On the Importance of Nuanced Taxonomies for LLM-Based Understanding of Harmful Events: A Case Study on Antisemitism

Karina Halevy,[1] Julia Mendelsohn,[2] Naomi Younger,[3] Tammi Rossman-Benjamin,[3]
Chan Young Park,[1] Yulia Tsvetkov,[4], Mona Diab,[1] 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

Monitoring the news at scale for incidents of hate, violence, and other toxicity is essential to understanding broad societal trends, including harms to marginalized communities. As large language models (LLMs) become a primary tool for understanding events at scale, they can be useful for elucidating these harms. However, labeling harmful events is challenging due to the subjectivity of labels such as "toxicity" and "hate." Motivated by the rise of antisemitism, this paper presents a case study of the capability of LLMs to discover reports of antisemitic events. We pilot the task of hateful event classification on the AMCHA Corpus, a continuously updated dataset with expert-labeled instances of 3 coarse-grained categories and 14 fine-grained types of antisemitism. We show that incorporating domain knowledge from fine-grained taxonomies is needed to make LLMs more effective at this task. Our experiments find that providing precise definitions from a fine-grained taxonomy of antisemitism can steer GPT-4 and Llama-3 to perform better on tagging antisemitic event descriptions to a limited extent, with GPT-4 achieving up to a 14% increase in mean weighted F1. However, our results suggest that LLMs are still far from perfect at understanding antisemitic events, suggesting avenues for future work on more creative LLM alignment and more policy work on creating precise definitions of antisemitism.

1 Introduction

Understanding and labeling hateful or harmful events from news reports can reveal broad societal trends (Pontiki et al., 2020) and harms toward marginalized communities.[1] For example, hate speech against ethnic minorities can increase psychological distress and vulnerability within those communities and embolden perpetrators of physical violence against community members.[2]

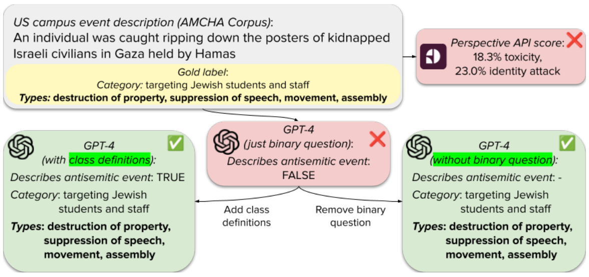
US campus event description (AMCHA Corpus):
An individual was caught ripping down the posters of kidnapped Israeli civilians in Gaza held by Hamas
Gold label:
Category: targeting Jewish students and staff
Types: destruction of property, suppression of speech, movement, assembly
Perspective API score:
18.3% toxicity,
23.0% identity attack
GPT-4
(just binary question):
Describes antisemitic event:
FALSE
GPT-4
(with class definitions):
Describes antisemitic event: TRUE
Category: targeting Jewish students and staff
Types: destruction of property, suppression of speech, movement, assembly
Add class definitions
Remove binary question
GPT-4
(without binary question):
Describes antisemitic event: -
Category: targeting Jewish students and staff
Types: destruction of property, suppression of speech, movement, assembly

Figure 1: An AMCHA Corpus entry where the Perspective API assigns low toxicity probabilities, and a context-free GPT-4 prompt, given only the description and its corresponding date and university, does not predict antisemitism described in the text. However, the interventions of adding definitions and removing the binary question both trigger GPT-4 to predict the gold label category and types.

Labeling whether a text describes a harmful event is non-trivial, as 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). Disagreement can occur in translating coarse concepts such as "harmful," "toxic," or "antisemitic" to or from specific subtypes, especially when translation guidelines are abstract or contested, tasks require domain knowledge (Kim et al., 2023), or texts contain implicit hate (ElSherief et al., 2021). LLMs may then operationalize an "average" perspective when in reality one of two annotators sees a harmful stereotype, erasing valuable disagreement (Pavlovic and Poesio, 2024; Richardson, 2021).

This work investigates approaches to address these challenges by adding fine-grained prior knowledge to LLM prompts. We stress-test LLMs' ability to perform nuanced classification for de-

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

[2] https://www.un.org/en/hate-speech/impact-and-prevention/why-tackle-hate-speech

scriptions 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). Despite its controversial nature,[3] studying antisemitism is 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 extract and release the AMCHA Corpus to the research community. The AMCHA Corpus is a unique challenge set of textual descriptions of antisemitic events from US university campuses, collected continuously from 2015 to the present. The corpus is annotated with 3 coarse-grained categories and 14 fine-grained types (see Section 2.3) by experts who have academic, professional, and lived experiences of antisemitism. This corpus is particularly challenging because the predominant focus of hate detection has been toxic comment classification, but hateful event understanding requires different forms of comprehension and reasoning (see Figure 1 for an example), not least because reports of harmful events likely do not contain verbiage from hate speech lexica.

Our work asks the following research questions:

1. How well do LLMs label the coarse-grained categories and fine-grained types of antisemitism included in the AMCHA Corpus?
2. To what extent can we steer LLMs to use various definitions of antisemitism?
3. Within texts labeled as antisemitic, which types and categories of antisemitic events are harder for LLMs to predict?
4. How much can in-context learning improve LLMs' antisemitic event classification?

Overall, we find that models with no context perform poorly on fine-grained classification, performing even worse for contested types than for more uncontroversial types. Furthermore, LLMs can be steered to a limited extent to improve fine-grained classification. Supplying definitions of categories and types is more helpful than providing in-context examples, though performance is still weak for several types even with these interventions. Comparing two recent LLMs, we see that OpenAI's GPT-4 (OpenAI et al., 2024) is more steerable towards improved fine-grained classification, while Meta's Llama 3-8B-Instruct[5] is more steerable toward improved coarse-grained classification.

Through these experiments, we illustrate the importance of specificity in annotation guidelines and label definitions for toxicity detection and argue for the utility of producing more precise definitions of concepts like antisemitism and toxicity.

2 The AMCHA Corpus

We scrape and release the AMCHA Corpus to the research community. In this section, we first motivate the use of this corpus: the first subsection provides a primer on definitions and recent trends of antisemitism. The second subsection gives an overview of the corpus' collection process and contents, and the third subsection specifies the typology used to define antisemitism within the corpus.

2.1 Background: Antisemitism

Antisemitism, an old yet constantly evolving form of hatred, often fails to fit neatly into modern notions of race, racism, religion, and religious discrimination (Museum). According to the International Holocaust Remembrance Association (IHRA), antisemitism is "a certain perception of Jews, which may be expressed as hatred toward Jews," including "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."[6] Precise taxonomies and typologies characterizing antisemitism vary, with the antisemitic nature of certain actions and rhetoric being actively debated (JDA, 2021). This debate thus motivates the need for more fine-grained categorization of potential forms of antisemitism.

The urgent need to address antisemitism on college campus settings, especially given recent surges of antisemitism on university campuses

[3] While antisemitism is controversial due to events in the Middle East, we believe it is important to study since antisemitism is continuously on the rise, and Jewish people are the 2nd-most targeted group in hate crimes (FBI, 2023). We take a descriptive stance to accommodate disagreement on definitions of antisemitism, and our methods can adapt to varying views (see Section 8).

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

[5] https://ai.meta.com/blog/meta-llama-3/

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

since Hamas' attacks on October 7, 2023,[7] have created a need for these incidents to be labelled 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, most research centered on antisemitism has focused on isolated rhetorical snippets rather than descriptions of events with larger contexts.

2.2 Dataset Overview

The AMCHA Corpus is a growing challenge set of 6,748 English-language entries[8] contextualized descriptions of antisemitic events that have occurred on higher education campuses, annotated for coarse- fine-grained categories of antisemitic expression. To our knowledge, the AMCHA Corpus is the only publicly accessible dataset of this kind that is also accompanied by a taxonomy with which it is labeled—other databases exist[9] but do not have fully defined taxonomies. The dataset is collected by the AMCHA Initiative through a continuous monitoring and screening procedure, summarized below and outlined in detail in Appendix C:

1. Collectors from the AMCHA Initiative monitor news publications and antisemitic groups.
2. An AMCHA team member then verifies that the post describes an event that harmed Jewish people and impacted a higher education campus community, and fact-checks the description. The verifier also verifies the factuality of the incident and its associated reporting.
3. Another AMCHA team member writes a short and long description of the event, and categorizes it according to their typology.

2.3 Typology

The AMCHA Initiative's definition of antisemitism is organized as follows (coarse-grained categories are bolded, fine-grained types are bulleted):[10] [11]

Targeting Jewish Students and Staff (**Targ**): 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:

- *Physical Assault* (**T**-*Phy*)—Physically attacking Jewish students or staff because of their Jewishness or perceived association with Israel.
- *Discrimination* (**T**-*Dis*)—Unfair treatment or exclusion of Jewish students or staff because of their Jewishness or perceived association with Israel.
- *Destruction of Jewish Property* (**T**-*Des*)—Inflicting damage or destroying property owned by Jews or related to Jews.
- *Genocidal Expression* (**T**-*Gen*) Using imagery (e.g. swastika) or language that expresses a desire or will to kill Jews or exterminate the Jewish people.
- *Suppression of Speech, Movement, or Assembly* (**T**-*Sup*)—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.
- *Bullying* (**T**-*Bul*)—Tormenting Jewish students or staff because of their Jewishness or perceived association with Israel.
- *Denigration* (**T**-*Den*)—Unfairly ostracizing, vilifying or defaming Jewish students or staff because of their Jewishness or perceived association with Israel.

Antisemitic Expression (**Expr**): Language, imagery or behavior deemed antisemitic by the U.S. State Department definition of antisemitism, or wholly consistent with that definition:

- *Historical Antisemitism* (**A**-*His*)—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" (U.S. State Department).
- *Condoning Terrorism Against Israel or Jews* (**A**-*Ter*)—Calling for, aiding or justifying the killing or harming of Jews.

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

[8] We use the downloaded dataset as of January 24, 2024.

[9] https://www.adl.org/resources/tools-to-track-hate/heat-map, https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime, https://en-humanities.tau.ac.il/roth/publications/db

[10] https://amchainitiative.org/categories-antisemitic-activity

[11] We recognize that this definition is not universal (Rosenfeld, 2021; Halbfinger et al., 2019) and that other significantly differing taxonomies exist with scholarly endorsement (JDA, 2021). We employ the AMCHA Initiative's taxonomy in this paper to leverage the affordances of the corpus' uniquely rich content, labels, and metadata for our event classification task. See Section 8 for further discussion.

Text	**Gold**	M-As	M-As-IcE	M-As-Def
A swastika was found in an anatomy lab on lab material.	**Expr**; **T**-*Gen*, **A**-*His*	**Expr**; **T**-*Des*, **T**-*Den*	**Expr**; **T**-*Des*	**Expr** **T**-*Gen*, **A**-*His*
According to the ADL, "An antisemitic post was made on University of Arizona YikYak stating, 'ZBT and Sammy and AEPI being able to run this school is more proof that *ws run the world (sic)."	**Targ**; **T**-*Den*, **A**-*His*	**Expr**; **T**-*Den*	**Targ**; **T**-*Bul*	**Expr**; **T**-*Den*, **A**-*His*
According to the ADL, "Pieces of paper with racist and homophobic slurs and swastikas were left in a comment box at the University of Utah."	**Targ**; **T**-*Gen*, **A**-*His*	**Expr**; **T**-*Den*	**Expr**; **A**-*His*	**Expr**; **T**-*Gen*, **A**-*His*
According to the ADL, "Swastika graffiti was located at a light rail station near the campus of Sacramento City College."	**Targ**; **T**-*Gen*, **A**-*His*	**Expr**; **T**-*Des*, **T**-*Den*	**Expr**; **T**-*Des*	**Expr**; **T**-*Gen*, **A**-*His*
According to the ADL, "A swastika was drawn on a sidewalk in front of a Queens College building."	**Targ**; **T**-*Gen*, **A**-*His*	**Expr**; **T**-*Des*, **T**-*Den*	**Expr**; **T**-*Des*	**Expr**; **T**-*Gen*, **A**-*His*

Table 1: Examples of entries from the AMCHA Corpus that encountered various classification errors when considering corpus labels as gold labels. Full names of abbreviated categories and types are specified in Section 2.3.

- *Denying Jews Self-Determination* (**A**-*Det*)—Denying Israel the right to exist or promoting the elimination of Israel as a Jewish state.
- *Demonization of Israel* (**A**-*Dem*)—Using symbols, images and tropes associated with classic antisemitism to characterize Israel, Israelis, Zionism or Zionists, such as claiming that Israelis are evil or blood-thirsty and deliberately murder children or that Zionism is white supremacy, or delegitimizing Israel by insinuating that Israel is an illegitimate state and does not belong in the family of nations.

Boycott, Divest, Sanction (BDS) Activity (BDS):*[12] These incidents contain the following activities:

- *Calls for BDS** (**B**-*Cal*)—Promoting BDS verbally or by writing, signing or publicizing resolutions, petitions, statements or op-eds calling for BDS.
- *BDS Votes** (**B**-*Vot*)—Considering, discussing or voting on resolutions calling for BDS.
- *BDS Events** (**B**-*Eve*)—Holding events which promote BDS.

We have placed asterisks next to categories and types in the provided typology that have been contested and not fully agreed upon by the authors, as we believe that observing performance changes on these particular examples helps shed light on the importance of working toward precise definitions of antisemitism. See Section 8 for further discussion.

3 Methods: Classification Setups

3.1 Experiments

We perform a set of experiments assessing how popular LLMs perform on the task of fine-grained antisemitic event classification. We experiment on two of the most capable LLMs currently available, the closed-source `gpt-4-1106-preview`[13] and the open-source Meta's `llama3-8b-instruct`.[14] In the interest of investigating whether this task is possible with fewer compute resources, we also experiment on lighter-weight hate and toxicity classifiers, specifically HateBERT (Caselli et al., 2021) and the Perspective API.[15] Our classification task is set up as follows: Given an input event description along with the date and university of the event (collectively, input d), model M must classify the coarse-grained categorical label c of the event, the set of n fine-grained type labels $t = t_1, \ldots, t_n$, as

[12] https://amchainitiative.org/BDS-background

[13] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[14] https://llama.meta.com/llama3/

[15] https://perspectiveapi.com/

well as, optionally, the binary antisemitism label l. In some experiments, we provide additional inputs such as definitions (DEF), in-context examples (ICE), as well as the antisemitism label l (AS).

Our first two experiments aim to measure the closed-book ability of M to understand, (implicitly) define, and detect various types of antisemitism in event descriptions. We first provide a system prompt, where we inform M that it is evaluating and analyzing event descriptions.

M-NOCTX provides a baseline for **RQ1** by formulating d into a prompt that asks M for l, then c, then t. M-AS helps us answer **RQ3** and **RQ4** by modifying M-NOCTX to eliminate the binary classification task, only prompting the model for c and t, with "Other" options for both. If the model chooses "Other," we additionally ask it to generate its own c and t values.

M-AS-ICE has the same task presentation as M-AS, but it gives additional insight into **RQ4** by prepending one randomly selected entry corresponding to each potential value of t from the corpus to create a few-shot learning setting.[16]

Our next set of experiments probes M's understanding of the definition of antisemitism by comparing how its outputs differ from our first four experiments when we supply the definitions from Section 2.3 in the prompt (**RQ2**, **RQ4**). M-DEF and M-AS-DEF use the same task setup as M-NOCTX and M-AS, respectively, but we also supply the full definitions of each candidate for c and t, as well as Wikipedia's general definition of antisemitism.

Appendix A provides exact prompt templates.

3.2 Evaluation Metrics

For the appropriate experiments, we report **binary detection rate**, **accuracy**, **precision**, **recall**, **F1**, and a weighted modification of F1 (**WF1**).

For M-NOCTX and M-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 our overall dataset as well as for each category and type, where the rate for a category or type is computed among the entries whose gold label contains that category or type.

For all experiments, we report accuracy, precision, recall, F1 per category and per type as well as the means of all these scores across categories and across types. Additionally, since not all types and categories have equal frequency in the dataset, we compute a **WF1** metric, representing an F1 score weighted by the frequencies of categories or types within the gold labels of the corpus. Precise definitions of our scores are detailed in Appendix D.

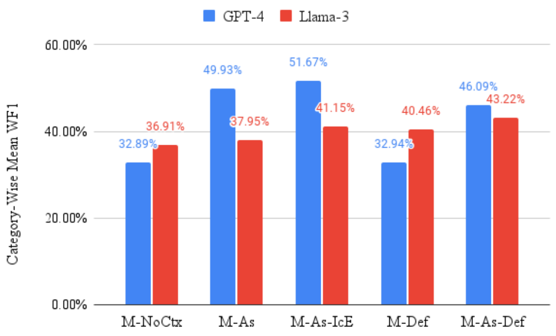

Figure 2: Category-wise mean WF1 scores for each experimental setup. With AS, GPT-4 has higher scores than Llama-3, but the opposite is true without AS.

3.3 Ensuring Validity of Classifier Setups

To ensure that our classifiers did not just default to classifying all event descriptions as antisemitic, we generate a control dataset of positive news stories about Jewish people and communities with inspiration from the methodology of Hartvigsen et al. (2022), and we verify that this set achieves a 0% false positive rate on the M-NOCTX and M-DEF setups, indicating a valid classifier. Full details on control set generation are given in Appendix E.

4 Results

Figures 2 and 3 summarize the category- and type-level mean WF1 scores of each model and experimental setup. In the following sections, we break the results down by research question and examine patterns for particular types.

4.1 Zero-Shot Baseline Performance (RQ1)

To investigate **RQ1**, we examine the results of M-NOCTX. Overall, we find that neither model achieves high WF1 scores across the board (mean of 32.89% across categories for GPT-4, 36.91% for Llama-3). From binary and coarse-grained perspectives, GPT-4 is less aligned with AMCHA's gold labels than Llama-3, but GPT-4 is more aligned than Llama-3 when it comes to fine-grained labels (28.16% mean WF1 across types for GPT-4 vs. 25.52% for Llama-3).

[16] In the interest of time, compute resources, and little observed benefit of adding examples, we do not perform the few-shot experiment with the task setup of M-NOCTX.

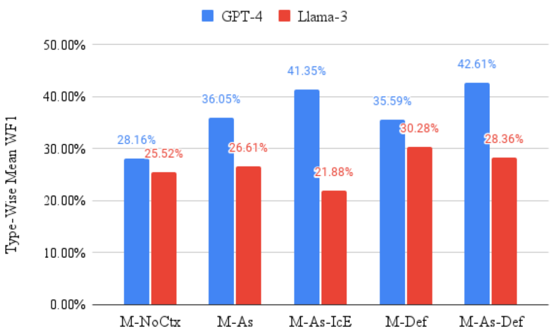

Figure 3: Type-wise mean WF1 scores for each experimental setup. GPT-4 has higher scores for fine-grained classification across the board.

Additionally, categories and types that are (a) particularly contentious or (b) heavily reliant on historical knowledge have the least alignment with AMCHA's labels. **BDS** has the worst detection rate (1.62% for GPT-4, 33.87% for Llama-3), and the types relating to criticism of Israel that cross the line into antisemitism according to AMCHA labels (e.g. **A**-*Dem*, **B**-*Cal*, **B**-*Eve*, **T**-*Sup*, **B**-*Vot*) also have significantly lower detection rates than the less contested types. In addition to these contested types, **T**-*Den* and **T**-*Gen* also have lower WF1 scores, indicating that GPT-4 detects many of those examples as antisemitic but does not properly categorize them.

Across the board, recall is lower than precision, except for **T**-*Den* and **T**-*Des*, indicating that GPT-4 mistook incidents involving historical tropes as **T**-*Des* and mistakes incidents targeting institutions or organizations as incidents targeting and denigrating individuals, possibly suggesting the need for more infusion of historical knowledge that would help differentiate these types from each other. For Llama-3, the lowest WF1 scores on top of the BDS-related types are for **A**-*His*, **T**-*Bul*, and **T**-*Sup*. For each of these types, the precision is significantly higher than the recall, indicating that Llama-3 correctly detects many examples as antisemitic but mistakes them for **B**-*Cal*, **T**-*Den*, or **T**-*Des* (the three types with higher recall than precision). We visualize the WF1 scores by type for both Llama-3 and GPT-4 in Figure 8.

4.2 Steering LLMs with Definitions of Antisemitism (RQ2)

When we add definitions of general antisemitism and its categories and types through M-Def, GPT-4's overall detection rate increases to **56.85%**, while Llama-3's decreases to **84.43%**, though this is still higher than GPT-4. We also observe slight increases in mean WF1 across categories in the setting without As (+2.32% for GPT-4, +2.05% for Llama-3) and slight increases in mean WF1 across types in the As setting (+6.56% for GPT-4, +1.75% for Llama-3). Llama-3 also shows additional improvements in category mean WF1 in the As setting and type mean WF1, but GPT-4 does not. Looking at confusion matrices by category, adding definitions shifts many predictions from "not antisemitic" to **BDS**.

Without As, both models' WF1 significantly improve for **B**-*Cal* and **T**-*Bul*, with Llama-3's WF1 also improving for **T**-*Den* and **T**-*Dis*. However, for Llama-3, the WF1 significantly decreases for **T**-*Sup* and **T**-*Des*. In the As setting, WF1 significantly increases for **A**-*His* and **T**-*Gen* (as well as **T**-*Des* for Llama-3 only) but significantly decreases for **A**-*Dem* in GPT-4 and for **B**-*Eve*, **B**-*Vot*, **A**-*Ter*, and **T**-*Phy* for Llama-3. As the examples in Table 1 highlight, adding these definitions corrects several cases of **A**-*His* and **T**-*Gen* that GPT-4 initially mistakes for **T**-*Den* or **T**-*Des*, indicating that adding definitions that point to historical knowledge helps both models—but especially GPT-4—operationalize that knowledge in their classification predictions. Figure 9 shows GPT-4's WF1 scores by type in M-As and M-As-Def.

These results suggest that GPT-4 can be steered to work with a particular definition of antisemitism to a limited extent, and so can Llama-3, but to an even lesser extent and only for certain historical context-heavy categories. Thus, it is useful for social scientists, policymakers, and other stakeholders to continue advocating for thorough and precise definitions of types of antisemitism that can be successfully detected by LLMs.

4.3 Changing the Premise of the Task: Removing the Binary Classification Question (RQ3)

Next, we examine **RQ3** by comparing M-noCtx with M-As and M-Def with M-As-Def to see the effect of instructing M to presuppose that the events described are antisemitic.

On average, presupposing that events are antisemitic improves coarse-grained detection for both models (+17.05% mean category WF1 for GPT-4 and +1.05% for Llama-3 without Def, +13.15% for GPT-4 and +2.75% for Llama-3 with Def). Average effects are visualized in Figure 4.

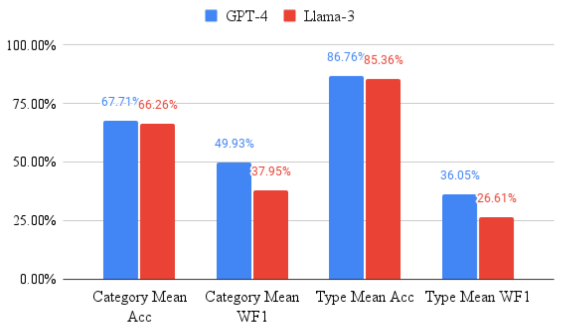

Figure 4: Category- and type-level mean WF1 and accuracy scores in the **M-As** setup. While accuracies are similar, accounting for class imbalance reveals higher WF1 scores for GPT-4. Disaggregated scores can be found in Appendix G.

By type, GPT-4 has higher WF1 scores on **A**-*Dem*, **B**-*Cal*, **B**-*Eve*, **T**-*Sup*, and **B**-*Vot* in both comparisons. Without Def, GPT-4 performs slightly worse on **A**-*His*, while the settings with Def see no significant decreases in WF1s by type.

For Llama-3, WF1 scores improve for **B**-*Vot* and **B**-*Cal* and decrease for **A**-*His* and **T**-*Phy*. The WF1 for **T**-*Bul* also decreases with Def. In general, the WF1 decreases are more significant from in the Def setup. The WF1 scores for M-As are displayed in Figure 10. Overall, while alignment increases at a coarse-grained level by removing the binary question, result are mixed and include significant WF1 drops at the fine-grained level.

4.4 Effects of In-Context Learning with Examples of Antisemitic Events (RQ4)

To answer **RQ4**, we investigate the results of M-As-IcE compared to M-As. The coarse-grained effect is a slight increase in category-wise mean WF1 (+1.74% for GPT-4, +3.2% for Llama-3). However, we observe from a confusion matrix that predictions overall shift more to **Expr** and away from **Targ**, indicating that few-shot learning biases models against predicting **Targ**.

Breaking the results down by type, we observe significant WF1 increases for **A**-*His*, **B**-*Eve*, and **B**-*Vot* and moderate decreases for **T**-*Gen* and **T**-*Sup* on GPT-4. For Llama-3, the fine-grained type-level alignment is significantly worse across the board, especially as demonstrated by lower WF1s for **A**-*Dem*, **B**-*Vot*, and **T**-*Phy*, showing that few-shot learning significantly hurts fine-grained classification alignment and that adding definitions seems like the preferable route for steering LLMs to be more effective on the AMCHA taxonomy. Figure 11 compares the WF1s from GPT-4 by type between M-noCtx and M-As-Def, showing the most dramatic overall difference in performance among all of our setups.

4.5 Experiments on Lighter-Weight Toxicity Classifiers

We additionally present the event descriptions from the AMCHA Corpus to the Perspective API[17] and to HateBERT (Caselli et al., 2021) to assess whether a more lightweight tool can be used to perform the preliminary binary classification task before we prompt an LLM for the multi-class and multi-label tasks. However, the mean Perspective API toxicity score for these entries is only 13%, with the highest-rated type only scoring a 27% mean. Additionally, HateBERT gives a binary "toxic" prediction on 0% of entries. This suggests that the event classification domain is different than the hate speech detection domain that HateBERT and Perspective are specialized for, necessitating LLMs or supervised in-domain classifiers.

5 Related Work

5.1 Detecting Antisemitism

Only a small subset of the literature on toxicity detection tackles antisemitism explicitly, and even fewer examine fine-grained subtypes of antisemitism. Most of these works focus on detecting online posts that are antisemitic for the purposes of online content moderation. Notable early work on antisemitism detection includes Warner and Hirschberg (2012), who work with a 9000-entry binary hate speech detection dataset sourced from American Jewish Congress-provided links to sites that may contain antisemitism, of which 90 are labeled antisemitic. Other domains in which binary antisemitism detection has been explored include 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), news sites like YouTube

[17] https://perspectiveapi.com/

(Khorramrouz et al., 2023; Barna and Knap, 2021), and synthetically generated text (Hartvigsen et al., 2022; Khorramrouz et al., 2023). To the best of our knowledge, our work is the first to examine antisemitism computationally from the perspective of hateful event description understanding rather than hate speech detection.

5.2 Detecting Other Forms of Toxicity and Creating Taxonomies of Hate

A few works have built some forms of fine-grained taxonomies of hate speech. 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, thus far, no work has examined LLMs' ability to operationalize fine-grained definitions of toxicity and detect when those definitions apply in event reports, especially for the specific case of antisemitism with a fine-grained taxonomy.

6 Conclusion and Discussion

In this work, we studied LLMs' abilities to detect fine-grained harmful event types in the context of antisemitism. We extracted and released the AMCHA Corpus, a set of real-world event descriptions pertaining to 3 categories and 14 types of antisemitism. We used the AMCHA Corpus to conduct a novel investigation of GPT-4 and Llama-3's abilities to perform fine-grained classification for the task of antisemitic event understanding, experimenting with additional prompt-level interventions such as adding definitions and in-context examples and removing the binary question.

Our findings show that while Llama has generally higher binary detection rates and can be steered to improve alignment of coarse-grained categories, GPT-4 has higher WF1 scores and appears to be more steerable toward effective fine-grained classification in line with AMCHA's definitions. Specifically, WF1 scores **B**-*Cal*, **A**-*His*, **T**-*Bul*, **T**-*Gen*, **B**-*Eve*, and **B**-*Vot* were able to be steered to improve by > 20% in some setting. While **T**-*Bul* and **T**-*Gen* were also steerable in Llama-3, it had much more concerning WF1 decreases in other categories that likely outweigh these improvements. We also observe that definitions tend to improve WF1 scores more than in-context examples. Our findings suggest that LLMs show promise as a tool that can be adapted to understand harmful events at scale and contribute to decreasing the human burden of exposure to distressing news and better grasping real-world manifestations of harm toward marginalized communities.

Future work Despite these improvements through definitions and in-context examples, we also observe that both models have significant room for improvement in the classification task. Future work should be done to improve both models' recall of BDS-related incidents and to improve their ability to differentiate between targeting individuals and targeting institutions or organizations, perhaps by creating a taxonomy that separates individual vs. institutional harm types at the categorical level. Overall, the fine-grained approach to this classification task can help resolve disputes about forms of antisemitism that are more contested. 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 actively represent different annotators' stances as in Deng et al. (2023).

7 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. 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 M-AS-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 the AMCHA Initiative's taxonomy. 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 from US-based campuses, 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.
8. Transforming events into finite text descriptions leads to loss of grounded or embodied event information. Future work could explore the classification of raw news articles describing events or explore automatically extracting event descriptions from raw news articles.

8 Ethics Statement

Environmental Statement: Our 7 experiments (5 with AMCHA Corpus, 2 for control data) cost $873.13 in total through Microsoft Azure's OpenAI API.[18] For Llama-3, we use the free Groq inference service.[19] As we use APIs for inference, our experiments only used CPUs. GPT-4 experiments took approximately 10 days, while Llama-3 experiments took approximately 4 days. Llama-3 has 8 billion parameters, while GPT-4's parameter count is undisclosed. We use `pandas`[20] to load our corpus and use `scikit-learn`[21] and `matplotlib`[22] 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. In particular, we acknowledge disagreements about whether opposition to Israel constitutes antisemitism, specifically as it relates to the inclusion of the BDS movement in our study. See Appendix F for an in-depth discussion on these disagreements.

Positionality Statement: The authors have diverse perspectives on and connections to antisemitism, including US- and Israel-raised Jewish, Egyptian-born Muslim, 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*.

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.

M Muhannad Ayyash. 2023. The toxic other: The palestinian critique and debates about race and racism. *Critical Sociology*, 49(6):953–966.

[18] https://azure.microsoft.com/en-us/products/ai-services/openai-service

[19] https://groq.com/

[20] https://pandas.pydata.org/

[21] https://scikit-learn.org/stable/

[22] https://matplotlib.org/stable

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.

Abigail B Bakan and Yasmeen Abu-Laban. 2024. Anti-palestinian racism, antisemitism, and solidarity: considerations towards an analytic of praxis. *Studies in Political Economy*, 105(1):107–122.

Omar Barghouti. 2021. Bds: Nonviolent, globalized palestinian resistance to israel's settler colonialism and apartheid. *Journal of Palestine studies*, 50(2):108–125.

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.

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

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.

FBI. 2023. 2022 FBI hate crimes statistics. https://www.justice.gov/crs/highlights/2022-hate-crime-statistics. Accessed: 2024-5-13.

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*.

Jennifer Hitchcock. 2023. Framing palestinian rights: A rhetorical frame analysis of vernacular boycott, divestment, sanctions (bds) movement discourse. *Rhetoric Society Quarterly*, 53(2):87–103.

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.

Rashid I Khalidi. 2021. The journal of palestine studies in the twenty-first century: An editor's reflections. *Journal of Palestine studies*, 50(3):5–17.

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

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.

United States Holocaust Memorial Museum. Antisemitism in history: From the early church to 1400.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

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.

Rebecca Ruth Gould. 2020. The ihra definition of antisemitism: defining antisemitism by erasing palestinians. *The Political Quarterly*, 91(4):825–831.

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.

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.

Dov Waxman, David Schraub, and Adam Hosein. 2022. Arguing about antisemitism: why we disagree about antisemitism, and what we can do about it. *Ethnic and Racial Studies*, 45(9):1803–1824.

Ben White. 2020. Delegitimizing solidarity: Israel smears palestine advocacy as anti-semitic. *Journal of Palestine Studies*, 49(2):65–79.

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

A LLM Prompts

Example user prompts for each setting can be found in the `experiments/event_discovery/log.txt` file at `https://anonymous.4open.science/r/antisemitism-event-understanding-ECF0/`. Our system prompt and exact user prompt formulation function can be found in experiments/event_discovery/utils.py within the same repository.

B Control Data Generation

Seed phrases for generation of control data, as well as their corresponding dates and a list of universities tracked by the AMCHA Initiative, 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}."
```

C Source Collection for AMCHA Corpus

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 the typology presented in Section 2.3.

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[23] 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.

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 Control Data Generation

One error we could not detect with the AMCHA Corpus alone was that of false positives at the binary level, as all incidents logged in the dataset

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

are labelled as antisemitic. To assess false positive rates on this task in GPT-4 and Llama-3, we run M-noCtx on a GPT-4-generated control set 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.
2. For each university U in the AMCHA Initiative's list of universities tracked for incidents:
   (a) For each seed phrase P and corresponding date d, we ask GPT-4 to generate a short, factual news article about the event described by P occurring on d at U. We create six such generations per tuple of $((P, d), U)$, amounting to $6|\{P\}| \cdot |\{U\}|$ entries in total.
3. We take a random sample of this generated set that is equal in size to the AMCHA Corpus used for our experiments.

The results of M-noCtx on this control set 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.

F Further Discussion on Disagreements Over AMCHA's Taxonomy

We use a taxonomy directly from the AMCHA Initiative, who categorize events based on the International Holocaust Remembrance Alliance (IHRA) definition of antisemitism. While the IHRA definition does not consider criticism of Israel to be inherently antisemitic, it states that ”denying the Jewish people their right to self-determination” and ”using the symbols and images associated with classic antisemitism (e.g., claims of Jews killing Jesus or blood libel) to characterize Israel or Israelis” are antisemitic. Based on these points, the AMCHA Initiative argues that the BDS movement is antisemitic because it ”aims to demonize, delegitimize, and destroy the Jewish nature of Israel, with the result of denying to Jews their right of national self-determination”, and that the movement invokes ”classic antisemitic tropes of Jewish evil, power, and mendacity”.[24]

However, both the IHRA definition and AMCHA's characterization of BDS as antisemitic is rejected by many Jewish and Palestinian advocacy organizations. The Jerusalem Declaration on Antisemitism (JDA, 2021) and Nexus Document (Nexus, 2023) specifically assert that BDS and other non-violent forms of political protest against the State of Israel or its policies are not in and of themselves antisemitic. The BDS movement self-identifies as inclusive and in opposition to all forms of racism including antisemitism.[25] They emphasize that the movement does not target Israeli or Jewish individuals on the basis of their identities. Rather, the movement targets the Israeli state, companies, and institutions based on international law and human rights violations. Both the BDS organization and its members refute claims of antisemitism by issuing condemnations of antisemitic incidents and highlighting Jewish support for BDS, especially through organizations such as Jewish Voice for Peace (Hitchcock, 2023).

Some scholars argue that framing BDS as antisemitic delegitimizes Palestinian solidarity movements (White, 2020; Bakan and Abu-Laban, 2024), stifles freedom of expression in support of Palestine (Ruth Gould, 2020; Bakan and Abu-Laban, 2024), suppresses academic scholarship about Palestine (Khalidi, 2021), and ultimately contributes to an erasure of Palestinian experiences (Ruth Gould, 2020; Ayyash, 2023). According to some scholars, the Israeli state and pro-Israel organizations call BDS antisemitic in order to neutralize the strategic challenge that BDS presents to the state and avoid taking accountability in policies that harm Palestinians (White, 2020; Barghouti, 2021). Barghouti, a co-founder of the BDS movement, asserts that just as ”there is nothing Jewish about Israel's regime of occupation, siege, ethnic cleansing, and apartheid, there is nothing inherently anti-Jewish, then, about a nonviolent, morally consistent human rights struggle to end this system of oppression” (2021). By focusing on BDS and related pro-Palestine activism in discourse about antisemitism, some believe that efforts to combat antisemitism

[24] https://amchainitiative.org/BDS-background/

[25] `https://bdsmovement.net/faqs#collapse16241`

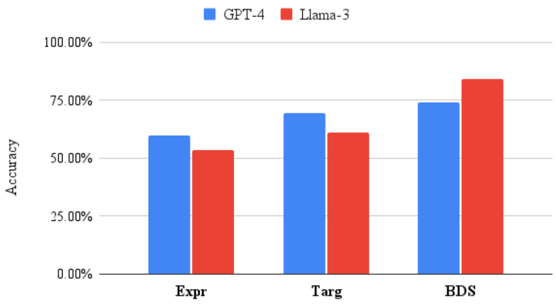

Figure 5: Category-level accuracy scores on **M-As**. While GPT-4 is more accurate on **Expr** and **Targ**, Llama-3 shows greater alignment with **BDS**.

risk neglecting rising threats from far-right white supremacists (White, 2020; Ayyash, 2023).

Defining antisemitism has been and will continue to be contentious, even within Jewish communities. People disagree on where to identify antisemitism: in perpetrators' intent, victims' perceptions, objective outcomes in the world, or in discourse (e.g., invoking antisemitic tropes even if unintentionally) (Waxman et al., 2022). It is precisely due to these disagreements that focusing on fine-grained categories is a key strength of our work. As there is no single correct answer about what constitutes antisemitism, we need adaptable and explainable methods for automated antisemitism detection. Ultimately, practitioners must decide whether to include certain categories in their definitions of hate and hateful events. Some may find it useful to exclude BDS activity or detect it as a related but distinct category from antisemitism.

We caution that such disagreements should not be weaponized to disregard concerns of antisemitism or dismiss academic research on antisemitism. On the contrary, we urge the NLP community to assume that all parties arrive to these discussions in good faith, and to take perceptions of antisemitism seriously even if they are not conclusive (Waxman et al., 2022). We hope that computational work on antisemitism embraces such diversity of perspectives in future datasets, models, and research team compositions.

G Additional Plots

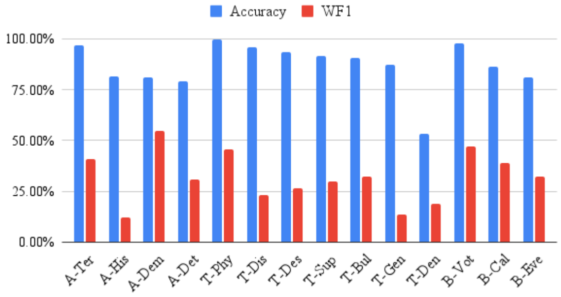

Figure 6: Type-level accuracy and WF1 scores on M-As for GPT-4. While accuracies are high due to class imbalance, WF1 scores are considerably lower, especially for **A**-*His*, **T**-*Gen*, and **T**-*Den*.

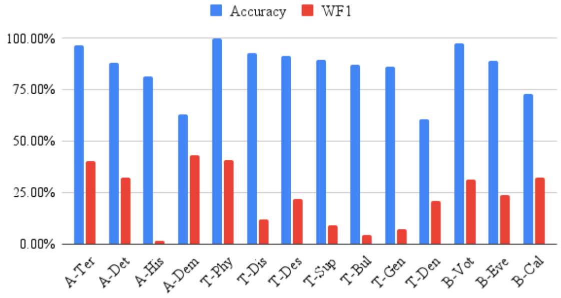

Figure 7: Type-level accuracies on **M-As** for Llama-3. While accuracy scores are high due to class imbalance, WF1 scores are especially low on **A**-*His*, **T**-*Gen*, and **T**-*Bul*, with the first two indicating a lack of Llama-3's ability to leverage domain knowledge that would reveal historical antisemitism or genocidal rhetoric.

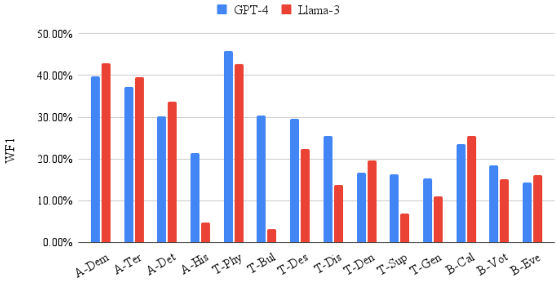

Figure 8: WF1 scores by fine-grained type on **M-noCtx**. GPT-4 has higher scores on a slight majority of types, though both models have low scores overall.

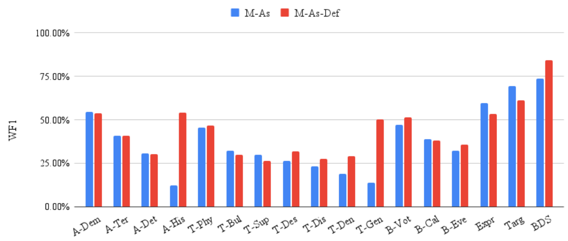

Figure 9: Comparison of WF1 scores by type on GPT-4 with vs. without adding precise definitions of our typology to the user prompt. Adding definitions improves WF1 scores for a majority of types.

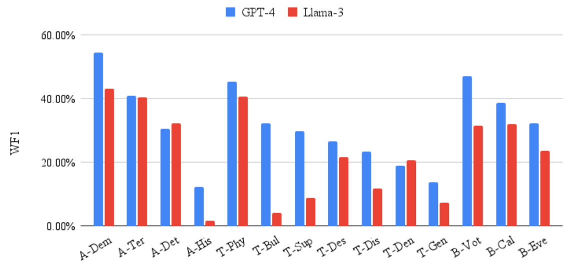

Figure 10: WF1 scores by type on GPT-4 and Llama-3 after changing the premise of the prompt to instruct the model to assume that the given text describes an antisemitic event (**M-AS**). GPT-4 shows more alignment than Llama-3 across the board.

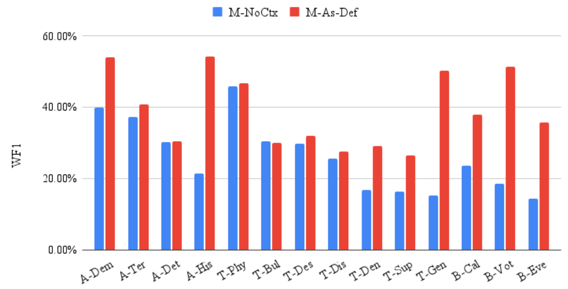

Figure 11: Type-level WF1 scores for M-NOCTX and M-AS-DEF for GPT-4. M-AS-DEF scores higher, showing promise for GPT-4's steerability.