Revisiting Implicitly Abusive Language Detection: Evaluating LLMs in Zero-Shot and Few-Shot Settings

Julia Jaremko
Digital Age Research Center (D!ARC)
Alpen-Adria-Universität Klagenfurt, Austria
julia.pardatscher@aau.at

Dagmar Gromann
Center for Translation Studies
University of Vienna, Austria
dagmar.gromann@univie.ac.at

Michael Wiegand
Digital Philology
Faculty of Philological and
Cultural Studies
University of Vienna, Austria
michael.wiegand@univie.ac.at

Abstract

Implicitly abusive language (IAL), unlike its explicit counterpart, lacks overt slurs or unambiguously offensive keywords, such as *bimbo* or *scum*, making it challenging to detect and mitigate. While current research predominantly focuses on explicitly abusive language, the subtler and more covert forms of IAL remain insufficiently studied. The rapid advancement and widespread adoption of large language models (LLMs) have opened new possibilities for various NLP tasks, but their application to IAL detection has been limited. We revisit three very recent challenging datasets of IAL and investigate the potential of LLMs to enhance the detection of IAL in English through zero-shot and few-shot prompting approaches. We evaluate the models' capabilities in classifying sentences directly as either IAL or benign, and in extracting linguistic features associated with IAL. Our results indicate that classifiers trained on features extracted by advanced LLMs outperform the best previously reported results, achieving near-human performance.

1 Introduction

Offensive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person.[1] Closely related terms, e.g. *abusive language* (Nobata et al., 2016), *cyber bullying* (Zhong et al., 2016) or *hate speech* (Waseem and Hovy, 2016), are compatible with the definition above. In the remainder, we use the term *abusive language* simply because most datasets we work with in this paper use that expression.

Implicitly abusive language (IAL) represents a subtle form of abusive language. Unlike explicitly abusive language (1)-(2), which includes overtly offensive expressions, IAL involves **subtler expressions that convey hostility without using explicit discriminatory language** (3)-(4).

(1) Negros are so dumb.
(2) my fag jew teacher told my principal that i have white power markings on my binder and backpack i was suspended for 1 week ! [sic]
(3) I think Whites should home school their children if at all possible.
(4) Have Muslims ever made a contribution to our society?

While explicitly abusive language has been the primary focus of recent studies, the subtler forms of IAL still pose a considerable challenge (van Aken et al., 2018). Automated detection of IAL is more complex than identifying explicit instances because it requires classifiers to go beyond simple keyword detection and understand the nuanced meaning behind linguistic patterns.

A common approach for detecting abusive language, whether explicit or implicit, is supervised learning (Wiegand et al., 2021b). This method relies heavily on manually annotated data to train models effectively (Plaza-del arco et al., 2023). However, creating sufficient annotated training data can be costly, especially for smaller organizations, as noted by Ding et al. (2022). This challenge is especially pronounced for IAL, where the subtle language requires domain-specific knowledge to interpret it, making annotation more complex and resource-intensive (Waseem et al., 2017).

Given these challenges, **large language models (LLMs)** offer a promising alternative to human annotators. They can generate text that closely

[1] http://thelawdictionary.org

Proceedings of the 31st International Conference on Computational Linguistics, pages 3879–3898
January 19–24, 2025. ©2025 Association for Computational Linguistics

mimics human language and perform various NLP tasks, such as translating languages, creating summaries and answering questions (Ding et al., 2022; Chen et al., 2023; Baktash and Dawodi, 2023).

In this paper, we use zero-shot and few-shot approaches to assess whether LLMs can directly recognize IAL or identify linguistic features correlated with IAL. The direct recognition of IAL is the most straightforward and a fairly simple method for automatically recognizing a specific linguistic phenomenon. It also allows us to investigate the extent to which LLMs inherently grasp IAL. By also considering the identification of linguistic features, we can explore how a set of specific linguistic categories, which are typically easier to detect than the complex phenomenon of IAL, can be utilized to identify IAL. Both direct recognition and the identification of linguistic features are **evaluated on three recent, challenging datasets** for IAL: one targeting abusive language towards identity groups, another focusing on euphemistically abusive language and a third on abusive comparisons.

Our **contributions** are as follows:

- We revisit three challenging recent datasets for IAL using zero-shot and few-shot approaches with LLMs.
- We show that advanced LLMs, such as GPT-4, can be a viable alternative to human annotators for extracting difficult linguistic features.
- We establish new benchmark performance with our approaches using LLMs, achieving near-human performance on all three datasets.

2 Related Work

The escalating prevalence of abusive language online and its detrimental effects underscore the urgency of abusive language detection in NLP (Gelber and McNamara, 2016; Mullen and Smyth, 2004; Müller and Schwarz, 2017). This challenge requires a nuanced understanding of both explicit and implicit forms. Fortuna and Nunes (2018) and Schmidt and Wiegand (2017) contribute to this understanding by providing a comprehensive overview of the current state of the automatic detection of abusive language detection, organizing prior research, methods and datasets, and highlighting the complexities and societal impacts associated with abusive language.

In the area of detecting abusive language detection, Hartvigsen et al. (2022) introduce ToxiGen, a large-scale dataset generated using GPT-3 (Brown et al., 2020a), demonstrating that the usage of LLMs can improve classifiers' ability to identify subtle, implicitly toxic language. Plaza-del arco et al. (2023) investigate the application of zero-shot learning with various LLMs for abusive language detection across three languages: English, Italian and Spanish. Their research indicates that zero-shot prompting with LLMs can match or exceed the performance of fine-tuned models, making it a viable option for under-resourced languages. Min et al. (2022) examine how LLMs perform in in-context learning by analyzing various tasks, including abusive language detection. Huang et al. (2023) investigate the potential of using ChatGPT, presumed to be GPT-3.5, to detect IAL and generate natural language explanations, demonstrating that GPT-3.5 provides clear and often convincing explanations.

Recent studies also explore the use of LLMs for data annotation across various NLP tasks. Ding et al. (2022) assess the effectiveness of GPT-3 in annotating datasets for a range of applications. They demonstrate that GPT-3 reduces the need for labor-intensive manual annotations, although improvements are needed in the quality of the data produced. Similarly, Gilardi et al. (2023) demonstrate that GPT-3.5 outperforms crowd-workers achieving higher accuracy and inter-annotator agreement at a fraction of the cost. Chiang and Lee (2023) explore the potential of using LLMs as an alternative to human evaluators for assessing the quality of text in various NLP tasks, finding that LLMs produce consistent and reproducible evaluations comparable to expert human assessments.

Building on prior research into LLMs for abusive language detection and human annotation substitution, this study focuses on IAL detection using zero-shot and few-shot prompting. Our key contribution lies in comparing LLM-driven IAL annotation with human annotations, addressing complex linguistic phenomena such as euphemisms, contradictory comparisons, absurd images, etc., which challenge even human annotators. Despite the complexity of the task and the novelty of the datasets, our findings demonstrate that LLMs can effectively handle these nuanced tasks.

3 Data and Tasks

In our experiment, we make use of three datasets, each tailored to a specific subtype of IAL. These datasets are structured for **binary text classifica-**

tion tasks where the objective is to differentiate between the subtype of IAL and non-abusive language **at the sentence level**. All datasets treat the text instances out of context. Moreover, for each of these datasets, a set of complex linguistic features have been explored. These datasets also have in common that they were crafted to ensure **minimal biases**. This was mainly achieved by producing difficult non-abusive instances that were aimed to be structurally and semantically similar to the abusive instances of the respective dataset, e.g. by employing *contrast sets* (Gardner et al., 2020). Thus, the datasets should not exhibit notable *spurious correlations* (Ramponi and Tonelli, 2022) that classifiers for IAL often overfit to (Wiegand et al., 2021b).

The following descriptions provide an overview of the three datasets used in this study, each representing distinct aspects of IAL.

Identity Groups. Wiegand et al. (2022) introduced a dataset on IAL focusing on negative stereotypes, targeted against specific identity groups, namely gay people, Jews, Muslims and women, using data extracted from Twitter (now X). The dataset consists of **2,221 sentences**, out of which 1,249 were manually labeled as abusive.

For their modelling approach, the authors identified exactly three distinct linguistic features to coincide with this type of abuse. These features were then used to build various classifiers. The first feature addresses the non-episodic nature of abusive language, which refers to habitual or generalizing statements (5), rather than references to specific incidents (6). The second feature assesses whether the target is depicted as a perpetrator, suggesting guilt or wrongdoing (7). The third feature examines whether the target is portrayed as a nonconformist, implying deviation from societal norms (8).

(5) Muslims are peddling fake news *as always*.
(6) Muslims pillage embassy of Saudi Muslims.
(7) Muslims *terrorize* the world daily.
(8) Muslims dislike peace.

Euphemistic Abuse. Wiegand et al. (2023) created a dataset to detect subtle, euphemistically abusive language that paraphrases explicit abuse (9)-(11). This dataset, generated through crowdsourcing, focuses on high lexical variability to challenge classifiers in identifying nuanced forms of abuse. It consists of **1,797 sentences**, out of which 640 were manually labeled as abusive.

(9) You would fit well in a brothel. (→ You are whorey).
(10) You are truly the best at doing nothing. (→ You are lazy.)
(11) You enjoy sharing your genitals with everyone you lay your eyes on. (→ You are licentious.)

Wiegand et al. (2023) identified six linguistic features of euphemistic abuse and trained a supervised classifier using these manually extracted features. One example of these features are extreme or absolute language (10) that may be conveyed by the usage of superlatives, generalizations or hyperbole. Another feature is the use of words linked to taboo subjects, such as physical or mental abnormalities, particular body parts, death, etc. (11).

Comparisons. The dataset by Wiegand et al. (2021a) focuses on detecting implicit abuse through comparative statements (12)-(14). It was also created through crowdsourcing and consists of **1,000 sentences**, out of which 500 were manually labeled as abusive.

(12) You make me feel like bringing up my lunch.
(13) You behave like a toddler on acid.
(14) You are as smart as a Neanderthal.

For classification, a combination of linguistic features similar to the approach of Wiegand et al. (2023) was proposed for supervised classification. Among the six manually compiled features[2] are, for example, absurd images (13) where comparisons invoke bizarre or highly unlikely scenarios to add a layer of ridicule or surreal critique. Another feature concerns contradictions (14). These generally occur when the characteristic of the comparison, e.g. *smart*, is contrary to the prototypical characteristics associated with the vehicle, e.g. a *Neanderthal*, is typically considered to be simpleminded. Such sarcastic comparisons are often perceived as abusive.

Tables 22 and 23 of Appendix A.6 provide a complete list of the manually extracted features proposed for euphemistic abuse and abusive comparisons that we consider in this work.

4 Method

We carefully selected LLMs to directly detect IAL (§4.2) and identify linguistic features predictive of IAL (§4.3). We also include a rule-based classifier (§4.5) and supervised classifiers (§4.6) that utilize the features extracted by the LLMs in their decision-making process.

[2] That work also examined automatically generated features that are notably simpler to produce relying, for example, on part-of-speech or frequency information.

4.1 Selection of LLMs and Settings

As LLMs for our experiments, we selected GPT-3.5 (Brown et al., 2020a), GPT-4 (OpenAI et al., 2023) and LLaMA-3 (Dubey et al., 2024). We chose GPT-3.5 and GPT-4 due to their advanced language understanding and generation capabilities, along with their strong performance across various NLP tasks (Chen et al., 2023; Baktash and Dawodi, 2023). To complement these proprietary models, we also included LLaMA-3 as a very recent open-source alternative. We chose the default settings[3] that were proposed by the platforms through which we use these models, i.e. OpenAI and Replicate (OpenAI Playground; Replicate, 2024).

4.2 LLM-Driven Direct Detection of IAL

One objective of this work is to evaluate the ability of LLMs to directly identify IAL. To achieve this, we employ prompts that directly address the task rather than correlated features. For this purpose, we have developed both zero-shot and few-shot prompting approaches.

Zero-shot prompting provides the LLM with a natural language description of the task without any examples or prior demonstrations of the task. The expectation is for the LLM to understand and execute the task based solely on its pre-training and the provided description. This approach is advantageous for its convenience and potential robustness, as it avoids the LLM learning spurious correlations from specific examples. However, it can also be challenging; without examples, the LLM might not clearly grasp the task requirements, particularly if the task format is ambiguous (Brown et al., 2020b).

Few-shot prompting equips the LLM with a few task-specific examples in addition to the task description. These examples act as a reference for the LLM on how the task should be performed.

Figure 1 illustrates the zero-shot prompt structure we devised for the task of detecting IAL with LLMs. The selected **identifiers**, *hateful*, *abusive*, *offensive*, *toxic* and *insulting*, reflect the wide spectrum of language constituting abusive language. Differences in prompt formulation can substantially influence the performance of LLMs (Zhang et al., 2021; Min et al., 2022; Plaza-del arco et al., 2023). By incorporating a range of terms, the aim is to capture the varied expressions of abusive language, thereby ensuring a more comprehensive assessment of the LLMs' capabilities in detecting IAL.

For the few-shot[4] solution, a balanced setup with 10 labeled examples was implemented, evenly split between the two classes, *abusive* and *non-abusive* language. The decision to use 10 examples was based on the findings of Min et al. (2022), which indicate that this is the point at which the performance of LLMs begins to plateau. To avoid artificially simplifying the task by providing LLMs with direct solutions, the examples chosen were not taken from the dataset itself. Instead, we treated the LLMs like human annotators, selecting examples from annotation guidelines and papers that present the datasets (Wiegand et al., 2021a; 2022; 2023). This ensured that the examples were both effective and prototypical, without revealing the correct answers of the test data directly.

4.3 LLM-Driven Feature Extraction

IAL, despite its subtle nature, possesses discernible characteristics that make it possible to detect. As outlined in §3, previous research accounted for that by establishing various linguistic features correlated with the particular subtypes of IAL. We tried to systematically extract these features with LLMs.

In our zero-shot prompting approach, we give the LLM a natural language description of the task[5] without including any examples. We included a definition of a feature if the concept to be annotated was not straightforward. For these cases, we used the definition from the paper that introduced the respective feature. For instance, for the prompt template of the feature *non-conformist views* (15), we included a definition, whereas for the feature *contradiction*, given the simplicity of the underlying concept, we did not include one (16).

(15) Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence "{sentence}" Does the author of this sentence think that {target[6]} are non-conformist?

(16) Is there a contradiction in the following sentence? "{sentence}"

In our few-shot prompting approach, we developed the prompt templates for LLM-driven feature extraction by adding 10 examples to the original

[3] The specific settings are provided in Appendix A.1.

[4] The complete prompt templates are provided in Appendices A.4 and A.5.

[5] Tables 19-21 of Appendix A.5 present the complete zero-shot prompt templates designed to extract the linguistic features of IAL.

[6] The placeholder {target} represents one of four identity groups: *gay people*, *women*, *Jews* or *Muslims*.

```
available_identifiers ← ["hateful", "abusive", "offensive", "toxic", "insulting"]
identifier ← selectIdentifierToUseForClassification(available_identifiers)
sentences ← getAllInstances(dataset)
for sentence in sentences do
    prompt ← 'Is this sentence {identifier}? "{sentence}" Answer either "YES" or "NO". No other answer permitted.'
end for
```

Figure 1: Zero-shot prompt structure for the task of direct detection of IAL.

```
procedure isImplicitlyAbusive(sentence)
    abusive ← FALSE
    if not (getAspect(sentence) == EPISODIC) then
        if hasPerpetrator(sentence) then
            abusive ← TRUE
        else if hasNonConformistView(sentence) then
            abusive ← TRUE
        end if
    end if
    return abusive
end procedure
```

Figure 2: Rule-based classifier for the dataset *Identity Groups*.

zero-shot prompt template. Table 1 demonstrates this method using the feature *non-conformist*. The examples were taken either from the annotation guidelines or from the paper presenting the respective dataset (Wiegand et al., 2021a; 2022; 2023).

4.4 Prompt Optimization and Response Preprocessing

To ensure that the LLM responses aligned with our requirements, we optimized the prompts through a process of trial and error. Strategies included explicit instructions such as: Answer either YES or NO. No other answer is permitted. Despite these efforts, the LLMs did not always follow the instructions perfectly, and minor deviations occasionally occurred.

To address these inconsistencies, we implemented preprocessing steps to standardize responses without altering the original outputs of the LLMs. These included converting text to uppercase, removing punctuation, and extracting YES or NO responses using regular expressions. In rare instances where the LLMs failed to provide a clear response, we replaced non-responses with the most frequent label generated by the respective LLM within each dataset. Each LLM-prompt-dataset combination was executed only once.

4.5 Feature Combination through Rule-Based Classifier (on the Dataset *Identity Groups*)

Wiegand et al. (2022) developed a rule-based classifier for the dataset *Identity Groups* by assessing sentences according to the linguistic features described in §3. The classifier (see Figure 2), operates by determining whether a sentence is non-episodic, which indicates that it describes habitual or routine behaviors rather than isolated events. It then checks if the sentence depicts the targeted group as either perpetrators involved in reprehensible actions or as non-conformists who deviate from societal norms. A sentence is classified as IAL if it satisfies the non-episodic condition and at least one of the latter conditions, such as (17) and (18); otherwise, it is labeled as non-abusive, such as (19) and (20).

(17) Lesbians pressure straight women into being lesbians all the time.
{*non-episodic*, *perpetrator*, *conformistic*} → *abuse*

(18) Muslims lack basic morality bcoz quran [sic] lacks morality lessons.
{*non-episodic*, *no-perpetrator*, *non-conformistic*} → *abuse*

(19) Jews censure Miley Cyrus for not raising Gypsy apartheid issue on 'Gypsy Heart' tour.
{*episodic*, *no-perpetrator*, *non-conformistic*} → *other*

(20) All women grapple with the voice of self- doubt [sic] in one way or another.
{*non-episodic*, *no-perpetrator*, *conformistic*} → *other*

Wiegand et al. (2022) instantiate this classifier in two different ways, either by manually extracting them, which should be considered an upper bound of human annotation, or automatically using a cascade of three automatic classifiers that implement each feature conventionally, i.e. using traditional NLP resources like word lists, rather than LLMs. Wiegand et al. (2022) refer to the latter as the *linguistically informed classifier*. In this paper, we introduce a third variant of the rule-based classifier that implements these three features using LLMs. This allows us to examine in how far the effectiveness of the rule-based classifier depends on the quality of the implementation of its three linguistic features.

Unfortunately, there is no similar rule-based clas-

Prompt Template
Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence “sentence” Does the author of this sentence think that the agents (in this case {target}) are non-conformists? Answer either “YES” or “NO”. No other answer is permitted. Here are examples: Sentence: “Lesbians pray to Satan.” Answer: “YES” Sentence: “Jews long for a safe Israel.” Answer: “NO” Sentence: “Muslims dislike peace.” Answer: “YES” Sentence: “Women abhor violence.” Answer: “NO” Sentence: “Lesbians disrespect God’s plan.” Answer: “YES” Sentence: “Muslims grieve for their brothers.” Answer: “NO” Sentence: “Muslims glorify violence.” Answer: “YES” Sentence: “Jews suffer from ethnic cleansing.” Answer: “NO” Sentence: “Muslims revert to stoning victims.” Answer: “YES” Sentence: “Nowadays, women are promoting self-assurance and valuing their own importance.” Answer: “NO”

Table 1: Few-shot prompt template for extracting the linguistic feature *non-conformist* using LLMs. The placeholder {target} represents one of four identity groups: *gay people*, *women*, *Jews* or *Muslims*.

sifier for the other two datasets, i.e. Euphemistic Abuse and Comparisons. Therefore, we examine this rule-based approach **only for the dataset *Identity Groups***.

4.6 Completion Combination through Logistic Regression

To investigate whether the different completions of the LLMs can be usefully combined for detecting IAL, three logistic regression models were built following different training scenarios.

The first model uses completions from our experiments on the direct detection of IAL (§4.2). It uses the completions by the LLMs to the five identifiers *hateful*, *abusive*, *offensive*, *toxic* and *insulting*. These responses, more specifically the individual class predictions, are used as features for training. With this setup, we aim to assess whether combining all identifiers provides some benefit compared to focusing on predictions from just one identifier.

The second model is trained using completions derived from the LLM-driven feature extraction (§4.3). While we have already introduced a method for combining features for the dataset *Identity Groups* through rule-based classification (§4.5), no equivalent method exists for the other two datasets. By combining features through logistic regression, we present a universally applicable technique.

The third model is a blend of the previous two. It combines the responses to the five prompts with the identifiers and completions from the LLM-driven linguistic feature extraction.

We employed logistic regression as implemented in *scikit-learn* (Pedregosa et al., 2011) and evaluated the resulting classifiers using **five-fold cross-validation** on each of the three datasets separately.

5 Results

5.1 Direct Detection of IAL

Table 2 presents the results of the direct detection of IAL using LLMs in both zero-shot and few-shot settings. It displays the average F1 scores and standard deviations for five different identifiers, with these metrics calculated from individual macro-averaged F1 scores across the selected LLMs, i.e. GPT-3.5, GPT-4 and LLaMA-3, as well as the three datasets, i.e. *Identity Groups*, *Euphemistic Abuse* and *Comparisons*.[7] Table 2 indicates that the performance of all five identifiers is relatively similar, with no identifier notably outperforming the others. The *offensive* identifier achieved on average the highest F1 score, along with the second-lowest standard deviation. As a result, we have selected it as the reference identifier for subsequent analysis. On average, the few-shot approach demonstrates better performance than the zero-shot approach.

Identifier	Zero-shot	Few-shot	Average
hateful	69.4 (±8.6)	72.3 (±9.7)	70.9 (±9.1)
insulting	70.7 (±7.8)	71.6 (±9.3)	71.2 (±8.5)
offensive	**70.9** (±8.2)	72.7 (±9.0)	**71.8** (±8.6)
toxic	70.8 (±8.6)	72.4 (±8.7)	71.6 (±8.6)
abusive	69.2 (±9.3)	**72.8** (±9.9)	71.0 (±9.6)

Table 2: Direct detection of IAL: Average F1 scores and standard deviations calculated from individual macro-averaged F1 scores across multiple LLMs (GPT-3.5, GPT-4 and LLaMA-3) and datasets (*Identity Groups*, *Comparisons* and *Euphemistic Abuse*).

Table 3 reports on the results of the direct detection of IAL on each dataset separately. It shows the macro-averaged F1 score only for the specific

[7] Detailed results for each identifier are provided in Tables 12-14 of Appendix A.2.

identifier *offensive*. The LLMs demonstrate varying levels of performance, with GPT-4 performing the best, followed by LLaMA-3, while GPT-3.5 lags notably behind. The Identity Groups dataset appears to be easier for the models, with all models performing better on this task. On average, the few-shot approach shows better results than zero-shot.

Dataset	Model	ZS	FS
Identity Groups	GPT-3.5	64.91	71.95
Identity Groups	GPT-4	81.26	**82.37**
Identity Groups	LLaMA-3	79.38	82.31
Comparisons	GPT-3.5	58.90	57.67
Comparisons	GPT-4	**75.93**	74.93
Comparisons	LLaMA-3	69.43	70.34
Euphemistic Abuse	GPT-3.5	59.76	58.73
Euphemistic Abuse	GPT-4	76.16	**78.01**
Euphemistic Abuse	LLaMA-3	71.62	75.63

Table 3: Direct detection of IAL: Macro-averaged F1 scores of the identifier *offensive* in the zero-shot (ZS) and few-shot (FS) approaches.

5.2 Feature Combination through Rule-Based Classifier (on the Dataset *Identity Groups*)

Table 4 presents the macro-averaged F1 scores for the rule-based classifier (§4.5) on the dataset *Identity Groups* using different ways of how the features are extracted. The results show that all but three classifiers achieved higher scores than the classifier that relied on manually extracted features by human annotators. These three exceptions used features extracted by GPT-3.5 and features extracted with the help of **traditional NLP resources**, such as word lists, as proposed by Wiegand et al. (2022). The latter result shows impressively that traditional resources are notably inferior to LLMs. All rule-based classifiers using features extracted by either GPT-4 or LLaMA-3 outperform the classifier using features extracted by human annotators by a large degree. These findings align with the work of Gilardi et al. (2023), who demonstrated that LLMs can outperform crowdworkers in several annotation tasks.

5.3 Feature Combination through Logistic Regression

Table 5 describes the exact configurations corresponding to the different identifiers used in subsequent tables. Table 6 describes classification methods underlying the best previously reported results taken from Wiegand et al. (2021a; 2022; 2023). Tables 7-9 report the performance of the logistic regression models that combine predictions of different kinds of LLM completions for the datasets *Identity Groups*, *Euphemistic Abuse* and *Comparisons*, respectively. Instead of reporting F-scores, we present the percentage change (based on macro-average F1) of one classifier relative to another.

Rule-based classifier using features ...	F1
extracted by GPT-3.5 zero-shot	59.6
extracted with traditional NLP resources*	71.9
extracted by GPT-3.5 few-shot	72.8
extracted by human annotators	**77.7**
extracted by GPT-4 zero-shot	82.8
extracted by LLaMA-3 zero-shot	83.3
extracted by LLaMA-3 few-shot	84.8
extracted by GPT-4 few-shot	86.2

*: this corresponds to the *linguistically informed classifier* as proposed by Wiegand et al. (2022)

Table 4: Dataset *Identity Groups*: Macro-averaged F1 scores for rule-based classifiers using (the same) features generated in different ways.

Classifier	Referring to results of ...
one identifier	the respective LLM at direct detection of IAL using the identifier *offensive*
combined ident.	logistic regression models trained on the LLM completions of all five identifiers: *hateful*, *insulting*, *offensive*, *toxic* and *abusive*
ling. features	logistic regression models trained on linguistic features for the respective dataset (§A.6) extracted by the respective LLM
ling. features + combined ident.	logistic regression models trained on linguistic features (extracted by the respective LLM) and LLM completions of all five identifiers
best previously reported result	best performing classifiers reported by respective previous work (Wiegand et al., 2021a, 2022, 2023), see also Table 6

Table 5: Description of classifiers used in Tables 7-9.

For the task of recognizing abuse directed at identity groups, referred to as the dataset *Identity Groups*, the results are presented in Table 7. It shows that, except for GPT-3.5 in the few-shot approach, all models that combine the completions of all five identifiers outperform the direct detection of abusive language results of the single identifier *offensive*. The capabilities of GPT-4 and LLaMA-3 in recognizing abusive language directly are close to those of models trained on linguistic features. The strongest classifier is the model trained on both linguistic features and combined identifiers. With the exception of GPT-3.5, all classifiers presented here outperform the best previously reported results (Wiegand et al., 2022).

Table 8 displays the same classifier configura-

Dataset	Model	Additional Details	F1
Identity Groups	Logis. Regr.	trained on (manually extracted) linguistic features	77.7
Euphem. Abuse	Logis. Regr.	trained on (manually extracted) linguistic features	75.6
Compar.	BERT	fine-tuned on dataset combined with (manually extracted) linguistic features	72.9

Table 6: Description of the best previously reported classifiers from Wiegand et al. (2021a; 2022; 2023).

tions as in Table 7, but applied to the dataset *Euphemistic Abuse*. Models using combined identifiers outperform the LLMs' direct detection capabilities with the single identifier *offensive*. The direct detection capabilities of the LLMs (*combined identifiers*) generally outperform the feature-based approaches (*linguistic features*). As in Table 7, the strongest classifiers are the models trained on both linguistic features and combined identifiers. Notably, the classifiers trained on features extracted by GPT-4 in both approaches and LLaMA-3 in the few-shot approach outperform the best previously reported results (Wiegand et al., 2023).

Comparison	LLM	ZS	FS
one ident. *vs.* comb. ident.	GPT-3.5	+3.1	-0.9
one ident. *vs.* comb. ident.	GPT-4	+3.7	+4.8
one ident. *vs.* comb. ident.	LLaMA-3	+1.0	+0.8
ling. feats. *vs.* comb. ident.	GPT-3.5	-7.2	+14.2
ling. feats. *vs.* comb. ident.	GPT-4	+3.8	+0.6
ling. feats. *vs.* comb. ident.	LLaMA-3	-0.5	-1.7
ling. feats. *vs.* ling. feats. + comb. ident.	GPT-3.5	+3.4	+15.1
ling. feats. *vs.* ling. feats. + comb. ident.	GPT-4	+5.1	+1.3
ling. feats. *vs.* ling. feats. + comb. ident.	LLaMA-3	+4.2	+1.1
best previously rep. result *vs.* ling. feats. + comb. ident.	GPT-3.5	-4.1	-7.4
best previously rep. result *vs.* ling. feats. + comb. ident.	GPT-4	+9.8	+11.8
best previously rep. result *vs.* ling. feats. + comb. ident.	LLaMA-3	+8.1	+9.8

Table 7: Dataset *Identity Groups*: Comparison of classifiers with percentual change in macro-averaged F1 scores in the zero-shot (ZS) and the few-shot (FS) approach. Table 5 offers a description of each classifier.

Comparison	LLM	ZS	FS
one ident. *vs.* comb. ident.	GPT-3.5	+3.2	+4.7
one ident. *vs.* comb. ident.	GPT-4	+0.4	+1.3
one ident. *vs.* comb. ident.	LLaMA-3	+0.5	+0.4
ling. feats. *vs.* comb. ident.	GPT-3.5	+2.2	+0.4
ling. feats. *vs.* comb. ident.	GPT-4	+9.7	+8.4
ling. feats. *vs.* comb. ident.	LLaMA-3	+1.8	+4.7
ling. feats. *vs.* ling. feats. + comb. ident.	GPT-3.5	+3.3	+1.6
ling. feats. *vs.* ling. feats. + comb. ident.	GPT-4	+11.9	+9.3
ling. feats. *vs.* ling. feats. + comb. ident.	LLaMA-3	+5.6	+5.7
best previously rep. result *vs.* ling. feats. + comb. ident.	GPT-3.5	-17.6	-17.7
best previously rep. result *vs.* ling. feats. + comb. ident.	GPT-4	+3.2	+5.3
best previously rep. result *vs.* ling. feats. + comb. ident.	LLaMA-3	-1.2	+1.4

Table 8: Dataset *Euphemistic Abuse*: Comparison of classifiers with percentual change in macro-averaged F1 scores in the zero-shot (ZS) and the few-shot (FS) approach. Table 5 offers a description of each classifier.

Table 9 displays the same classifier configurations as in Table 7, but applied to the dataset *Comparisons*. Models using combined identifiers continue to outperform the LLMs' direct detection capabilities with the single identifier *offensive*. The direct detection capabilities of the LLMs are generally stronger than the feature-based approaches. As in Tables 7 and 8, the strongest classifiers are the models trained on both linguistic features and combined identifiers. The classifiers trained on features extracted by GPT-4 in both approaches and LLaMA-3 in the few-shot approach outperform the best previous classifier.

Comparison	LLM	ZS	FS
one ident. *vs.* comb. ident.	GPT-3.5	+11.5	+8.2
one ident. *vs.* comb. ident.	GPT-4	+0.0	+2.0
one ident. *vs.* comb. ident.	LLaMA-3	+4.0	+0.5
ling. feats. *vs.* comb. ident.	GPT-3.5	-1.2	+6.7
ling. feats. *vs.* comb. ident.	GPT-4	+18.5	+18.3
ling. feats. *vs.* comb. ident.	LLaMA-3	+5.5	+0.2
ling. feats. *vs.* ling. feats. + comb. ident.	GPT-3.5	-0.5	+8.4
ling. feats. *vs.* ling. feats. + comb. ident.	GPT-4	+17.5	+17.0
ling. feats. *vs.* ling. feats. + comb. ident.	LLaMA-3	+4.4	+1.9
best previously rep. result *vs.* ling. feats. + comb. ident.	GPT-3.5	-9.3	-13.0
best previously rep. result *vs.* ling. feats. + comb. ident.	GPT-4	+3.3	+3.7
best previously rep. result *vs.* ling. feats. + comb. ident.	LLaMA-3	-0.7	+1.5

Table 9: Dataset *Comparisons*: Comparison of classifiers with percentual change in macro-averaged F1 scores in the zero-shot (ZS) and the few-shot (FS) approach. Table 5 offers a description of each classifier.

Table 10 compares the strongest classifiers from this study with previously reported results. We also included the human baselines from Wiegand et al.'s work (2021a; 2022; 2023), which used the

Classifier	**Identity Groups**	**Euphem. Abuse**	**Compar.**
best previously reported result	77.7	75.6	72.9
human baseline	81.8	78.3	**77.6**
best classifier in this work	**86.9**	**79.5**	75.6

Table 10: Comparison of macro-averaged F1 scores between the strongest classifiers from this study and those from previous work (described in Table 6).

judgment of a single annotator randomly sampled from the crowdsourced gold-standard annotation for abusive language detection. This individual judgment may differ from the gold standard label, which is the majority vote of five annotators. In those previous work, these human baselines were considered an upper bound. Except for the *Comparisons* dataset, all upper bound human baselines were surpassed. Although these results initially seemed counterintuitive to us, we were able to explain them by closely examining the specific implementation of the human baselines. Since they are based on the judgment of a single randomly selected annotator, it may not be entirely solid or reliable; the performance of this annotator could vary widely depending on their individual ability and understanding. (A more robust baseline would involve averaging the judgments of multiple randomly selected annotators.) Nonetheless, we believe the quality of the existing human baseline allows us to conclude that our best classifiers are much closer to human performance than the best previously reported results.

6 Conclusion

This study evaluated the performance of LLMs, specifically GPT-3.5, GPT-4 and LLaMA-3, in detecting IAL using zero-shot and few-shot learning approaches. It also examined whether these models could effectively extract linguistic features predictive of IAL, functioning as annotators.

The few-shot learning approach generally outperformed the zero-shot approach, demonstrating the advantages of providing some (limited) training examples. GPT-4 achieved the best overall performance, slightly outperforming LLaMA-3, while GPT-3.5 lagged notably behind both.

Classifiers that combine linguistic features derived from LLM outputs generally outperformed those extracted by human annotators. This indicates that LLMs can provide annotations that are superior to manual efforts. The most effective classifiers were those that combined linguistic features with LLM outputs for the direct detection of abusive language. These classifiers outperformed the best previously reported results on most datasets and reach near-human performance.

7 Limitations

The dataset *Identity Groups* may not fully capture the nuances of real-world IAL, as it contains only atomic negative sentences that oversimplify the complexity of IAL. Moreover, the datasets *Euphemistic Abuse* and *Comparisons*, created via crowdsourcing, might lack authenticity, potentially missing the natural usage of abusive language. Consequently, we sought an alternative to the three previously described datasets. We identified the *ISHate* dataset (Ocampo et al., 2023), which offers a broader range of instances that better reflect real-world scenarios, making it a more representative resource. Ocampo et al. (2023) compiled 7 existing standard datasets designed to detect both explicit and IAL (de Gibert et al., 2018; Basile et al., 2019; ElSherief et al., 2021; Hartvigsen et al., 2022; Hammer, 2017; Chung et al., 2019; Fanton et al., 2021). We attempted to replicate the task from Ocampo et al. (2023) using LLMs to classify instances as non-abusive speech, IAL or explicitly abusive language. However, our experiments with *ISHate* did not yield meaningful conclusions, as the results were notably lower than those presented by Ocampo et al. (2023) and inconsistent with our other datasets. This inconsistency may arise from two main issues:

First, the dataset exhibits labeling irregularities, as seen in examples such as (21)-(23), which both we, the authors of this paper, and the best models reported in this study would categorize as abusive language, but which are labeled as non-abusive speech in the dataset. According to Ocampo et al. (2023) all instances labeled as non-abusive speech in the original datasets were not modified; only the instances labeled as abusive language were expanded. We assume that label inconsistencies may have arisen due to differences in the annotation guidelines followed by the individual datasets.

(21) little black kid gets his ass kicked - YouTube Because I always enjoy these videos.

(22) 88 In the end is victory

(23) Whiteteacher puts Negro in his place!

Second, as noted by Wiegand et al. (2021b), when datasets like *ISHate* combine data from different domains, they become stylistically diverse and less coherent, which can inadvertently lead classifiers to detect style rather than the substance of abuse, thereby reducing the dataset's reliability. This issue is particularly pronounced when there are disparities in class distributions across the combined datasets, potentially skewing the classifier's learning process. If datasets with significantly different proportions of abusive and non-abusive language are merged, a classifier might learn to differentiate between the datasets themselves rather than accurately distinguishing between abusive and non-abusive speech (Wiegand et al., 2021b). Previously reported supervised classifiers may be more prone to overfitting these artifacts compared to the zero-shot and few-shot approaches we examined in this paper.

Another limitation of this work is its time sensitivity; as newer LLMs are developed, the conclusions drawn here could quickly become outdated.

Furthermore, the vast array of possible prompts introduces variability in outcomes, which this study could not fully explore due to time constraints.

Moreover, recent research has identified data leakage as a concern when evaluating LLMs on publicly available datasets (Li et al., 2024). While the datasets in our study are publicly accessible, potentially allowing inclusion in model training and artificially enhancing performance, we consider data leakage unlikely. Overlap would likely result in higher performance than observed. Furthermore, the datasets are specific and small, making them impractical for fine-tuning.

Lastly, the use of closed models like GPT-4 raises concerns[8] about transparency and reproducibility, as proprietary data and methods limit the ability to validate results independently.

8 Ethics Statement

This research shows that LLMs can replace human annotators for the detection of both IAL and linguistic features predictive of IAL. This could negatively impact the workforce of human annotators, who often rely on microtask income for financial stability, particularly in economically disadvantaged regions (Posch et al., 2022; Ross et al., 2010). However, using LLMs could also alleviate the psychological burden on human annotators who are frequently exposed to harmful content, reducing emotional distress (Steiger et al., 2021).

Besides, the democratization of LLMs could lower the barriers to entry for smaller organizations by reducing the cost and complexity of data annotation, promoting innovation and equality (Ding et al., 2022). However, LLMs may perpetuate biases from their training data, potentially amplifying harmful stereotypes (Bender et al., 2021; Wang et al., 2023). To address these biases, diverse training data and continuous monitoring of model outputs are essential, though proprietary restrictions limit transparency in the models' training data.

Acknowledgements

The authors were partially supported by the Austrian Science Fund (FWF): P 35467-G.

References

Jawid Ahmad Baktash and Mursal Dawodi. 2023. GPT-4: A Review on Advancements and Opportunities in Natural Language Processing. *ArXiv*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, Virtual Event, Canada. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

[8] https://hackingsemantics.xyz/2023/closed-baselines/

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *ArXiv*, abs/2303.00293.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Albert Li. 2022. Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu

Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Association for Computing Machinery Computing Surveys*, 51(4).

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1320, Online.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social identities*, 22(3):324–341.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Hugo Lewi Hammer. 2017. Automatic detection of hateful comments in online discussion. In *Industrial*

Networks and Intelligent Systems, pages 164–173, Cham. Springer International Publishing.

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.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Empirical Methods in Natural Language Processing*.

Brian Mullen and Joshua M. Smyth. 2004. Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine*, 66:343–348.

Karsten Müller and Carlo Schwarz. 2017. Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153, Republic and Canton of Geneva, Switzerland.

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

OpenAI, :, 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, Mo 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, 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 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 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. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI Playground. 2024. Playground. https://platform.openai.com/playground. Accessed: 2024-04-25.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Lisa Posch, Arnim Bleier, Fabian Flöck, Clemens Lechner, Katharina Kinder-Kurlanda, Denis Helic, and Markus Strohmaier. 2022. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *Human Computation*, 9(1):22–57.

Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 3027–3040, Seattle, WA, USA.

Replicate. 2024. Run llama 3 with an api. https://replicate.com/blog/run-llama-3-with-an-api?input=python. Accessed: 2024-08-26.

Joel Ross, Lilly Irani, Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 2863–2872, New York, NY, USA. Association for Computing Machinery.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Miriah Steiger, Timir Bharucha, Sukrit Venkatagiri, Martin Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL – Student Research Workshop*, pages 88–93, San Diego, CA, USA.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612, Seattle, United States. Association for Computational Linguistics.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 16280–16297, Singapore. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double perturbation: On the robustness of robustness and counterfactual bias evaluation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3899–3916, Online. Association for Computational Linguistics.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3952–3958, New York City, NY, USA.

A Appendix

This appendix offers additional details on specific aspects of our research that could not be fully covered in the main paper due to space limitations. It is optional and serves as illustrative material that is not essential for understanding the main content of the paper.

A.1 LLM Hyperarameter Settings

In our experiments, we utilized the LLMs in a standard, potentially optimal configuration as envisioned by the platform providers (OpenAI Playground; Replicate, 2024). Table 11 shows the selected settings.

Model	Access Method	Temper.	Top p
GPT-3.5-turbo	OpenAI API	1.0	1.0
GPT-4	OpenAI API	1.0	1.0
LLaMA-3 70B	Replicate API	0.6	0.9

Table 11: Overview of model settings and access methods.

A.2 Additional Results for LLM-Driven Direct Detection of IAL

Tables 12-14 present the results of the direct detection of IAL for each of the three selected datasets: Identity Groups, Euphemistic Abuse and Comparisons. The scores in these tables were used to calculate the averages presented in Table 2.

Identifier	LLM	Zero-shot	Few-shot
hateful	GPT-3.5	62.63	69.69
insulting	GPT-3.5	62.09	66.78
offensive	GPT-3.5	64.91	71.95
toxic	GPT-3.5	65.97	69.80
abusive	GPT-3.5	55.65	70.54
hateful	GPT-4	84.58	86.29
insulting	GPT-4	83.23	84.21
offensive	GPT-4	81.26	82.37
toxic	GPT-4	82.74	83.54
abusive	GPT-4	79.53	84.78
hateful	LLaMA-3	80.57	82.83
insulting	LLaMA-3	79.04	82.58
offensive	LLaMA-3	79.38	82.31
toxic	LLaMA-3	80.63	81.92
abusive	LLaMA-3	80.24	82.76

Table 12: Direct detection of IAL on the dataset *Identity Groups*: Individual macro-averaged F1 scores for all five identifiers.

Identifier	LLM	Zero-shot	Few-shot
hateful	GPT-3.5	60.18	58.64
insulting	GPT-3.5	60.27	57.47
offensive	GPT-3.5	59.76	58.73
toxic	GPT-3.5	57.45	59.49
abusive	GPT-3.5	60.23	58.29
hateful	GPT-4	70.33	76.11
insulting	GPT-4	76.57	76.71
offensive	GPT-4	76.16	78.01
toxic	GPT-4	75.37	78.48
abusive	GPT-4	72.59	78.36
hateful	LLaMA-3	68.93	73.31
insulting	LLaMA-3	68.82	73.23
offensive	LLaMA-3	71.62	75.63
toxic	LLaMA-3	70.71	73.19
abusive	LLaMA-3	72.85	75.73

Table 13: Direct detection of IAL on the dataset *Euphemistic Abuse*: Individual macro-averaged F1 scores for all five identifiers.

A.3 Additional Results for Feature Combination through Logistic Regression

Tables 15-17 show the results of individual classifiers that we built in which features were combined through logistic regression. The specific results were used to calculate the percentage changes in macro-averaged F1 scores displayed in Tables 7-9.

A.4 LLM-Driven Direct Detection of IAL

Table 18 shows the final prompt template for the direct detection of IAL with LLMs in the few-shot approach. LLMs are prompted to classify each sentence as either abusive or non-abusive lan-

Identifier	**LLM**	**Zero-shot**	**Few-shot**
hateful	GPT-3.5	59.39	57.60
insulting	GPT-3.5	64.71	59.22
offensive	GPT-3.5	58.90	57.67
toxic	GPT-3.5	60.49	59.59
abusive	GPT-3.5	57.12	56.36
hateful	GPT-4	69.09	76.18
insulting	GPT-4	71.74	73.91
offensive	GPT-4	75.93	74.93
toxic	GPT-4	74.76	75.01
abusive	GPT-4	71.55	75.98
hateful	LLaMA-3	69.05	70.24
insulting	LLaMA-3	69.75	70.30
offensive	LLaMA-3	70.30	72.35
toxic	LLaMA-3	69.43	70.34
abusive	LLaMA-3	73.13	72.77

Table 14: Direct detection of IAL on the dataset *Comparisons*: Individual macro-averaged F1 scores for all five identifiers.

Classifier	**LLM**	**Zero-shot**	**Few-shot**
one identifier	GPT-3.5	64.91	71.95
comb. identifiers	GPT-3.5	66.89	71.32
ling. feats.	GPT-3.5	72.08	62.47
ling. feats. + comb. identifiers	GPT-3.5	74.53	71.92
one identifier	GPT-4	81.26	82.37
comb. identifiers	GPT-4	84.26	86.30
ling. feats.	GPT-4	81.20	85.78
ling. feats. + comb. identifiers	GPT-4	85.33	**86.86**
one identifier	LLaMA-3	79.38	82.31
comb. identifiers	LLaMA-3	80.20	82.98
ling. feats.	LLaMA-3	80.62	84.39
ling. feats. + comb. identifiers	LLaMA-3	84.01	85.32
best previously reported result		77.7	

Table 15: Dataset *Identity Groups*: Comparison of macro-averaged F1 scores for direct detection of IAL for the identifier *offensive* (*one identifier*), logistic regression models (*combined identifiers*, *linguistic features* and *linguistic features* + *combined identifiers*) and the best previously reported result taken from Wiegand et al. (2022).

Classifier	**LLM**	**Zero-shot**	**Few-shot**
one identifier	GPT-3.5	59.76	55.37
comb. identifiers	GPT-3.5	61.69	61.49
ling. feats.	GPT-3.5	60.36	61.27
ling. feats. + comb. identifiers	GPT-3.5	62.32	62.26
one identifier	GPT-4	76.16	76.38
comb. identifiers	GPT-4	76.47	79.00
ling. feats.	GPT-4	69.74	72.88
ling. feats. + comb. identifiers	GPT-4	78.01	**79.62**
one identifier	LLaMA-3	71.62	75.63
comb. identifiers	LLaMA-3	71.96	75.94
ling. feats.	LLaMA-3	70.72	72.56
ling. feats. + comb. identifiers	LLaMA-3	74.68	76.69
best previously reported result		75.60	

Table 16: Dataset *Euphemistic Abuse*: Comparison of macro-averaged F1 scores for direct detection of IAL for the identifier *offensive* (*one identifier*), logistic regression models (*combined identifiers*, *linguistic features* and *linguistic features* + *combined identifiers*) and the best previously reported result taken from Wiegand et al. (2023).

Classifier	**LLM**	**Zero-shot**	**Few-shot**
one identifier	GPT-3.5	58.90	64.95
comb. identifiers	GPT-3.5	65.65	62.42
ling. feats.	GPT-3.5	66.45	58.52
ling. feats. + comb. identifiers	GPT-3.5	66.10	63.46
one identifier	GPT-4	75.93	74.66
comb. identifiers	GPT-4	75.95	76.41
ling. feats.	GPT-4	64.07	64.61
ling. feats. + comb. identifiers	GPT-4	75.31	**75.60**
one identifier	LLaMA-3	70.30	72.35
comb. identifiers	LLaMA-3	73.13	72.73
ling. feats.	LLaMA-3	69.31	72.62
ling. feats. + comb. identifiers	LLaMA-3	72.38	74.01
best previously reported result		72.90	

Table 17: Dataset *Comparisons*: Comparison of macro-averaged F1 scores for direct detection of IAL for the identifier *offensive* (*one identifier*), logistic regression models (*combined identifiers*, *linguistic features* and *linguistic features* + *combined identifiers*) and the best previously reported result taken from Wiegand et al. (2021a).

guage based on whether the content matches any of the specified identifiers: *hateful*, *abusive*, *offensive*, *toxic* or *insulting*. The examples used in this prompt template are taken from the annotation guidelines and the paper presenting the dataset *Identity Groups* (Wiegand et al., 2022).

A.5 Prompt Templates for LLM-Driven Feature Extraction

Tables 19-21 present the prompt templates designed to extract the linguistic features discussed in §4.3 using LLMs.

A.6 Linguistic Features for Detecting IAL

Tables 22 and 23 describe the feature sets that were used for the datasets *Euphemistic Abuse* (Wiegand et al., 2023) and *Comparisons* (Wiegand et al., 2021a), respectively. The feature set for the dataset *Identity Groups* only comprises three features and they have already been sufficiently described in the main paper (§3 & §4.5).

Prompt Template
Is this sentence {identifier}? “{sentence}” Answer either “YES” or “NO”. No other answer permitted. Here are examples: Sentence: “Islam forces societies to regress.” Answer: “YES” Sentence: “Jews are concerned by the recent antisemitic attacks in Europe.” Answer: “NO” Sentence: “I cannot believe how perverted and sick Jews are.” Answer: “YES” Sentence: “Muslims fight against prejudice.” Answer: “NO” Sentence: “Women are very manipulative.” Answer: “YES” Sentence: “Gay people dislike being stereotyped.” Answer: “NO” Sentence: “Gay people are sooo dramatic.” Answer: “YES” Sentence: “Women fight against domestic violence.” Answer: “NO” Sentence: “Put on a wig and lipstick and be who you really are.” Answer: “YES” Sentence: “In Poland, women are protesting against the government’s prohibition on abortion.” Answer: “NO”

Table 18: Prompt template used for direct detection of IAL with LLMs in a few-shot learning approach. The model is prompted to respond with either *YES* or *NO* based on whether the content matches any of the specified identifiers: *hateful*, *abusive*, *offensive*, *toxic* or *insulting*.

Feature	Prompt
Non-conformist	Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence “{sentence}” Does the author of this sentence think that {target} are non-conformist? Answer either YES or NO. No other answer permitted.
Non-episodic	Classify the following sentence as either “EPISODIC” or “NON-EPISODIC”: “{sentence}” No other answer permitted. Answer “EPISODIC” if the sentence expresses information about a singular, distinctive or particular event. Answer “NON-EPISODIC” if the sentence expresses information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal or standard.
Perpetrator	A perpetrator is a person who commits an illegal, criminal, harmful or evil act. Consider the following sentence: “{sentence}” Does the author of this sentence think that {target} are perpetrators? Answer either YES or NO. No other answer permitted.

Table 19: Dataset *Identity Groups*: List of prompts used for LLM-driven extraction of linguistic features for the detection of IAL.

Feature	Prompt
Negated Antonym	Does the following sentence use the negated antonym of an abusive word? “{sentence}” Answer either “yes” or “no”; no other answers are permitted.
Extremes	Does the following sentence use extreme or absolute language such as superlatives, generalizations or hyperbole? “{sentence}” Answer either “yes” or “no”; no other answers are permitted.
Lexicalization	Does the following sentence use derogatory idioms that one could also potentially find in a dictionary? “{sentence}” Answer either “yes” or “no”; no other answers are permitted.
Opposing Sentiments	The pairing of opposing sentiments in a sentence is typically a contradiction and although it may appear positive on the surface (since the positive polar expression is usually the more salient expression), it is often meant in a derogatory way. Given this definition, does the following sentence use a pairing of opposing sentiments? “{sentence}” Answer either “yes” or “no”; no other answers are permitted.
Taboo	Taboo is a proscription of behavior that affects everyday life. Topics considered taboo include: bodies and their effluvia (sweat, snot, feces, menstrual fluid, etc.); the organs and acts of sex, micturition, and defecation; diseases, death and killing (including hunting and fishing); naming, addressing, touching and viewing persons and sacred beings, objects and places; and food gathering, preparation and consumption. Based on this definition, does the following sentence address a taboo topic? “{sentence}” Answer either “yes” or “no”; no other answers are permitted.
Unusual Properties	We define unusual utterances as sentences where the addressed person is attributed unusual properties or displays some unusual behavior. This could be strange hobbies, preferences or beliefs. The addressed person could also cause unusual situations or events or unusual behavior on the part of the speaker. The unusual property may also be conveyed by the usage of non-standard language, i.e. unusual imagery or some creative wording. The intention of the speaker is to alienate the addressed person from the reader. Based on this definition, does the following sentence describe an unusual property, behavior or situation? “{sentence}” Answer either “yes” or “no”; no other answers are permitted.

Table 20: Dataset *Euphemistic Abuse*: List of prompts used for LLM-driven extraction of linguistic features for the detection of IAL.

Feature	Prompt
Absurd	An absurd sentence is a sentence that describes an image that is extremely rarely or never observed in real life. Based on this definition, is the following sentence absurd? "{sentence}" Answer with either "yes" or "no"; no other answers are permitted.
Contradiction	Is there a contradiction in the following sentence? "{sentence}" Answer either "yes" or "no"; no other answers are permitted.
Dehumanization	A dehumanizing comparison is defined as directly comparing a person or their inherent mental or physical attributes to a non-human entity. Based on this definition, is the following comparison dehumanizing? "{sentence}" Answer either "yes" or "no"; no other answers are permitted.
Evaluation vs. Emotional Frame of Mind	An evaluative comparison involves the author negatively assessing a specific aspect of the addressed person (the addressed person is determined by the second-person pronoun "you"), often by criticizing their behavior or outward appearance. On the other hand, a non-evaluative comparison describes the emotional state of the addressed person without necessarily passing judgment. Based on these definitions, is the following sentence evaluative? "{sentence}" Answer either "yes" or "no"; no other answers are permitted.
Figurativeness vs. Literalness	Is the following comparison figurative? "{sentence}" Answer "yes" or "no"; no other answers are permitted.
Taboo	Taboo is a proscription of behavior that affects everyday life. Topics considered taboo include: bodies and their effluvia (sweat, snot, feces, menstrual fluid, etc.); the organs and acts of sex, micturition and defecation; diseases, death and killing (including hunting and fishing); naming, addressing, touching and viewing persons and sacred beings, objects and places; and food gathering, preparation and consumption. Based on this definition, does the following sentence address a taboo topic? "{sentence}" Answer either "yes" or "no"; no other answers are permitted.

Table 21: Dataset *Comparisons*: List of prompts used for LLM-driven extraction of linguistic features for the detection of IAL.

Feature	Description	Examples
Negated Antonym	Negated antonyms use negation to soften or disguise insults by stating the opposite of an overtly negative term. They employ words like *not* or *lack* to subtly convey a negative meaning, making the insult less direct while still implying the same derogatory intent.	You are not beautiful. There is nothing of interest in your life. You lack humility.
Opposing Sentiments	Opposing sentiments involve pairing contradictory emotions in a sentence to create a sarcastic or derogatory effect. While the expression may seem positive due to a prominent positive phrase, the overall intention is often negative or abusive, aiming to provoke a specific reaction from the reader.	You are excellent at breaking things. You must love having people hate you. You are unique in your ability to disappoint.
Taboo	Abusive language frequently uses words linked to taboo subjects, such as physical or mental abnormalities, particular body parts, death, etc. to convey offensiveness.	I'd prefer you were in a grave. You would fit well in a brothel. Your smell greeted me five minutes before you arrived.
Extremes	Extreme or absolute language can manifest through various linguistic forms, such as the use of superlatives, generalizations or hyperbole.	You are truly the best at doing nothing. You are not very good at anything. If you get any thinner, you'll be transparent.
Lexicalizat.	Lexicalizations are derogatory idioms that could potentially be found in dictionaries.	You are not the sharpest tool in the box. You are a thorn in my side. You don't have a backbone.
Unusual Properties	This feature involves attributing odd or uncommon traits, behaviors or situations to a person to create a derogatory or mocking effect. These attributions might include peculiar hobbies, strange preferences, or causing unusual events, often conveyed through imaginative language or creative phrasing.	Your main hobby must be letting life pass you by. Your heart made an iceberg look warm. You are the leader of Boredville.

Table 22: Linguistic features for the detection of euphemistically abusive language identified by Wiegand et al. (2023).

Feature	**Description**	**Examples**
Figurativeness vs. Literalness	Figurative comparisons involve entities that are fundamentally different and cannot be reversed without changing the meaning. In contrast, literal comparisons are reversible and highlight prominent shared properties between the compared entities. Literal comparisons maintain their meaning when their components switch places, whereas figurative comparisons do not.	Encyclopedias are like goldmines. (Figurative) Encyclopedias are like dictionaries. (Literal) Your words are like fire. (Figurative) You have the face of a sad person. (Literal)
Dehumanization	A dehumanizing comparison involves directly comparing a person or their intrinsic mental or physical traits to a non-human entity.	You walk like a giraffe. You sing like a dying bird. Your eyes are like a sack of potatoes.
Taboo	Abusive language frequently uses words linked to taboo subjects, such as physical or mental abnormalities, particular body parts, death, etc. to convey offensiveness.	You eat like you have worms. You are sweating like a dog in heat. You make me feel like bringing up my lunch.
Absurd Images	This feature refers to descriptions of scenes that are very rarely or never observed in reality.	You behave like a toddler on acid. Your manners are like a bull in a china shop. You cook like you read the instructions backwards.
Contradiction	This feature involves comparisons that use characteristics directly opposing the typical traits of the entities being compared. These contradictions, often seen as a form of sarcasm, convey abuse by highlighting a quality that starkly contrasts with the usual attributes of the subject.	You are as thin as an elephant. You are as smart as a Neanderthal. You are as modern as a caveman.
Evaluation vs. Emotional Frame of Mind	Evaluative comparisons involve a negative judgment about a specific trait or behavior of the target, such as criticizing their appearance or actions, and are often seen as abusive. In contrast, comparisons that describe the emotional state of the target, such as indicating pain or exhaustion, do not necessarily imply criticism and are less likely to be perceived as abusive.	You look like an overfed cat. (Evaluation) You look like a shocked cat. (Frame of mind) You walk like a giraffe. (Evaluation) You look like you're lost. (Frame of mind)

Table 23: Linguistic features for the detection of implicitly abusive comparisons identified by Wiegand et al. (2021a).