in-context examples, which provide crucial historical knowledge. With definitions and examples, the existing performance gap between implicit and explicit types also widens from the baseline, but unlike the baseline, there is concrete improvement for both clusters.

Llama-3.2 has similar 97% binary detection rates across all configurations and both clusters, and the pattern of implicit types performing worse without context is more significant (13.8% WF1 for implicit vs. 30.1% for explicit on NOCTX, 13.0% vs. 30.6% on ASSUMED). Like with GPT-4o, definitions help moderately with implicit types, augmenting the WF1s to 31.8% implicit vs. 30.9% explicit for DEF and 30.4% vs. 28.6% for ASSUMED-DEF. Similarly to GPT-4o, in-context examples help slightly, but much less than definitions (15.6% implicit vs. 31.6% explicit WF1 on ASSUMED-ICE). Thus, a pattern emerges that **binary detection of implicit types is nearly perfect for both models, but disambiguation is a bottleneck that can be moderately alleviated by definitions.**

[11] This subsection only discusses the AMCHA Corpus because none of the ADL H.E.A.T. Map categories inherently require historical knowledge to understand.

7 Detecting New Antisemitic Events

In this section, we present a real-world use case of our best-performing, most comprehensive classification prompt to examine whether our classifier generalizes to surfacing unseen events.

7.1 Experimental Setup

We apply our best performing setup by mean fine-grained WF1 (GPT-4o, DEF) with the AMCHA taxonomy on a set of campus newspaper articles dated from October 1, 2022 to December 25, 2024 (∼1 year before October 7, 2023 until the day of data collection) from 5 of the universities with the most frequent incident counts according to AMCHA's data: the Harvard Crimson,[12] the Stanford Daily,[13] the Michigan Daily,[14] the Daily Illini,[15] and the Columbia Spectator.[16] We selected these five out of the top 10 universities with most frequent incidents based on feasibility of web scraping, for a total of 5,275 articles analyzed. Of these, some authors who are experts on antisemitism annotated 225 entries (roughly evenly distributed across publications) according to the AMCHA taxonomy to benchmark our classifier's performance.

7.2 Campus Newspaper Results

Out-of-domain performance vs. Expert Labels. We first analyze the results of benchmarking GPT-4o's labels against human expert labels, focusing on publications for which there was at least one rhetoric-oriented and one action-oriented incident or one implicit and one explicit incident. For the Michigan Daily, we observe a detection rate of 62.5% for rhetoric-based types and 0% for action-based types; the corresponding WF1s were 29.17% and 0%. We observe a similar pattern for the Columbia Spectator: 100% vs. 70% detection rate, and 26.67% vs. 11.11% WF1. For the Harvard Crimson, the detection rate for both clusters was 100%, but the WF1 was also higher for rhetoric-oriented (25%) than action-oriented types (16.67%). The Stanford Daily followed a similar pattern: 50% vs. 0% detection rate, 16.67% vs. 0% WF1. Overall, this confirms our results in §6.3 that **GPT-4o is resoundingly better at detecting and classifying antisemitic rhetoric than action**.

Clustering by historical knowledge required, we observe that GPT-4o tends to do better with entries that require historical knowledge, detecting such examples at a rate of 100% for the Michigan Daily and Stanford Daily, with WF1s of 33.33% for both. The detection rates for explicit types for these papers were 37.5% and 0%, and the WF1s were 7.14% and 0%. This is surprising given the results in Section 6.4 showing that models perform better on types that do not require historical knowledge.

Looking at coarse-grained categories, we consistently observe higher precision than recall for category-level means (35.71% vs. 8.93% for Michigan Daily, 45.83% vs. 18.83% for Columbia Spectator, 100.00% vs. 75.00% for Harvard Crimson, 75% vs. 34.09% for Stanford Daily). However, this pattern is reversed for type-level means, suggesting that our classifier is most useful for surfacing a lower bound of antisemitic incidents and classifying them at a coarse-grained level.

Surfacing Antisemitic Events on Campuses. To showcase the utility of our setups for surfacing antisemitic event reports, we include results from GPT-4o's higher-precision predictions on Harvard and Stanford's newspapers, on the full 1716 and 950 articles scraped. Within the Harvard Crimson, we find 139 antisemitic incidents reported over the last 2 years: 94 under **Expression** and 45 under **Targeting**. In the Stanford Daily, we find 43 incidents (26 for **Expression** and 17 for **Targeting**). The **Targeting** incidents suggest a significant number of actively hostile actions toward Jewish students and staff, suggesting a need for critical interventions to ensure the safety of Jewish people on these campuses. The **Expression** incidents point to a need for university-wide anti-hate education. Flagged events are also often of broader national importance and involve larger organizations in handling situations (e.g. lawsuits,[17] Congressional hearings,[18] FBI reports[19]), which can help practitioners find the right organizations to contact and collaborate with to further combat antisemitism and co-occurring forms of hate.[20]

8 Related Work

Detecting Antisemitism A small subset of toxicity detection literature tackles antisemitism, and

[12] https://www.thecrimson.com/
[13] https://stanforddaily.com/
[14] https://www.michigandaily.com/
[15] https://dailyillini.com/
[16] https://www.columbiaspectator.com/
[17] http://tiny.cc/crimson-lawsuit
[18] http://tiny.cc/crimson-stefanik
[19] http://tiny.cc/stanford-fbi
[20] http://tiny.cc/stanford-report

fewer works examine its subtypes. Most works focus on detecting binary online antisemitism for content moderation, e.g. for websites (Warner and Hirschberg, 2012), Gab (Bagavathi et al., 2019; Sap et al., 2020), Twitter (Smedt, 2021; Ron et al., 2023; Chew, 2021; Arviv et al., 2021; ElSherief et al., 2021; Sap et al., 2020; Chandra et al., 2021; Mihaljević and Steffen, 2023; Steffen et al., 2023; Jikeli et al., 2022; Ozalp et al., 2020; ADL, 2018), Facebook (Smedt, 2021), Instagram (Vargas et al., 2022), Telegram (Mihaljević and Steffen, 2023; Steffen et al., 2023), 4chan /pol/ (González and Zannettou, 2023; Ali and Zannettou, 2022), Stormfront (Sap et al., 2020), YouTube (Khorramrouz et al., 2023; Barna and Knap, 2021), and synthetically generated text (Hartvigsen et al., 2022; Khorramrouz et al., 2023). To our knowledge, we are the first to examine antisemitism computationally from the perspective of hateful *event description* understanding rather than hate *speech* detection.

Taxonomies of Hate A few works have built hate speech taxonomies (Talat et al., 2017; Zufall et al., 2022; Khurana et al., 2022). For example, Sap et al. (2020) built bottom-up explanations of harmful stereotypes and tagged a dataset with particular stereotypes invoked and identity groups targeted in each entry. ElSherief et al. (2021) augmented this work by adding a subset of data that tagged and described implicit forms of hate. Others have assessed how LLM performance on hate detection varies by prompt construction and taxonomy definition (Pavlovic and Poesio, 2024) and developed frameworks to enhance the robustness of LLMs as hate speech detectors and annotators (Kumar et al., 2024). However, no work has yet examined LLMs' ability to operationalize fine-grained definitions of toxicity and detect when those definitions apply in event reports, especially for antisemitism.

9 Conclusion and Discussion

This work introduces a novel task of detecting fine-grained harmful event types and evaluates the utility of off-the-shelf LLMs in contexts of antisemitism. To test our hypotheses, we use the ADL H.E.A.T. Map corpus and contribute a novel AMCHA Corpus, both drawn from real-world incident databases with various levels of categorical tags. We experiment with GPT-4o and Llama-3.2 with prompt-level interventions such as adding definitions, in-context examples, and a binary assumption of antisemitism. We also experiment with generalizing our classifier to surface unseen event reports from freshly scraped newspaper data on university campuses with frequent incidents.

Our findings show that rhetoric-oriented types tend to be more accurately classified than action-oriented types, supporting the hypothesis that LLMs are more familiar with hate speech detection than embodied event understanding. Among our experimental configurations, adding definitions helps most with boosting rhetoric detection, while adding in-context examples helps moderately with both clusters, and adding the binary assumption of antisemitism also helps with both clusters. However, this result may only hold if the taxonomy is sufficiently fine-grained—while we do find this pattern with Llama-3.2, it does not hold for GPT-4o.

We also surprisingly find that while our results on the AMCHA Corpus show poorer classification rates for implicit types (that heavily rely on historical knowledge), the reverse is true for in-the-wild scraped newspaper data. Within the AMCHA Corpus, we find that adding definitions helps disambiguate among implicit types. In general, both models have high binary detection rates but relatively lower rates of classifying the incidents correctly, with in-context examples helping on both datasets but with definitions having mixed effects. For fine-grained types, adding definitions, binary assumptions, and in-context examples help augment WF1 performance. We also show that for a few campus publications, our classifier has high precision such that it can surface a lower bound of antisemitic incidents from previously unseen corpora. Our findings suggest that LLMs have some potential to be adapted to understand harmful events at scale, answer computational social science questions about correlations between hateful events and broader political or economic trends, and decrease the human burden of exposure to distressing news while using the narrative nature of event reports to better grasp real-world manifestations of harm toward marginalized communities.

Future work Future work should improve both models' classification of incidents that are action-oriented and/or require historical knowledge to understand. Future work can also generalize our study to other forms of hate with multiple stakeholders who have differing perspectives, possibly through creating annotator-specific taxonomies with definitions that can steer LLMs to represent different annotators' stances as in Deng et al. (2023).

10 Limitations

We acknowledge the following limitations:

1. The AMCHA Corpus also contains descriptions of college student voices and presidential and student government statements on antisemitism. Though our focus is on antisemitic incidents, this work can be extended to create a taxonomy of types of responses to antisemitism and assess LLMs' abilities to classify these responses and distinguish them from descriptions of outright hateful events.

2. LLMs can be sensitive to small variations in prompt formatting (Sclar et al., 2023). Future work can assess how robust LLM responses are under slight variations of our prompts.

3. In the interest of time and compute budget, our work only explored two models and added one example per type for ASSUMED-ICE. Future work can expand to more model sizes, families, and instruction/chat-tuning levels to investigate effects of these variables on classification.

4. We only examine two taxonomies. However, several other taxonomies of antisemitism appear in related work (JDA, 2021), and future work can create datasets corresponding to those taxonomies and investigate classification on those datasets.

5. Our data consists of English-only event descriptions in the US, and the dataset curators operated under a US-centric socio-cultural lens. Future work should explore antisemitism in other cultures.

6. We focus on antisemitism. Future work could explore other forms of toxicity and create fine-grained taxonomies for them. Datasets describing hateful events against other ethnic minorities are also scarce, so more work should be done to collect such datasets.

7. Though a strength of the AMCHA Corpus is that AMCHA Initiative team members, who are experts in antisemitism, created and labeled it, this also means that we have limited information available as to the details of some steps in the collection process. For example, we do not have complete information on how the team handled submissions that were rejected from inclusion in the corpus, and the sampling strategy of manually tracking news and social media may induce biases. Future work could explore more statistics-based sampling strategies, both for generating positive examples and control sets.

11 Ethics Statement

Environmental Statement: Our experiments (control data generation, AMCHA and ADL H.E.A.T. Map predictions, news data predictions) cost $350.12 in total through OpenAI's API. For Llama-3.2, we use HuggingFace (free) on 8 A100 GPUs. Experiments took approximately 1 day per model per dataset. Llama-3.2 has 3 billion parameters, while GPT-4o's parameter count is undisclosed. We use `pandas`[21] to load our corpus and use `scikit-learn`[22] and `matplotlib`[23] to compute and visualize evaluation metrics. Results are reported on single runs of each entry.

We also acknowledge that studying antisemitism is fraught, as much as it matters for creating more informed humans and models and safer online and offline spaces.While we believe our work is meaningful to study LLM alignment and steerability in toxicity detection, we acknowledge that some types in the taxonomy have contested associations with antisemitism and that some scholars and activists have criticized this taxonomy and how the AMCHA Initiative operationalizes it. We do not promote this taxonomy as universal; rather, we present a research case study on LLM capabilities for this novel event classification task that the corpus and taxonomy are computationally suited for.

Positionality Statement: The authors have diverse perspectives on and connections to antisemitism, including US- and Israel-raised Jewish, and Korea- and Europe-born former Christian perspectives.

References

ADL. 2018. Quantifying hate: A year of anti-semitism on twitter.

Moonis Ali and Savvas Zannettou. 2022. Analyzing antisemitism and islamophobia using a lexicon-based approach. In *ICWSM Workshops*.

[21] `https://pandas.pydata.org/`
[22] `https://scikit-learn.org/stable/`
[23] `https://matplotlib.org/stable`

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Eyal Arviv, Simo Hanouna, and Oren Tsur. 2021. It's a thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 61–70.

Arunkumar Bagavathi, Pedram Bashiri, Shannon Reid, Matthew Phillips, and Siddharth Krishnan. 2019. Examining untempered social media: analyzing cascades of polarized conversations. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 625–632.

Ildikó Barna and Árpád Knap. 2021. An exploration of coronavirus-related online antisemitism in hungary using quantitative topic model and qualitative discourse analysis. *Intersections. East European Journal of Society and Politics*, 7(3):80–100.

Adrien Bibal, Nathaniel Gerlek, Goran Muric, Elizabeth Boschee, Steven C. Fincke, Mike Ross, and Steven N. Minton. 2025. Automating annotation guideline improvements using LLMs: A case study. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 129–144, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Daniel L Byman. 2021. How hateful rhetoric connects to real-world violence. https://www.brookings.edu/articles/how-hateful-rhetoric-connects-to-real-world-violence/. Accessed: 2024-4-29.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 148–157, New York, NY, USA. Association for Computing Machinery.

Peter A Chew. 2021. Quantifying polish anti-semitism in twitter: A robust unsupervised approach with signal processing. In *Conference of the Computational Social Science Society of the Americas*, pages 11–22. Springer.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

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.

David Feldman and Marc Volovici. 2023. *Antisemitism, Islamophobia and the Politics of Definition*. Palgrave Critical Studies of Antisemitism and Racism. Springer International Publishing.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Felipe González and Savvas Zannettou. 2023. Understanding and detecting hateful content using contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 257–268.

David M. Halbfinger, Michael Wines, and Steven Erlanger. 2019. Is b.d.s. anti-semitic? a closer look at the boycott israel campaign.

Bernard Harrison and Lesley Klaff. 2021. The IHRA definition and its critics. In Alvin H Rosenfeld, editor, *Contending with Antisemitism in a Rapidly Changing Political Climate*, pages 9–43. Indiana University Press.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Herf. 2021. IHRA and JDA: Examining definitions of antisemitism in 2021. *Fathom*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

JDA. 2021. The jerusalem declaration on antisemitism.

Jerusalem. 2021. The jerusalem declaration on antisemitism. https://jerusalemdeclaration.org/. Accessed: 2024-5-3.

Gunther Jikeli, David Axelrod, Rhonda Fischer, Elham Forouzesh, Weejeong Jeong, Daniel Miehling, and Katharina Soemer. 2022. Differences between antisemitic and non-antisemitic English language tweets. *Computational and Mathematical Organization Theory*.

Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. 2023. A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, 24(4):2598–2615.

Adel Khorramrouz, Sujan Dutta, Arka Dutta, and Ashiqur R. KhudaBukhsh. 2023. Down the toxicity rabbit hole: Investigating PaLM 2 guardrails.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Dohee Kim, Yujin Baek, Soyoung Yang, and Jaegul Choo. 2023. Towards formality-aware neural machine translation by leveraging context information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7384–7392, Singapore. Association for Computational Linguistics.

Brian Klug. 2023. Defining antisemitism: What is the point? In David Feldman and Marc Volovici, editors, *Antisemitism, Islamophobia and the Politics of Definition*, pages 191–209. Springer International Publishing, Cham.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2024. Gen-z: Generative zero-shot text classification with contextualized label descriptions. In *The Twelfth International Conference on Learning Representations*.

Helena Mihaljević and Elisabeth Steffen. 2023. How toxic is antisemitism? potentials and limitations of automated toxicity scoring for antisemitic online content.

Nexus. 2023. The nexus project - israel and antisemitism. `https://nexusproject.us/`. Accessed: 2024-5-3.

Sefa Ozalp, Matthew L. Williams, Pete Burnap, Han Liu, and Mohamed Mostafa. 2020. Antisemitism on twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society*, 6(2):2056305120916850.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Derek Penslar. 2022. Who's afraid of defining antisemitism? *Antisemitism Studies*, 6(1):133–145.

Maria Pontiki, Maria Gavriilidou, Dimitris Gkoumas, and Stelios Piperidis. 2020. Verbal aggression as an indicator of xenophobic attitudes in Greek Twitter during and after the financial crisis. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 19–26, Marseille, France. European Language Resources Association.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(1):204.

Pekka Räsänen, James Hawdon, Emma Holkeri, Teo Keipi, Matti Näsi, and Atte Oksanen. 2016. Targets of online hate: Examining determinants of victimization among young finnish facebook users. *Violence and victims*, 31(4):708–725.

Sharon Richardson. 2021. Against generalisation: Data-driven decisions need context to be human-compatible. *Business Information Review*, 38(4):162–169.

Gal Ron, Effi Levi, Odelia Oshri, and Shaul Shenhav. 2023. Factoring hate speech: A new annotation framework to study hate speech in social media. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 215–220, Toronto, Canada. Association for Computational Linguistics.

Arno Rosenfeld. 2021. Leading jewish scholars say bds, one-state solution are not antisemitic.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Tom De Smedt. 2021. Online anti-semitism across platforms.

Elisabeth Steffen, Helena Mihaljevic, Milena Pustet, Nyco Bischoff, María 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. In *ICWSM*, volume 17, pages 1082–1092.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

UN. 2018. Hate speech and real harm. https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A LLM Prompts

Example user prompts for each setting can be found in the `experiments/event_discovery/log.txt` file in our GitHub repository, which will be released upon paper publication. Our system prompt and exact user prompt formulation function can be found in `experiments/event_discovery/utils.py` within the same repository. An example prompt is given below for the DEF setting on a news article:

```
Antisemitism is defined as hostility to,
prejudice towards, or discrimination
against Jews. The following text,
enclosed in single quotes, describes an
event that happened on 2024-11-07 00:29:00
at Stanford University.
'[article text goes here]'
Does this quoted text describe an
antisemitic event? If so, what category
of antisemitism does the event fall under,
and what type of expression does it fall
under?
Give an answer in the following format:
{'describes_antisemitic_event': bool,
'category': char,
'type_of_expression': int list,}.
For 'category', put the letter
corresponding to the strongest option
below for the quoted text:
A: 'Does not describe an antisemitic
incident'
B: 'antisemitic expression' - Language,
imagery or behavior deemed antisemitic
by the U.S. State Department definition of
antisemitism, or wholly consistent with
that definition
C: 'targeting Jewish students and staff'
- Incidents that directly target Jewish
students on campus or other Jewish
members of the campus community for
harmful or hateful action based on their
Jewishness or perceived support for Israel
For 'type_of_expression', list all numbers
corresponding to the the following options
that apply to the quoted text:
0: 'Does not describe an antisemitic
incident'
1: 'physical assault' - Physically
attacking Jewish students or staff because
of their Jewishness or perceived association
with Israel
2: 'discrimination' - Unfair treatment or
exclusion of Jewish students or staff
because of their Jewishness or perceived
association with Israel
3: 'destruction of Jewish property' -
Inflicting damage or destroying property
owned by Jews or related to Jews
4: 'genocidal expression' - Using imagery
(e.g. swastika) or language that expresses
a desire or will to kill Jews or exterminate
the Jewish people
5: 'suppression of speech/movement/assembly'
- Preventing or impeding the expression of
Jewish students, such as by removing or
defacing Jewish students' flyers, attempting
to disrupt or shut down speakers at Jewish
or pro-Israel events, or blocking access to
```

```
Jewish or pro-Israel student events
6: 'bullying' - Tormenting Jewish students
or staff because of their Jewishness or
perceived association with Israel
7: 'denigration' - Unfairly ostracizing,
vilifying or defaming Jewish students or
staff because of their Jewishness or
perceived association with Israel
8: 'historical' - Using symbols, images and
tropes associated with historical
antisemitism, including by making
"mendacious, dehumanizing, demonizing,
or stereotypical allegations about Jews as
such, or the power of Jews as a
collective-especially but not exclusively,
the myth about a world Jewish conspiracy or
of  Jews controlling the media, economy,
governments, or other societal
institutions"
9: 'condoning terrorism' - Calling for,
aiding or justifying the killing or
harming of Jews
```

B Control Data Generation

Seed phrases for generation of control data, as well as their corresponding dates, a list of universities tracked by the AMCHA Initiative, and a list of states where incidents have been reported by the ADL, are listed at https://docs.google.com/spreadsheets/d/1yNhjQHrfQhk6k3zfJy-CbED1zb5zXPZ8c7thU1FU8sM/edit?usp=sharing. Following the setup of Hartvigsen et al. (2022) in their synthetic hate speech data generation task, we use a temperature of 0.9 for our generations. We use the following user prompt for a given seed phrase P, date d, and university U:

```
"Write a short (<300 tokens), objective
 news article about a {P} that happened
  on {d} at {U}."
```

For a given P, d, city C, and state S, we use the following prompt:

```
"Write a short (<300 tokens), objective
 news article about a {P} that happened
  on {d} in {C}, {S}."
```

C Additional Details About AMCHA Corpus

C.1 Source Collection

The AMCHA Initiative's monitored sources include (a) a list of anti-Zionist campus groups, (b) a list of campus news publications, (c) a list of popular Jewish news publications, (d) a list of Google Alert keywords,(e) a list of antisemitism trackers on social media, and (f) submissions from a reporting form. Lists of monitored sources are available upon request.

If an event passes the initial verification step, another team member (the "descriptor," a different person than the verifier) writes both a short ("Short Description") and long description ("Description") of the event. Depending on the organization responsible for the event, the descriptor may customize a pre-built description template (templates may not have been used verbatim prior to April 2024, as the volume and nature of events in the past year have necessitated some adjustments in the data creation procedure). The descriptions always conclude with links to the source(s) reporting the event and to any photo or video evidence linked to the event. The descriptor also tags the event with a Category and Classification from AMCHA's taxonomy.

All AMCHA entries are sourced from news and social media platforms that are already viewable to the public. However, as an additional step to protect the privacy of individuals potentially named in the corpus, we use Microsoft's Presidio package[24] to replace names of people with the <PERSON> tag. We manually review and remove names from any entries where the call to the Presidio engine fails.

C.2 Dataset Statistics and Details

Table 2 summarizes the label distribution of the AMCHA Corpus as of October 10, 2024. We also analyze the label distribution over the ten years of incidents collected in this dataset, finding a relative decrease in BDS activity, a relative increase in targeting Jewish students and staff, and a peak of antisemitic expression around 2021.

The types under **Targeting** are: Physical Assault (*Assault*), Discrimination (*Discrimination*), Destruction of Jewish property (*Destruction*), Genocidal expression (*Genocidal*), Suppression of speech/movement/assembly (*Suppression*), Bullying (*Bullying*), Denigration (*Den-*

[24] https://microsoft.github.io/presidio/text_anonymization/

Coarse-Grained Category	Frequency
Expression	16.94%
Targeting	83.06%
Fine-Grained Type	**Frequency**
Terrorism	16.03%
Assault	2.63%
Destruction	8.89%
Historical	33.06%
Suppression	27.55%
Genocidal	22.02%
Denigration	31.97%
Discrimination	7.41%
Bullying	32.09%

Table 2: Label distribution of AMCHA Corpus. Note that fine-grained type percentages add up to more than 100% because each entry can have multiple labels.

igration). The types under **Expression** are: Historical antisemitism (*Historical*), Condoning terrorism against Israel or Jews (*Terrorism*). Definitions of taxonomy components can be found at https://amchainitiative.org/categories-antisemitic-activity. To focus our analysis on clear-cut, relatively uncontroversial incidents of antisemitism (JDA, 2021), we omit entries labeled exclusively with types pertaining to anti-Israel sentiment (i.e., *BDS Activity*, *Demonization*, and/or *Denying Jews Self-Determination*) and are left with nine total fine-grained types.[25]

For the ADL H.E.A.T. Map, we have 2.15%, 62.84%, and 35.01% of cases corresponding to *Assault*, *Harassment*, and *Vandalism*, respectively, as of December 25, 2024.

D Reporting Scores

For a coarse-grained category, a prediction is considered a true positive for the purpose of these metrics if its gold categorical label is that category (positive) and the model predicts that category on the multi-class classification portion of this task (true). For a fine-grained type, a prediction is a true positive if its gold type label contains that type (positive) and the model predicts that type among its predicted types (true), as this portion is a multi-label problem. The scores are then computed across the whole dataset, which includes negatives.

For coarse-grained categories, the WF1 is defined as follows: given a multi-class classification task with potential categories $C_1...C_m$, where the dataset contains n_i entries with gold label C_i for $i \in \{1...m\}$, the WF1 across categories is

$$WF1 = \frac{\sum_{i=1^m}(n_i \cdot F1(C_i))}{\sum_{i=1}^{m} n_i}, \quad (1)$$

where $F1(C_i)$ is the F1 score defined in the previous paragraph for category C_i. For fine-grained types, the WF1 is defined in two steps. First, given a multi-label classification task with potential labels $t_1...t_m$, where the jth entry of the dataset D has gold labels $s_{j,1}...s_{j,k}$ and predicted labels $u_{j,1}...u_{j,q}$, and $s_{j,i} \in \{t_1...t_m\}$ and $u_{j,i} \in \{t_1...t_m\}$ $\forall i \in 1...k$, create a dataset D' where the jth entry becomes k separate entries, and the ith separate entry of the original jth entry has gold label $s_{j,i}$ and predicted labels $u_{j,1}...u_{j,q}$. Now, for each type t_i, we define

$$WF1(t_i) = F1(t_i, D'), \quad (2)$$

or the F1 score as defined in the previous paragraph but computed on D' instead of the original dataset D. Then, the overall WF1 across types is

$$WF1 = \frac{\sum_{i=1}^{m} n_i \cdot WF1(t_i)}{\sum_{i=1}^{m} n_i}. \quad (3)$$

E Additional Experiments

E.1 Control Data Ceneration

One error we could not detect with the AMCHA Corpus or ADL H.E.A.T. Map alone was that of false positives at the binary level, as all incidents logged in the datasets are labelled as antisemitic. To assess false positive rates on this task, we run all experimental setups on GPT-4o-generated control sets of positive events that relate to Jewish and/or Israeli people. We generate this control set through the following procedure:

1. The first author, who has a background in Jewish history and is ethnically Jewish, manually crafts a list of seed phrases that describe positive events that relate to Jewish and/or Israeli communities in some way. For each seed phrase, the author then manually selects a reasonable date (arbitrarily within a reasonable range) on which the event described could have occurred. The list of seed phrases and events is in Appendix B.

[25] We recognize that AMCHA's taxonomy is not universal (Rosenfeld, 2021; Halbfinger et al., 2019) and that other significantly differing taxonomies exist with scholarly endorsement (JDA, 2021). We employ the taxonomy in this paper to leverage the advantages of the corpus' uniquely rich content, labels, and metadata for event classification.

2. For each university U in the AMCHA Initiative's list of universities tracked for incidents (or, for the H.E.A.T. Map, for each city x state pair $U = (C, S)$ found in the database):

   (a) For each seed phrase P and corresponding date d, we ask GPT-4o to generate a short, factual news article about the event described by P occurring on d at U. We create $n = 6$ such generations per tuple of $((P, d), U)$ for the AMCHA Corpus and $n = 1$ such generation per tuple for the H.E.A.T. Map, amounting to $n \cdot |\{P\}| \cdot |\{U\}|$ entries in total.

3. We take a random sample of this generated set that is equal in size to the corresponding corpus used for our experiments.

The results of both control sets on GPT-4o on all setups show 100% accuracy: there were no cases in which either model answered that the described event was antisemitic or fit into any category or type of antisemitism. However, there were significant proportions of false positives with Llama-3.2-3B-Instruct.

F Additional Plots and Tables

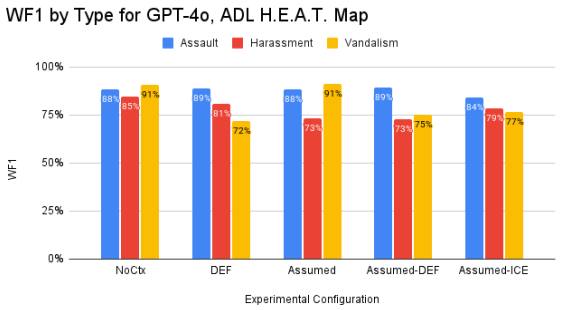

Figure 7: Mean WF1 by category for GPT-4o on the ADL H.E.A.T. Map.

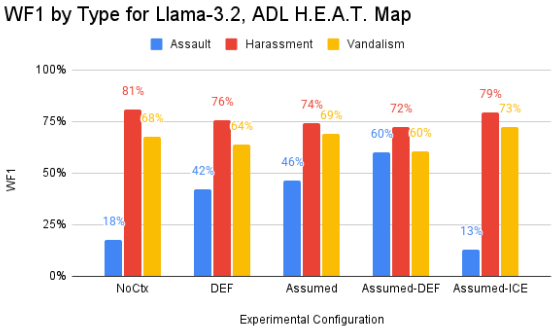

Figure 8: Mean WF1 by category for Llama-3.2 on the ADL H.E.A.T. Map.

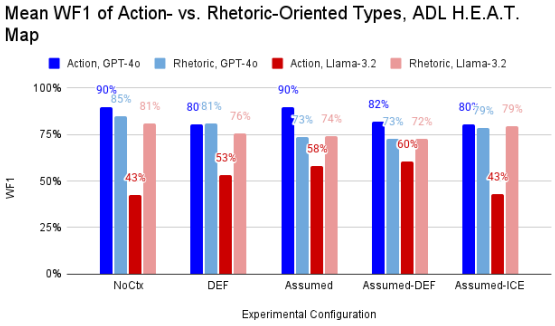

Figure 9: Mean WF1 of action- vs. rhetoric-oriented categories on ADL H.E.A.T. Map.

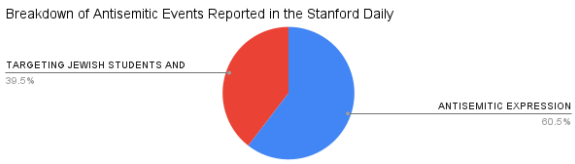

Figure 10: Breakdown of antisemitic events reported in the Stanford Daily.

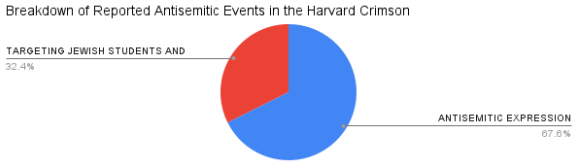

Figure 11: Breakdown of antisemitic events reported in the Harvard Crimson.

Model	Implicit or Explicit	NoCtx	DEF	Assumed	Assumed-DEF	Assumed-ICE
GPT-4o	**Implicit**	**99.07%**	**99.25%**	**99.91%**	**99.64%**	**99.62%**
	Explicit	94.08%	94.23%	98.48%	96.49%	99.49%
Llama-3.2	**Implicit**	99.66%	100.00%	100.00%	100.00%	97.99%
	Explicit	99.85%	99.95%	99.95%	99.96%	99.22%

Table 3: Binary detection rates across implicit vs. explicit type clusters for the AMCHA Corpus.

Publication	# Articles Scraped	# Articles Predicted w/ Anti-semitic Events	# Articles Predicted as **Expression/ Targeting**	Binary F1 on Human Annotations	Mean WF1 Over Categories on Human Annotations	Mean WF1 Over Types on Human Annotations
Harvard Crimson	1716	139	94/45	69.57%	68.89%	55.56%
Stanford Daily	950	43	26/17	40%	29.88%	16.67%
Daily Illini	224	2	1/1	10%	0%	0%
Michigan Daily	1307	31	25/6	35.9%	24.24%	27.78%
Columbia Spectator	1078	149	74/75	40%	18.55%	40.99%

Table 4: Summary of campus newspaper articles scraped and classified.

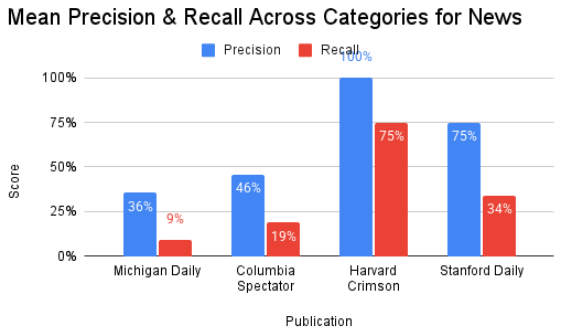

Figure 12: Mean precision and recall scores across coarse-grained categories for freshly scraped news data.

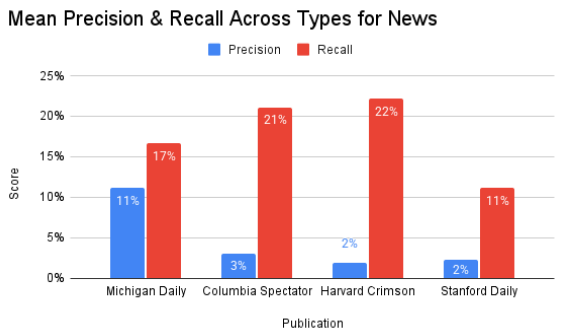

Figure 13: Mean precision and recall scores across fine-grained types for freshly scraped news data.

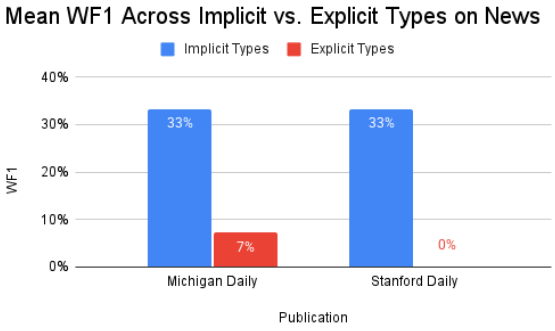

Figure 14: Mean WF1 scores across implicit vs. explicit type clusters for freshly scraped news data.

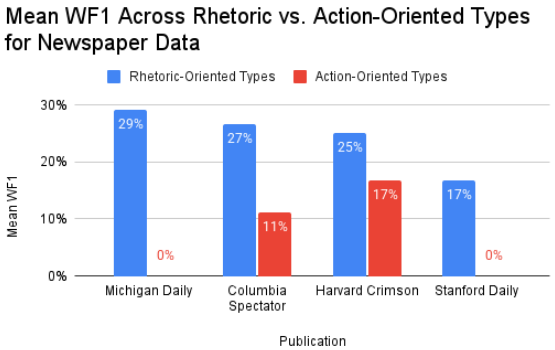

Figure 15: Mean WF1 scores across rhetoric- vs. action-oriented type clusters for freshly scraped news data.

Class/Category	% Antisemitic	Accuracy	Precision	Recall	F1	WF1
Overall	28.22%	28.63%	71.79%	25.74%	37.09%	N/A
Expression	19.94%	34.36%	59.29%	19.03%	27.87%	N/A
Targeting	54.76%	86.82%	43.33%	35.71%	38.00%	N/A
Category Mean	37.35%	60.59%	51.31%	27.37%	32.93%	28.31%
Assault	None	100.00%	0.00%	0.00%	0.00%	0.00%
Discrimination	None	97.76%	0.00%	0.00%	0.00%	0.00%
Destruction	100.00%	68.45%	1.54%	20.00%	2.86%	13.33%
Genocidal	100.00%	62.87%	14.00%	40.00%	20.00%	26.67%
Suppression	13.33%	64.06%	2.22%	8.00%	3.48%	8.89%
Bullying	50.00%	75.54%	1.33%	10.00%	2.35%	8.00%
Denigration	50.00%	60.75%	0.00%	0.00%	0.00%	0.00%
Historical	100.00%	64.71%	0.00%	0.00%	0.00%	0.00%
Terrorism	62.50%	76.86%	13.82%	50.00%	16.97%	43.33%
Type Mean	57.00%	74.55%	3.66%	14.22%	5.07%	28.20%

Table 5: Mean performance metrics of GPT-4o on news data against human expert labels.

Setup	Category	% Antisemitic	Accuracy	WF1
NoCtx	Overall	93.58%		
	Expression	98.26%	40.73%	
	Targeting	92.63%	34.90%	
	Category Mean	95.44%	37.81%	35.77%
DEF	Overall	93.76%		
	Expression	94.91%	40.95%	
	Targeting	93.53%	36.44%	
	Category Mean	94.22%	38.70%	37.71%
Assumed	**Expression**	99.33%	38.71%	
	Targeting	97.93%	37.10%	
	Category Mean	98.63%	37.90%	38.58%
Assumed-DEF	**Expression**	97.32%	41.47%	
	Targeting	95.50%	38.19%	
	Category Mean	96.41%	39.83%	40.13%
Assumed-ICE	**Expression**	99.87%	39.57%	
	Targeting	99.48%	39.16%	
	Category Mean	99.67%	39.37%	41.33%

Table 6: Full performance statistics on the AMCHA Corpus for GPT-4o by category and overall.

Setup	Type	%	Accuracy	F1	WF1
NoCtx	*Terrorism*	98.02%	94.67%	84.49%	63.49%
	Assault	97.41%	99.34%	86.76%	45.02%
	Destruction	96.68%	95.24%	66.45%	43.93%
	Historical	99.18%	79.68%	61.81%	43.83%
	Suppression	86.26%	79.66%	53.93%	39.34%
	Genocidal	98.97%	75.71%	57.18%	36.02%
	Denigration	92.41%	48.12%	52.92%	32.84%
	Discrimination	92.05%	93.76%	42.83%	30.98%
	Bullying	95.76%	74.85%	38.01%	28.31%
	Type Mean	**95.19%**	**82.34%**	**60.49%**	**38.76%**
DEF	*Terrorism*	94.34%	93.83%	81.08%	60.04%
	Assault	96.55%	99.34%	87.55%	44.16%
	Destruction	96.68%	94.97%	63.00%	42.95%
	Historical	99.31%	80.82%	74.70%	49.66%
	Suppression	88.89%	78.75%	51.17%	36.66%
	Genocidal	99.18%	84.67%	73.59%	44.34%
	Denigration	94.26%	69.21%	63.14%	42.59%
	Discrimination	93.88%	90.95%	48.52%	32.90%
	Bullying	94.98%	77.76%	49.82%	34.32%
	Type Mean	**95.34%**	**85.59%**	**65.84%**	**42.91%**
Assumed	*Terrorism*	99.01%	95.37%	86.20%	65.70%
	Assault	99.14%	99.34%	87.34%	45.15%
	Destruction	100.00%	94.94%	66.77%	43.16%
	Historical	99.93%	76.46%	52.60%	39.41%
	Suppression	95.80%	81.70%	61.95%	44.21%
	Genocidal	99.90%	77.46%	55.47%	35.78%
	Denigration	98.09%	50.75%	55.60%	34.64%
	Discrimination	98.78%	92.40%	44.99%	31.10%
	Bullying	98.52%	75.56%	40.77%	30.10%
	Type Mean	**98.80%**	**82.67%**	**61.30%**	**39.46%**
Assumed-DEF	*Terrorism*	97.03%	94.58%	82.49%	63.47%
	Assault	98.28%	99.34%	87.45%	44.20%
	Destruction	98.98%	95.06%	64.72%	43.34%
	Historical	99.59%	79.57%	73.22%	49.24%
	Suppression	91.77%	78.55%	49.25%	35.58%
	Genocidal	99.69%	88.03%	78.00%	46.68%
	Denigration	95.74%	72.24%	64.13%	43.20%
	Discrimination	96.02%	91.07%	50.50%	33.31%
	Bullying	97.60%	77.80%	49.87%	34.51%
	Type Mean	**97.19%**	**86.25%**	**66.63%**	**43.44%**
Assumed-ICE	*Terrorism*	99.86%	95.44%	85.49%	69.04%
	Assault	99.14%	99.37%	88.03%	44.30%
	Destruction	99.74%	95.49%	73.64%	44.73%
	Historical	99.66%	81.72%	64.83%	46.39%
	Suppression	99.42%	83.24%	65.93%	48.08%
	Genocidal	99.59%	80.20%	63.21%	40.17%
	Denigration	99.57%	63.72%	62.91%	40.71%
	Discrimination	99.39%	90.23%	48.63%	33.89%
	Bullying	99.29%	76.01%	43.12%	31.33%
	Type Mean	**99.52%**	**85.05%**	**66.20%**	**43.61%**

Table 7: Full performance statistics on the AMCHA Corpus for GPT-4o by fine-grained types.

Setup	Category	%	Accuracy	WF1
NoCtx	Overall	99.71%		
	Expression	100.00%	32.36%	
	Targeting	99.65%	32.06%	
	Category Mean	99.82%	32.21%	32.14%
DEF	Overall	99.91%		
	Expression	100.00%	22.15%	
	Targeting	99.89%	22.06%	
	Category Mean	99.95%	22.11%	14.89%
Assumed	**Expression**	100.00%	70.09%	
	Targeting	99.92%	70.02%	
	Category Mean	99.96%	70.06%	72.68%
Assumed-DEF	**Expression**	100.00%	63.88%	
	Targeting	99.89%	63.79%	
	Category Mean	99.95%	63.83%	68.19%
Assumed-ICE	**Expression**	97.46%	57.07%	
	Targeting	98.77%	56.49%	
	Category Mean	98.11%	56.78%	61.73%

Table 8: Full performance statistics on the AMCHA Corpus for Llama-3.2 by category and overall.

Setup	Type	%	Accuracy	F1	WF1
NoCtx	*Physical*	100.00%	95.33%	44.32%	22.71%
	Destruction	100.00%	84.58%	47.93%	26.53%
	Historical	99.73%	46.89%	27.04%	19.96%
	Suppression	99.51%	77.94%	54.08%	41.96%
	Genocidal	99.59%	64.31%	10.16%	7.62%
	Denigration	99.86%	48.44%	54.63%	34.78%
	Discrimination	100.00%	61.81%	18.17%	10.12%
	Bullying	99.86%	67.37%	25.56%	19.35%
	Type Mean	99.81%	70.78%	38.74%	27.40%
DEF	*Physical*	100.00%	93.81%	43.24%	22.39%
	Destruction	100.00%	83.33%	44.19%	24.95%
	Historical	100.00%	28.30%	30.32%	20.65%
	Suppression	99.67%	71.86%	55.09%	38.33%
	Genocidal	100.00%	84.33%	70.48%	42.94%
	Denigration	100.00%	71.50%	61.36%	40.88%
	Discrimination	100.00%	60.95%	20.65%	12.21%
	Bullying	100.00%	72.47%	38.12%	28.29%
	Type Mean	99.96%	72.86%	46.72%	33.32%
Assumed	*Physical*	100.00%	98.10%	65.29%	34.88%
	Destruction	100.00%	84.56%	45.65%	25.47%
	Historical	100.00%	58.82%	24.90%	19.07%
	Suppression	100.00%	78.71%	56.51%	43.79%
	Genocidal	100.00%	68.34%	8.76%	6.98%
	Denigration	99.86%	60.45%	59.18%	38.03%
	Discrimination	100.00%	42.31%	16.70%	9.13%
	Bullying	99.93%	70.09%	21.35%	17.71%
	Type Mean	99.96%	71.59%	39.91%	26.88%
Assumed-DEF	*Physical*	100.00%	96.64%	56.47%	29.14%
	Destruction	100.00%	83.38%	42.51%	23.83%
	Historical	100.00%	37.55%	30.94%	21.51%
	Suppression	99.84%	73.33%	51.32%	37.10%
	Genocidal	100.00%	84.01%	58.55%	39.34%
	Denigration	99.86%	71.63%	54.49%	36.64%
	Discrimination	100.00%	47.35%	19.04%	10.92%
	Bullying	100.00%	70.61%	35.46%	25.78%
	Type Mean	99.97%	72.44%	43.30%	30.59%
Assumed-ICE	*Physical*	100.00%	98.19%	63.64%	34.23%
	Destruction	98.98%	81.00%	44.36%	24.04%
	Historical	97.74%	61.13%	30.10%	23.50%
	Suppression	98.93%	77.28%	61.01%	44.00%
	Genocidal	98.25%	70.79%	8.78%	7.43%
	Denigration	99.57%	64.31%	60.21%	40.21%
	Discrimination	98.78%	78.10%	19.50%	10.89%
	Bullying	99.15%	70.59%	21.06%	17.43%
	Type Mean	98.95%	76.54%	40.87%	28.59%

Table 9: Full performance statistics on the AMCHA Corpus for Llama-3.2 by type.

Setup	Category	%	Accuracy	WF1
NoCtx	Overall	97.63%		
	Assault	98.97%	99.49%	88.32%
	Harassment	96.55%	78.48%	84.59%
	Vandalism	99.49%	93.34%	90.71%
	Type Mean	98.34%	90.44%	86.81%
DEF	Overall	90.78%		
	Assault	98.97%	99.49%	88.78%
	Harassment	88.24%	74.56%	80.87%
	Vandalism	94.82%	83.39%	71.88%
	Type Mean	94.01%	85.81%	77.89%
Assumed	*Assault*	98.97%	99.47%	88.24%
	Harassment	72.40%	68.81%	73.44%
	Vandalism	99.62%	93.61%	90.98%
	Type Mean	90.33%	87.30%	79.90%
Assumed-DEF	*Assault*	98.97%	99.49%	89.10%
	Harassment	73.74%	66.60%	72.66%
	Vandalism	98.93%	85.03%	74.94%
	Type Mean	90.55%	83.71%	73.81%
Assumed-ICE	*Assault*	98.97%	99.20%	84.21%
	Harassment	84.69%	72.71%	78.66%
	Vandalism	99.94%	85.25%	76.52%
	Type Mean	94.53%	85.72%	78.03%

Table 10: Full performance statistics on the ADL H.E.A.T. Map for GPT-4o.

Setup	Category	%	Accuracy	WF1
NoCtx	Overall	98.39%		
	Assault	98.97%	80.23%	17.53%
	Harassment	97.61%	73.74%	80.91%
	Vandalism	99.75%	67.04%	67.58%
	Type Mean	98.77%	73.67%	74.88%
DEF	Overall	98.45%		
	Assault	100.00%	94.23%	42.13%
	Harassment	97.71%	67.86%	75.53%
	Vandalism	99.68%	60.56%	63.94%
	Type Mean	99.13%	74.22%	70.76%
Assumed	*Assault*	98.97%	95.82%	46.46%
	Harassment	84.34%	72.62%	74.15%
	Vandalism	97.98%	68.83%	69.19%
	Type Mean	93.76%	79.09%	71.82%
Assumed-DEF	*Assault*	97.94%	97.37%	60.20%
	Harassment	97.18%	71.91%	72.45%
	Vandalism	99.87%	54.30%	60.50%
	Type Mean	98.33%	74.53%	68.00%
Assumed-ICE	*Assault*	100.00%	73.90%	13.11%
	Harassment	92.82%	77.22%	79.28%
	Vandalism	99.43%	73.70%	72.53%
	Type Mean	97.42%	74.94%	75.50%

Table 11: Full performance statistics on the ADL H.E.A.T. Map for Llama-3.2.