

Improving Hate Speech Detection with Large Language Models: Supplementary Materials

Natalia Umansky¹, Maël Kubli¹, Ana Kotarcic², Laura Bronner³, Selina Kurer³, Philip Grech³, Dominik Hangartner³, Fabrizio Gilardi¹, and Karsten Donnay¹

¹Department of Political Science, University of Zurich

²Department of History, University of Zurich

³Immigration Policy Lab, ETH Zurich

1 GPT-4o-mini Classification Pipeline

In this study, we used OpenAI’s ChatGPT (API model 4o-mini-2024-07-18) to perform hate speech classification. The process is grounded in a predefined codebook, which serves as a fundamental guide for ChatGPT’s analysis and categorization of comments based on specific themes.

1.1 Codebook and classification Prompt

The codebook encompasses a set of categories or themes against which ChatGPT was tasked to evaluate each comment. The model was presented with a German language prompt for classification, we here provide an English translation; the original German prompt can be found in Appendix 3. The system prompt is structured as follows for the zero-shot classifications:¹

¹Note that we prompted the GPT-4o-mini classification with language referring to “annotation.” This is a deliberate choice of terminology because we wanted to contextualize the LLM to closely emulate the task of a human coder.

Instructions:

You will read comments posted to articles on the Internet and determine whether they are hate speech or not. There is no universal definition of hate speech. A widely accepted definition, also used by the United Nations, defines hate speech as the use of language that attacks or devalues a person or group based on their identity - e.g. based on their gender, age, sexuality, religion, nationality, skin color or origin, mental or physical impairment. A broader definition of hate speech can also include derogatory language against a person or group based on their social status (e.g. education or income) or their political views. In addition, there is toxic/hateful language that, while not necessarily hate speech, nevertheless affects the culture of discussion and promotes hostility. This includes general insults, threats or vulgar statements that are not directed against a person or group based on their identity.

Steps:

1. Read the comment.
2. If a comment does not contain hate speech, please answer question 1 with 'False'. If a comment contains hate speech, please answer question 1 with 'True'.
3. Only if the comment contains hate speech, please also answer question 2, in which you indicate which group(s) of people the comment is directed against.

Question 1 (label):

Classify the comments as follows:

- True - The comment contains hate speech
- False - The comment does not contain hate speech

Question 2 (target):

If the comment contains hate speech or toxic language, indicate which target group the hate speech is directed against:

- **Gender:** The hate speech is directed against gender.
- **Age:** The hate speech is directed against age.
- **Sexuality:** The hate speech is directed against sexuality.
- **Religion:** The hate speech is directed against religion.
- **Nationality:** The hate speech is directed against nationality, skin color, or origin.
- **Disability:** The hate speech is directed against mental or physical disability.
- **Toxic:** Does not target any specific group, but uses toxic language.
- **Status:** The hate speech is directed against social status, education, income, or professional group.
- **Politics:** The hate speech is directed against political beliefs.
- **Appearance:** The hate speech is directed against appearance or physical attributes.
- **Other:** The hate speech is directed against something else.

Format for the output:

Always specify the result in JSON format, do not specify anything else. Here is an example of the format you must use:

```
{"Label": "<LABEL>", "reason": "Reasoning or evidence for the <LABEL> and <TARGET> chosen."}
```

This prompt guides the model through the classification process, instructing it to discern whether a comment constitutes hate speech, and, if so, to identify the specific target group affected by it, including if it is toxic or not. For the fine-tuning it is not possible to easily evaluate the fine-tuning with a two-stage answering process. Hence, we

opted for a slight change in procedure, which differs from the instruction set seen by the human annotator groups, as well as from the zero-shot prompt seen by GPT-4o. Instead of tasking the LLM to label a comment as Hate Speech or Non Hate Speech and then asking it for the target group, or if it contains only toxic speech. We opted to ask the LLM directly if a given comment is Hate Speech, Toxic Speech, or neither of those two. Thus the system prompt is slightly different for the fine-tuning, but it is identical to the prompt used for the zero-shot model when evaluating performance, i.e., the model performance after fine-tuning can be directly compared to the zero-shot model. The modified prompt for fine-tuning looks as follows:

Instructions:

You will read comments posted to articles on the Internet and determine whether they are hate speech or not. There is no universal definition of hate speech. A widely accepted definition — also employed by the United Nations — defines hate speech as the use of language that attacks or devalues a person or group based on their identity. These characteristics and attributes are not limited to fixed categories but can refer to any properties by which people are characterized or described – for example, gender, age, sexuality, religion, nationality, skin color, origin, mental or physical impairment, social status, political views, appearance, or other relevant attributes. Any discriminatory statement directed at a person or group based on one or more of these or other relevant characteristics/attributes is to be classified as HATE SPEECH (1).

In addition, there is toxic/hateful language that, while not necessarily hate speech, nevertheless affects the culture of discussion and promotes hostility. Toxic Speech encompasses unsound, insulting, threatening, or vulgar expressions that have an aggressive effect without specifically referring to characteristics or attributes of a person or group. Such comments are to be classified as TOXIC SPEECH (2).

Steps:

1. Read the comment.
2. If a comment contains NEITHER Hate Speech NOR Toxic Speech, reply with NO HATE SPEECH (0). If a comment devalues a person or group based on certain characteristics or attributes, reply with HATE SPEECH (1). If a comment contains demeaning expressions that disrupt the discussion atmosphere or are hostile, but are not to be classified as targeted discrimination against persons or groups based on their characteristics or attributes (Hate Speech), please classify it as TOXIC SPEECH (2).

Question 1 (label):

Classify the comments as follows:

- HATE SPEECH – The comment contains hate speech
- TOXIC SPEECH – The comment contains toxic speech
- NO HATE SPEECH – The comment contains neither hate speech nor toxic speech

1.2 Operational Procedure

The models individually processed each comment and then classified it based on the themes and the conceptual rules defined in the codebook. Specifically, classification entails feeding a comment as input to the model along with the codebook. In this context, the codebook served as a priming sequence to direct the model’s attention towards the desired facets of the comment text, essentially, the attributes we wished to encode. Primed with the themes and the rules from the codebook, ChatGPT-4o-mini Turbo then generated an output representing the annotated comment (in the format we specified in the prompt). We utilized the standard temperature setting for ChatGPT for classification (0.2) for the zero-shot model.

2 Model Fine-Tuning

The zero shot GPT-4o-mini model was fine-tuned using the same stratified sample of 100 and 250 comments for each annotator group, maintaining the (gold standard) label distribution characteristics of the original data set. This stratified approach guaranteed that both the 100-comment and 250-random comment subsets faithfully represented the label proportions observed in the full dataset. We utilized OpenAI’s API to fine-tune the GPT-4o-mini model providing the two respective samples as training data respectively.

In the fine-tuning step, we configured our model to run for five epochs ($n_{epochs} = 5$), batch size to 20, and set the temperature parameter as 0.0 ($temp = 0.0$) instead of 0.2. The temperature parameter directly influences the randomness of the model’s output, with our specified value promoting a more deterministic response. Thus, setting this value to the absolute minimum reduces variability to a minimum, which allows us to run the classification only once since the results should be equal in almost all cases over multiple runs. We are able to assume such model behavior as we already encountered very low variability in the zero-shot classifications, which we ran three times (Krippendorff’s α 0.948 for Hate Speech and 0.855 for Toxic Speech). A single epoch constitutes a full pass of our training data through the model. In simple words, during an epoch, the model weights are fine-tuned to minimize the difference between the predicted and actual class labels. It is essential to note that this process was undertaken independently for both the 100 and 250-row datasets, forming two distinct fine-tuned models for each annotation group.

The complete fine-tuning pipeline is as follows. First, we pre-processed the training dataset following the template provided by OpenAI.² After pre-processing, we uploaded the data to the API and initialized the model with the pre-trained weights. In the second step, we provided the model with instructions for our particular classification task. These instructions are given in a.json-friendly format that the language model understands. This also guides the model’s understanding and learning, in alignment with our classification

²Template: <https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset>.

aims. The instructions comprise both system-level and user-level prompts. The system-level prompt offers overarching guidance for the entire fine-tuning process. It outlines the task at hand: reading comments from internet articles and determining whether they contain Hate Speech or Not Hate Speech. In our case, the system prompt thus contains the complete codebook (see Appendix 1.1).

The user-level prompt offers granular instructions for handling individual instances within the dataset. It directs the model on processing and interpreting each comment, guiding its classification decisions. It reads as follows for our task (we here provide an English translation; the original German prompt can be found in Appendix 3):

User prompt:

Task:

Step 1: Read the comment..

Step 2: If a comment does not contain Hate Speech or Toxic Speech, respond with NO HATE SPEECH (0).

If a comment devalues a person or group based on specific characteristics or attributes, respond with HATE SPEECH (1).

If a comment contains derogatory statements that disrupt the discussion atmosphere or are hostile but do not qualify as targeted discrimination against individuals or groups based on their characteristics or attributes (Hate Speech), classify it as TOXIC SPEECH (2).

Text to classify: {text}

By incorporating both system- and user-level prompts, we ensure that the model receives comprehensive guidance throughout the fine-tuning process. The system-level prompt sets the context and provides overarching guidance, while the user-level prompt offers specific instructions for handling individual instances, ultimately guiding the model’s understanding and aligning it with our classification aims.

In the last step, GPT-4o-mini is fine-tuned through multiple epochs of iterative training during which the model dynamically adjusts its parameters based on the training data. The fine-tuned model is then evaluated using the original user prompt.

3 German Prompt Texts

The following corresponds to the German language ChatGPT prompt provided in its English translation in Appendix 1.1.

3.1 Codebook and zero shot classification prompt

Anleitung:

Sie werden Kommentare zu Artikeln aus dem Internet lesen und bestimmen, ob es sich um Hate Speech handelt oder nicht. Es gibt keine allgemeingültige Definition von Hate Speech. Eine weit verbreitete – auch von den Vereinten Nationen verwendete – Definition versteht unter Hate Speech den Gebrauch von Sprache, durch den eine Person oder eine Gruppe aufgrund eines Merkmals oder einer Eigenschaft angegriffen oder abgewertet wird. Merkmale und Eigenschaften umfassen dabei nicht nur fest definierte Kategorien, sondern können sich auf jegliche Eigenschaften beziehen, anhand derer Menschen charakterisiert oder beschrieben werden – beispielsweise Geschlecht, Alter, Sexualität, Religion, Nationalität, Hautfarbe, Herkunft, geistige oder körperliche Beeinträchtigung, sozialer Status, politische Einstellung, Aussehen oder andere relevante Merkmale. Zusätzlich gibt es toxischen/hasserfüllten Sprachgebrauch (Toxic Speech), der zwar nicht notwendigerweise Hate Speech im engeren Sinne ist, aber die Diskussionskultur beeinträchtigt und Feindseligkeit fördert. Toxic Speech umfasst unsachliche, beleidigende, bedrohliche oder vulgäre Äußerungen, die aggressiv wirken, ohne sich gezielt auf Merkmale oder Eigenschaften einer Person oder Gruppe zu beziehen.

Schritte:

1. Lesen Sie den Kommentar.
2. Wenn ein Kommentar KEINE HATE SPEECH enthält, beantworten Sie Frage 1 bitte mit 'False'. Wenn ein Kommentar HATE SPEECH oder TOXIC SPEECH enthält, beantworten Sie die Frage 1 bitte mit 'True'.
3. Nur wenn der Kommentar HATE SPEECH oder TOXIC SPEECH enthält, beantworten Sie bitte auch Frage 2, in der Sie angeben, gegen welche Personen-Gruppe(n) sich der Kommentar richtet oder im Falle von toxischer Sprache wählen sie Toxisch.

Frage 1 (Label):

Klassifiziere die Kommentare folgendermassen:

- True - Der Kommentar enthält Hate Speech oder Toxic Speech
- False - Der Kommentar enthält keine Hate Speech und auch keine Toxic Speech

Frage 2 (target):

Falls der Kommentar Hate Speech oder Toxische Sprache enthält gib an gegen welche Zielgruppe sich die Hatespeech richtet:

- **Geschlecht** - Die Hate Speech richtet sich gegen das Geschlecht
- **Alter** - Die Hate Speech richtet sich gegen das Alter
- **Sexualität** - Die Hate Speech richtet sich gegen die Sexualität
- **Religion** - Die Hate Speech richtet sich gegen die Religion
- **Nationalität** - Die Hate Speech richtet sich gegen die Nationalität oder Hautfarbe oder Herkunft
- **Beeinträchtigung** - Die Hate Speech richtet sich gegen die geistliche oder körperliche Beeinträchtigung

- **Toxisch** - Richtet sich gegen keine Gruppe, benutzt aber toxische Sprache
- **Status** - Die Hate Speech richtet sich gegen den sozialen Status, Bildung, Einkomme oder Berufsgruppe
- **Politik** - Die Hate Speech richtet sich gegen die politische Einstellung
- **Aussehen** - Die Hate Speech richtet sich gegen das Aussehen oder körperliche Merkmale
- **Andere** - Die Hate Speech richtet sich gegen etwas anderes

Format:

Gib das Ergebnis immer im JSON-Format an, gib nichts anderes an. Hier ist ein Beispiel für das Format, das du verwenden musst:

```
{"Label": "<LABEL>", "reason": "Reasoning or evidence for the <LABEL> and <TARGET> chosen."}
```

3.2 Fine-Tuning System prompt

Anleitung:

Sie werden Kommentare zu Artikeln aus dem Internet lesen und bestimmen, ob es sich um Hate Speech handelt oder nicht. Es gibt keine allgemeingültige Definition von Hate Speech. Eine weit verbreitete – auch von den Vereinten Nationen verwendete – Definition versteht unter Hate Speech den Gebrauch von Sprache, durch den eine Person oder eine Gruppe aufgrund eines Merkmals oder einer Eigenschaft angegriffen oder abgewertet wird. Merkmale und Eigenschaften umfassen dabei nicht nur fest definierte Kategorien, sondern können sich auf jegliche Eigenschaften beziehen, anhand derer Menschen charakterisiert oder beschrieben werden – beispielsweise Geschlecht, Alter, Sexualität, Religion, Nationalität, Hautfarbe, Herkunft, geistige oder körperliche Beeinträchtigung, sozialer Status, politische Einstellung, Ausse-

hen oder andere relevante Merkmale. Jede diskriminierende Aussage, die sich gegen eine Person oder Gruppe auf Basis einer oder mehrerer dieser oder anderer relevanter Merkmale/Eigenschaften richtet, ist als HATE SPEECH (1) zu klassifizieren.

Zusätzlich gibt es toxischen/hasserfüllten Sprachgebrauch (Toxic Speech), der zwar nicht notwendigerweise Hate Speech im engeren Sinne ist, aber die Diskussionskultur beeinträchtigt und Feindseligkeit fördert. Toxic Speech umfasst unsachliche, beleidigende, bedrohliche oder vulgäre Äußerungen, die aggressiv wirken, ohne sich gezielt auf Merkmale oder Eigenschaften einer Person oder Gruppe zu beziehen. Solche Kommentare sind als TOXIC SPEECH (2) zu klassifizieren.

Schritte:

1. Lesen Sie den Kommentar.
2. Wenn ein Kommentar KEINE Hate Speech und KEINE Toxic Speech enthält, antworten Sie mit KEINE HATE SPEECH (0).
3. Wenn ein Kommentar eine Person oder Gruppe aufgrund bestimmter Merkmale oder Eigenschaften abwertet, antworten Sie mit HATE SPEECH (1).
4. Wenn ein Kommentar abwertende Äußerungen enthält, die die Diskussionsatmosphäre stören oder feindselig sind, aber nicht als gezielte Diskriminierung von Personen oder Gruppen aufgrund ihrer Merkmale oder Eigenschaften (Hate Speech) einzuordnen sind, klassifizieren Sie ihn bitte als TOXIC SPEECH (2).

Frage (Label):

Klassifiziere die Kommentare folgendermassen:

- HATE SPEECH - Der Kommentar enthält Hate Speech

- TOXIC SPEECH - Der Kommentar enthält Toxic Speech
- KEINE HATE SPEECH - Der Kommentar enthält keine Hate Speech und keine Toxic Speech

And the following is the additional user-level prompt employed during model fine-tuning provided in its English translation in Appendix 2:

User Prompt:

Aufgabe

Schritt 1: Lesen Sie den Kommentar.

Schritt 2: Wenn ein Kommentar KEINE Hate Speech und KEINE Toxic Speech enthält, antworten Sie mit KEINE HATE SPEECH (0). Wenn ein Kommentar eine Person oder Gruppe aufgrund bestimmter Merkmale oder Eigenschaften abwertet, antworten Sie mit HATE SPEECH (1). Wenn ein Kommentar abwertende Äußerungen enthält, die die Diskussionsatmosphäre stören oder feindselig sind, aber nicht als gezielte Diskriminierung von Personen oder Gruppen aufgrund ihrer Merkmale oder Eigenschaften (Hate Speech) einzuordnen sind, klassifizieren Sie ihn bitte als TOXIC SPEECH (2).

Text zum klassifizieren:

{text}

4 GPT-4o-mini Classification Results

4.1 Confusion Matrices

This section presents the confusion matrices for GPT-4o-mini across different fine-tuning datasets, providing a comparative overview of classification performance. The results illustrate how dataset selection and prompt variations influence true positives, false positives, and false negatives in distinguishing between hate speech, toxic speech, and

non-hate speech categories. Bold values in the tables highlight the highest performance metrics, emphasizing the impact of fine-tuning strategies on model accuracy.

Table 1: Performance comparison of GPT-4o-mini across various fine-tuning datasets, highlighting the confusion matrix outcomes of the classifications with two classes (Hate Speech, No Hate Speech), where Toxic is part of Hate Speech. Bold values denote the highest values achieved in each category, showcasing the impact of dataset selection on model accuracy. Note: GPT-4o (A) refers to the model supplied with the prompt identical to that used by human annotators; GPT-4o (B) is using an adjusted prompt where the target group is no longer explicitly asked for, resulting in a single-stage classification for toxic, hate speech, and neither as used for fine-tuning; and GPT-4o (C) employs the same prompt as (A) but incorporates a short UN definition of Hate Speech.

Group		Fine-Tuning	Size	TP	TN	FP	FN
GPT-4o (A)				226	150	106	12
GPT-4o (B)				94	237	19	144
GPT-4o (C)				180	194	62	58
GPT-4o	Appen	✓	100	190	172	84	48
GPT-4o	Appen	✓	250	153	196	60	85
GPT-4o	Citizen Science	✓	100	203	168	88	35
GPT-4o	Citizen Science	✓	250	187	193	63	51
GPT-4o	NGO	✓	100	146	230	26	92
GPT-4o	NGO	✓	250	133	238	18	105
GPT-4o	Prolific	✓	100	168	217	39	70
GPT-4o	Prolific	✓	250	173	218	38	65
GPT-4o	Research Assistants	✓	100	180	213	43	58
GPT-4o	Research Assistants	✓	250	199	216	40	39

Table 2: Performance comparison of GPT-4o-mini across various fine-tuning datasets, highlighting the flattened confusion matrix outcomes of the classifications with three classes (Hate Speech [1], Toxic Speech [2], No Hate Speech [0]). Note: GPT-4o (A) refers to the model supplied with the prompt identical to that used by human annotators; GPT-4o (B) is using an adjusted prompt where the target group is no longer explicitly asked for, resulting in a single-stage classification for toxic, hate speech, and neither as used for fine-tuning; and GPT-4o (C) employs the same prompt as (A) but incorporates a short UN definition of Hate Speech.

Group	Fine-Tuning	Size	TP (0)	FP (0)/FN (1)	FP (0)/FN (2)	FP (1)/FN (0)	TP (1)	FP (1)/FN (2)	FP (2)/FN (0)	FP (2)/FN (1)	TP (2)
GPT-4o (A)			150	4	8	77	125	31	29	15	55
GPT-4o (B)			112	3	5	19	83	11	125	58	78
GPT-4o (C)			194	13	45	61	131	44	1	0	5
GPT-4o	✓	100	172	32	16	5	43	32	79	69	46
GPT-4o	✓	250	196	66	19	22	59	25	38	19	50
GPT-4o	✓	100	168	14	21	86	130	63	2	0	10
GPT-4o	✓	250	193	19	32	63	120	48	0	5	14
GPT-4o	✓	100	230	44	48	21	90	14	5	10	32
GPT-4o	✓	250	238	54	51	14	74	17	4	16	26
GPT-4o	✓	100	217	39	31	39	105	63	0	0	0
GPT-4o	✓	250	218	36	29	38	108	65	0	0	0
GPT-4o	✓	100	213	30	28	40	111	33	3	3	33
GPT-4o	✓	250	216	19	20	36	119	33	4	6	41

4.2 Classification Errors

We also investigated patterns in the classification errors in further details. Table 3 reports false positives and false negatives for each annotation group and fine-tuning size. As we increase the training set from 100 to 250 examples, the average false positive rate decreases from 21.9% at 100 examples to 17.1% at 250 examples (a drop of 4.8 pp). This is due to a notable decrease in false positive classifications for three groups (Appen, Citizen Science, NGO), the others see only a slight decrease (Prolific, Research Assistants).

At the same time, the average false-negative rate increases from 25.5% to 29.0%. This pattern reflects a classic precision–recall trade-off: the drop in false positives raises precision, but the simultaneous rise in false negatives lowers recall. In practical moderation terms, this means that fewer harmless comments are mistakenly removed, yet more hateful content may slip through, leaving the overall F-1 score largely unchanged. Most notably false negatives increase in three out of five groups with more fine-tuning examples (Appen, Citizen Science, NGO). The Prolific and Research Assistants classifiers alone see a decrease in false negatives when the number of fine-tuning examples is increased from 100 to 250. This further illustrates at the level of classification errors that only the two groups with the highest level of annotation quality can alleviate false negatives and, thus, increase overall performance. These error patterns mirror the precision–recall balance of the underlying label sets.

We might assume that these patterns of mismatched classifications are related to the type of hate speech, i.e., which group is targeted. As part of the gold standard annotation, our experts assigned robust target group labels that we can use to discover the source of false negative identifications specifically. Table 4 summarizes these patterns for all annotator groups and fine-tuning sizes. Every classifier misses more nationality-based hate speech comments than any other category. The fewest nationality misses occur for the model fine-tuned on Citizen Science, with 100 examples (11 comments). The most occur for the LLM fine-tuned with Appen at 250 examples (53 comments). Toxic speech represents the second-highest share of misclassifications. Its counts range from 16 misses

Table 3: False Positive and False Negative Rates by Classifier

Group	Fine-Tuning Size	FP	FN	FP Rate	FN Rate
Appen	100	84	48	32.8%	20.2%
Appen	250	60	85	23.4%	35.7%
Citizen Science	100	88	35	34.4%	14.7%
Citizen Science	250	63	51	24.6%	21.4%
NGO	100	26	92	10.2%	38.7%
NGO	250	18	105	7.0%	44.1%
Prolific	100	39	70	15.2%	29.4%
Prolific	250	38	65	14.8%	27.3%
Research Assistants	100	43	58	16.8%	24.4%
Research Assistants	250	40	39	15.6%	16.4%

for the model fine-tuned with 100 annotations from Appen to 51 misses for that fine-tuned on 250 annotations from the NGO. Other target groups, such as gender and politics, each account for at most nine misses per run. Target groups such as politics, age, religion, and “other” appear only much less frequently and overall decrease as we reach 250 fine-tuning examples.

Taken together these results overall suggest that adding in-group annotations can reduce both false positives and negatives shifting the remaining misses toward specific content. A qualitative, in-depth look at examples for false positive classifications suggests that the classifiers across the board struggle most with firmly worded, critical statements e.g., directed at specific politicians or societal groups. Those do not rise to the level of toxic or hate speech per our definition (and experts labels) but are structurally very similar in that they address a person or group specifically and voice targeted criticism. Nationality remains the toughest category and toxic speech also drives many of the remaining false negatives, while low-volume categories such as age and religion become essentially error-free in our example already with modest amounts of data.

Table 4: False Negative Classifications by Target Group for Fine-Tuned GPT-4o-mini

Group	Fine Tuning	Target	FN	Fine Tuning	Target	FN
Appen	100	age	1	250	age	1
Appen	100	nationality	26	250	nationality	53
Appen	100	politics	3	250	politics	3
Appen	100	gender	2	250	gender	7
Appen	100	religion	0	250	religion	0
Appen	100	status	0	250	status	1
Appen	100	other	0	250	other	1
Appen	100	toxic	16	250	toxic	19
Citizen Science	100	age	1	250	age	1
Citizen Science	100	nationality	11	250	nationality	13
Citizen Science	100	politics	1	250	politics	1
Citizen Science	100	gender	0	250	gender	0
Citizen Science	100	religion	0	250	religion	0
Citizen Science	100	status	0	250	status	2
Citizen Science	100	other	0	250	other	1
Citizen Science	100	toxic	21	250	toxic	32
NGO	100	age	1	250	age	1
NGO	100	nationality	28	250	nationality	32
NGO	100	politics	7	250	politics	4
NGO	100	gender	2	250	gender	9
NGO	100	religion	0	250	religion	1
NGO	100	status	4	250	status	5
NGO	100	other	2	250	other	2
NGO	100	toxic	48	250	toxic	51
Prolific	100	age	1	250	age	0
Prolific	100	nationality	29	250	nationality	27
Prolific	100	politics	4	250	politics	2
Prolific	100	gender	2	250	gender	5
Prolific	100	religion	0	250	religion	0
Prolific	100	status	2	250	status	1
Prolific	100	other	1	250	other	1
Prolific	100	toxic	31	250	toxic	29
Research Assistants	100	age	1	250	age	1
Research Assistants	100	nationality	22	250	nationality	14
Research Assistants	100	politics	1	250	politics	1
Research Assistants	100	gender	3	250	gender	2
Research Assistants	100	religion	0	250	religion	1
Research Assistants	100	status	2	250	status	0
Research Assistants	100	other	1	250	other	0
Research Assistants	100	toxic	28	250	toxic	20

Note: No false negative classifications for target groups “sexuality”, “appearance” or “disability”.

5 Inter-Coder Reliability Measures

Our annotation protocol was implemented in two stages for human coders. In the first stage, coders determined whether a comment was harmful or non-harmful. In a second stage, for comments identified as harmful, coders further distinguished between hate speech and toxic speech and additionally annotated the target groups when it was hate speech. Although target group information enriches the context of the annotations, in this paper, we focus on the measures for the harmful vs. non-harmful decision and for differentiating hate speech from toxic speech. In other words, the target group annotations are not used in the current experiments but are part of the full coding scheme used by human coders. To reflect this process, we report intercoder reliability measures for the harmful vs. non-harmful speech, the toxic vs. non-toxic speech separately as well as the differentiation between all hate speech, toxic speech and neither of those two which is non-harmful speech.

For the harmful vs. non-harmful decision, the reliability measures are derived by aggregating the hate and toxic labels into a single “harmful” category. This mirrors the first stage of the human coding process. Table 5 displays the performance for the harmful versus non-harmful (or “hate vs. no hate”) classification; Table 6 shows the performance for the toxic versus non-toxic distinction; and Table 7 presents the reliability for the three-class scenario. Across all tasks, expert annotators lead in overall reliability, although NGO coders sometimes achieve comparable raw agreement with lower Krippendorff’s Alpha values. Note that we here report the values for the (independent) expert annotations prior to deliberation where they then reached full consensus on the final gold standard annotations.

In the harmful vs. non harmful classification (Table 5), the experts show the highest reliability, reaching an overall agreement of 0.822 and a Krippendorff’s Alpha of 0.761. Research assistants have an agreement of 0.632 and Alpha of 0.506, with stronger performance in identifying non-hate speech (Alpha 0.246) than hate speech (Alpha 0.190). NGOs perform similarly, with an agreement of 0.660 and Alpha of 0.458, again trailing

the expert group. Citizen Science participants and Prolific annotators record moderate agreement levels of 0.508 and 0.549, respectively, and corresponding Alpha values of 0.335 and 0.391. Both show slightly better reliability for non-hate speech than hate speech. Appen stands out for its relatively weak consistency, yielding an agreement of 0.310 and Alpha of 0.080.

Table 5: Performance Comparison Across Annotator Groups. The table compares annotation quality among different Groups relative to the expert group. A represents overall agreement, while α represents Krippendorff’s Alpha as a measure of reliability. A (Hate Speech) and α (Hate Speech) correspond to agreement and reliability for hate speech annotations, respectively. Similarly, A (No Hate Speech) and α (No Hate Speech) represent the same measures for non-hate speech annotations.

Platform	A	α	A HS	A No HS	α HS	α No HS
Appen	0.310	0.080	0.294	0.324	0.017	0.053
Citizen Science	0.508	0.335	0.437	0.574	0.157	0.191
Experts	0.822	0.761	0.790	0.852	0.219	0.104
NGO	0.660	0.458	0.500	0.809	0.334	0.332
Prolific	0.549	0.391	0.492	0.602	0.194	0.199
Research Assistants	0.632	0.506	0.571	0.688	0.190	0.246

In the toxic vs. non-harmful classification (Table 6), experts again post strong results with an agreement of 0.816 and an Alpha of 0.645. NGOs exhibit the highest raw agreement at 0.852 but maintain a lower Alpha of 0.435, suggesting uncertainty on more ambiguous cases. Research assistants achieve an agreement of 0.763 and Alpha of 0.330, while Citizen Science participants reach 0.555 agreement and Alpha of 0.334, exceeding Appen’s weaker reliability of 0.352 agreement and Alpha of 0.071. Prolific is not reported in this table because of insufficient toxic cases.

For the three-class scenario encompassing hate speech, toxic speech, and non-harmful (Table 7), experts again show the highest reliability at 0.761 agreement and 0.728 Alpha. NGOs drop to 0.617 agreement and 0.433 Alpha, Citizen Science participants and Research Assistants remain in the low to moderate range, and Prolific stands at 0.549 agreement with an Alpha of 0.391. Appen shows the greatest inconsistency with a 0.269 agreement and an Alpha of 0.129. Although NGOs and some other groups occasionally rival or surpass experts in raw agreement, no set of non-expert annotators consistently

Table 6: Performance Comparison Across Annotator Groups. The table compares annotation quality among different Groups relative to the expert group. A represents overall agreement, while α represents Krippendorff’s Alpha as a measure of reliability. A (Toxic Speech) and α (Toxic Speech) correspond to agreement and reliability for toxic speech annotations, respectively. Similarly, A (Not Toxic Speech) and α (Not Toxic Speech) represent the same measures for non-toxic speech annotations. We do not report numbers for Prolific since there are not enough cases of toxic for reliable values.

Platform	A	α	A Toxic	A Not Toxic	α Toxic	α Not Toxic
Appen	0.352	0.071	0.266	0.372	0.016	0.047
Citizen Science	0.555	0.334	0.500	0.568	0.211	0.360
Experts	0.816	0.645	0.723	0.838	0.261	0.212
NGO	0.852	0.435	0.649	0.900	0.426	0.33
Prolific	-	-	-	-	-	-
Research Assistants	0.763	0.330	0.532	0.818	0.363	0.056

matches the expert Alpha values across all categories.

Table 7: Performance Comparison Across Annotator Groups. The table compares annotation quality among different Groups relative to the expert group. A represents overall agreement, while α represents Krippendorff’s Alpha as a measure of reliability. A (Hate Speech) and α (Hate Speech) correspond to agreement and reliability for Hate Speech annotations, respectively. Similarly, A (Toxic Speech) and α (Toxic Speech) represent the same measures for non-toxic speech annotations and the same goes for all that are neither toxic nor hate speech.

Platform	A	α	A HS	A Toxic	A Neither	α HS	α Toxic	α Neither
Appen	0.269	0.129	0.188	0.255	0.320	0.111	0.082	0.084
Citizen Science	0.455	0.308	0.347	0.309	0.570	0.078	0.217	0.189
Experts	0.761	0.728	0.639	0.702	0.852	0.302	0.212	0.090
NGO	0.617	0.433	0.410	0.436	0.436	0.307	0.350	0.350
Prolific	0.549	0.391	0.514	0.457	0.602	0.181	0.201	0.200
Research Assistants	0.538	0.435	0.444	0.340	0.664	0.122	0.272	0.172

We would like to emphasize that even though the overall inter-coder reliability of our expert annotators were high, they nonetheless – prior to reaching consensus through deliberation – did not always agree on whether speech was a targeted attack and who was targeted. Among the 69 cases initially flagged as hate speech by at least one expert, 47 were deemed hate speech after deliberation, 8 non-targeted toxic and the other 14 non-harmful content. For the 55 cases initially flagged as harmful by at least one expert, after deliberation 20 were deemed non-targeted toxic speech, 5 as hate speech and the other 30 as non-harmful content.

6 Effects of Hate Speech Definitions on GPT’s Classification Performance

We conducted an additional analysis to assess the potential impact of different definitions of hate speech on the classification performance of GPT-4o-mini. Specifically, we compared the model’s performance under three distinct conditions: a narrowly framed definition of hate speech modeled on common UN usage, a broader definition that also includes offensive language aimed at individuals or groups based on social or political status, and a human annotator prompt that is grounded in the stricter UN criteria. All of these classifications were then evaluated against annotations by human experts consistently following the narrower UN standard.

Table 8 provides the results for a binary classification task (hate vs. non-hate). The model exhibits varied performance across prompts. When using the human annotator prompt, GPT-4o-mini achieves the highest recall at 0.950 but a relatively low precision of 0.681, producing an accuracy of 0.761 and an F1-Score of 0.793. The UN Hate Speech Definition Prompt takes a more balanced approach, reaching an accuracy of 0.757, a recall of 0.756, a precision of 0.744, and an F1-Score of 0.750. This could suggest that the LLM has seen (or been trained on) a definition of hate speech similar to the UN definition, e.g., when implementing guardrails. The Fine Tuning Hate Speech Prompt yields the highest precision at 0.832, although it manifests a much lower recall of 0.395, leading to an accuracy of 0.670 and an F1-Score of 0.536.

Table 8: Performance of GPT-4o-mini annotations under different hate speech definition prompts compared to human expert annotations (Binary Classification).

Hate Speech Def.	Accuracy	Recall	Precision	F1-Score
Human Annotator Prompt	0.761	0.950	0.681	0.793
Fine Tuning Hate Speech Prompt	0.670	0.395	0.832	0.536
UN Hate Speech Definition Prompt	0.757	0.756	0.744	0.750

Table 9 summarizes the results for the more challenging three-way classification (hate, offensive, and non-hate). Although all metrics decrease compared to the more straightforward

Table 9: Performance of GPT-4o-mini annotations under different hate speech definition prompts compared to human expert annotations (Three-Way Classification).

Hate Speech Def.	Overall Accuracy	Recall	Precision	F1-Score
Human Annotator Prompt	0.668	0.680	0.673	0.650
Fine Tuning Hate Speech Prompt	0.553	0.615	0.656	0.560
UN Hate Speech Definition Prompt	0.668	0.574	0.719	0.518

ward binary scenario, a similar performance pattern emerges. Both the human annotator prompt and the UN Hate Speech Definition Prompt attain an accuracy of 0.668, but they diverge in their respective trade-offs: the former yields a recall of 0.680 and a precision of 0.673, while the latter drops to 0.574 recall yet improves on precision at 0.719. The Fine Tuning Hate Speech Prompt shows the weakest outcomes overall, achieving an accuracy of 0.553 and an F1-Score of 0.560, driven by comparatively imbalanced recall and precision values.

Although the lower recall in the classification differentiating between Hate Speech and Toxic Speech may seem concerning at first glance, it primarily reflects the increased complexity of distinguishing among three classes of which two are very nuanced rather than a fundamental weakness of the model. This decrease in recall is partly due to the heightened difficulty of accurately separating hate speech from toxic speech when definitions are finely parsed. Notably, the consistency of precision and F1 score across different prompts suggests that the model’s overall robustness in aligning with expert annotations remains intact. Although different prompts slightly shift the balance between recall and precision, the model classification performance does not change dramatically. This consistency highlights the robustness of GPT-4o-mini in handling different operational definitions of hate speech, indicating that changes in conceptual formulation have only a modest influence on its ability to match expert-coded gold standards, although also modest changes may matter in practice for small but meaningful performance improvements (or lack thereof) of such classifiers.