European Journal of Political Research (2026), 1–12
doi:10.1017/S1475676525100546

RESEARCH NOTE

Improving hate speech detection with large language models

Natalia Umansky[1], Maël Kubli[1], Ana Kotarcic[2], Laura Bronner[3], Selina Kurer[3], Philip Grech[3], Dominik Hangartner[3], Fabrizio Gilardi[1] and Karsten Donnay[1]

[1]Department of Political Science, University of Zurich, Zürich, Switzerland, [2]Department of History, University of Zurich, Zürich, Switzerland and [3]Immigration Policy Lab, ETH Zurich, Zürich, Switzerland
Corresponding author: Fabrizio Gilardi; Email: gilardi@ipz.uzh.ch

(Received 8 May 2024; revised 24 July 2025; accepted 30 September 2025)

Abstract
Efforts to curb online hate speech depend on our ability to reliably detect it at scale. Previous studies have highlighted the strong zero-shot classification performance of large language models (LLMs), offering a potential tool to efficiently identify harmful content. Yet for complex and ambivalent tasks like hate speech detection, pre-trained LLMs can be insufficient and carry systemic biases. Domain-specific models fine-tuned for the given task and empirical context could help address these issues, but, as we demonstrate, the quality of data used for fine-tuning decisively matters. In this study, we fine-tuned GPT-4o-mini using a unique corpus of online comments annotated by diverse groups of coders with varying annotation quality: research assistants, activists, two kinds of crowd workers, and citizen scientists. We find that only annotations from those groups of annotators that are better than zero-shot GPT-4o-mini in recognizing hate speech improve the classification performance of the fine-tuned LLM. Specifically, fine-tuning using the highest-quality annotator group – trained research assistants – boosts classification performance by increasing the model's precision without notably sacrificing the good recall of zero-shot GPT-4o-mini. In contrast, lower-quality annotations do not improve and may even decrease the ability to identify hate speech. By examining tasks reliant on human judgment and context, we offer insights that go beyond hate speech detection.

Keywords: hate speech; text analysis; large language models

Introduction

The regulation of online speech is not merely a technical matter – it is a normative and political one. Hate speech in particular has become a pressing concern in the digital age, necessitating not only scalable detection mechanisms but also critical reflection on how such content is governed, regulated, and contested within broader democratic frameworks. Regarding the ability to identify such content on a large scale, recent advancements in Natural Language Processing (NLP), especially the rapid progress in the area of Large Language Models (LLMs), present new opportunities for addressing this issue and efficiently (Kotarcic et al. 2022; Parker and Ruths 2023). These models have demonstrated significant zero-shot capabilities across various NLP tasks, highlighting their potential utility in flagging online harmful content at scale (Gilardi et al. 2023; Kasneci et al. 2023). Yet, reliably detecting hate speech and other harmful online content using automated approaches remains challenging.

© The Author(s), 2026. Published by Cambridge University Press on behalf of European Consortium for Political Research. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

Hate speech refers to toxic (that is, abusive, pejorative, or otherwise harmful) content that targets individuals as representatives of (protected) groups (United Nations 2020, 8). This definition situates hate speech within the broader category of harmful speech, distinguishing it from toxic speech that is offensive but lacks group-based targeting. We refer to speech that is neither hate speech nor toxic as non-harmful. Hateful statements may be (intentionally) ambivalent or subtle, and there is typically no reliable ground-truth for what constitutes offending content. Moreover, pre-trained LLMs may carry, and even exacerbate, systemic biases, driven by the characteristics of the datasets they were trained on (Ferrara 2023; Kasneci et al. 2023; Ray 2023; Shaikh et al. 2022). A domain-specific model fine-tuned for the task of detecting hate speech and sensitive to the particular context can help address these issues and improve model performance, even with a small number of annotations (Fui-Hoon Nah et al. 2023; Kasneci et al. 2023; Kocoń et al. 2023). In this paper, we contribute to these efforts by fine-tuning GPT-4o-mini on a diverse corpus of online comments in German, each annotated by coders from five groups with varying annotation quality: research assistants, activists, two kinds of crowd workers, and citizen scientists. With expert annotations as the gold standard for what constitutes hate speech, we study how the performance of fine-tuned GPT-4o-mini varies, relative to the zero-shot model, depending on the quality of the annotations used for fine-tuning.

We first demonstrate that annotation quality for what constitutes hate speech notably differs across our annotator groups. This includes significant variation in inter-coder agreement across groups. In addition, the agreement of different groups with the expert-based gold standard varies strongly between the best-performing group, trained research assistants, and the worst-performing group, citizen scientists. These differences in annotation quality then directly influence whether the performance of the LLM can actually be improved through fine-tuning or not. Specifically, we identify two key characteristics that set high-quality fine-tuning data apart.

First, only when using annotations from groups with a balanced ability to recover the gold standard, ie with relatively similar precision and recall, does fine-tuning enhance or at least match the zero-shot performance of GPT-4o-mini. Second, and most importantly, we find that only fine-tuning with data from groups that are better than zero-shot GPT-4o-mini at recognizing hate speech actually boosts classification performance. Specifically, fine-tuning using the highest-quality annotator group – trained research assistants – noticeably improves classification performance compared to zero-shot GPT-4o-mini by increasing precision without sacrificing recall. In practice, deciding when harmful content is targeted or not can be difficult even for well-trained experts. In all of our analyses, we therefore first consider harmful vs. non-harmful speech more broadly, and then treat hate speech (targeted attacks) and toxic speech (non-targeted attacks) as distinct forms of harmful speech. Our findings hold across both specifications.

Taken together, our results underscore the nuanced impact of data quality when fine-tuning LLMs to enhance hate speech detection. We demonstrate both that GPT-4o-mini performance can be noticeably improved by fine-tuning using only a small number of labeled data and that the quality of annotations is crucial. Fine-tuning helps when human coders perform better than zero-shot GPT-4o-mini, but it has no impact or may even hurt classification performance otherwise. In line with the emerging literature on supervised fine-tuning of LLMs for real-world classification tasks (Alizadeh et al. 2025; Ma et al. 2023), this finding suggests a move away from sheer quantity of labels to focusing on their quality instead. While data quality is a known concern in practice and has been widely acknowledged in the informal literature, there is to date a dearth of systematic evaluations of its impact. Our insights into what exactly constitutes 'high-quality' data and how this then translates to the performance of fine-tuned LLMs are, therefore, not only methodologically highly relevant but also have direct implications for their real-world application.

Enhancing the quality and precision of detection is a critical prerequisite for any kind of intervention that aims to manage the harms of online discourse, acknowledging that such interventions are inherently political and shaped by contested norms and power relations. The

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

development of task-specific and context-sensitive machine learning classifiers for hate speech detection demands significant time, funding, and computing resources, along with extensive training data (Laurer et al. 2024). In contrast, LLMs can be rapidly deployed and efficiently fine-tuned to achieve excellent classification performance, as long as comparably small but high-quality datasets are used. Given the small size of datasets needed for fine-tuning, it is feasible to rely on the comparably more costly annotations from trained research assistants to enable the fine-tuning of LLMs for complex classification tasks. Crowd work annotations, in comparison, are more affordable but, as we demonstrate, not of sufficient quality to boost classification performance through fine-tuning. Similarly, labels from volunteer contributors, such as citizen scientists or NGO activists, are also of insufficient quality.

While the primary aim of this study is to improve hate speech detection through fine-tuning LLMs, we also seek to draw broader insights into fine-tuning methodologies for tasks characterized by ambiguity and subjectivity. Hate speech, as a socially and politically constructed category, resists fixed definitions, making it an ideal case to examine how fine-tuning with varied annotation quality influences model performance. By empirically analyzing how different types of human judgment affect the fine-tuning of LLMs, we provide practical guidance for using machine learning in settings where contextual sensitivity is essential. In doing so, we also contribute to ongoing debates about the societal implications of automated content moderation and the role of human judgment in shaping its outcomes.

Concepts, data, and methods

Conceptual foundations

The term 'hate speech' is inherently contested (Gagliardone et al. 2015). Philosophical work treats hate speech not merely as offensive talk but as a practice that undermines equal civic standing (Waldron 2014). From a speech-act angle, certain forms of speech can subordinate targets by enacting social hierarchies rather than just describing them (Langton 1993). This insight can be refined by distinguishing constitutive harms (the very act of denigration) from consequential harms (downstream violence or exclusion) (Maitra and McGowan 2012). Together, these strands motivate treating hate speech as harmful in itself, even when no immediate violent outcome occurs. Building on these accounts, and consistent with the United Nations Strategy and Plan of Action on Hate Speech (United Nations 2020), we define hate speech as any communication that attacks or uses pejorative or discriminatory language with reference to a protected characteristic.

Operationalizing hate speech

Distinguishing between hate speech targeting individuals as representatives of (protected) groups and non-targeted toxic speech is conceptually important but can in practice be very challenging, even for highly trained expert coders. Following Kotarcic et al. (2022), we use a two-step coding protocol designed to make this distinction as explicit as possible. The protocol is used for all groups and our expert annotators, but also in the LLM-based classifications. In the first step, the protocol distinguishes *non-harmful speech* from *harmful speech*. In the second step, it then distinguishes within the broader category of harmful speech between targeted *hate speech* and non-targeted *toxic* speech based on whether a target can be identified. The first step yields a broader (binary) distinction between harmful and non-harmful speech, whereas the first and second steps jointly produce a more fine-grained (multi-class) categorization in three categories (hate speech, toxic speech, and non-harmful speech). We consider the performance for both tasks separately throughout our analysis, allowing us to test how the complexity of the tasks (for annotators and the LLM) influences what can be achieved by fine-tuning.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

Hate speech corpus

This study leverages a unique dataset of online comments in German encompassing a diverse set of hateful and non-hateful content. The comments stem from the German-language editions of two large Swiss newspapers that operate with relatively strict moderation guidelines. By collaborating directly with their community-management teams, we gained access not only to comments that were publicly visible, but also to entries the moderators had already hidden or deleted. The corpus covers a broad range of salient Swiss topics and inherently reflects the pluralistic nature of Swiss politics. At the same time, the data remain culturally coherent: all comments are in German, share the same domestic legal constraints, and are produced under the outlets' house rules.

Overall, the dataset includes annotations by 204 individuals across five distinct groups as follows: research assistants (3 people), crowd workers from Appen[1] (17 people) and Prolific[2] (53 people), individuals participating in a Citizen Science challenge (24 people), and a large group of activists in the volunteer community of a partner NGO (107 people). The selection of coder groups originated from a collaboration with the NGO.[3] While the activists' contributions were relevant for understanding how motivated, non-professional annotators perform, their annotations proved insufficient for our research needs, prompting the inclusion of alternative groups with varying levels of training and experience.

Each comment in our data was independently annotated by at least three individual coders from each group. Given the lack of an objective ground truth, three experts, all co-authors of this paper, annotated a gold standard set in line with the hate speech definition adopted in this paper. One of the three experts has longstanding experience with marginalized groups, allowing for a more balanced annotation of the ground truth dataset. In this step, the experts first labeled each comment independently. The initial labels assigned by the experts agreed for close to 80% of comments (Krippendorff's alpha: 0.728). For the remaining 20%, the experts deliberated until they reached full consensus on each comment.[4] The harmonized labels attained from this procedure were employed as the study's gold standard. Even so, it is important to note that hate speech annotations are inherently subjective and that there is no such thing as an absolute ground truth.

Our initial corpus consisted of 500 unique user-produced comments obtained from the discussion section of collaborating Swiss national media outlets and included published, moderated, and deleted comments (Kotarcic et al. 2022). The comments were posted under more than 60 different news articles spanning a broad range of topics, including COVID-19, international politics, current affairs, local news, sports, and lifestyle. Most comments are rather short, fully formulated sentences. They may include emojis, but no other visual content. Because of platform settings, six comments could not be consistently annotated across all groups, resulting in a total of 494 unique comments for analysis. To ensure consistency across annotator groups, we randomly sampled three annotations per comment for groups where some comments were annotated more than three times (notably Appen and the NGO), preserving the per-comment distribution of positive and negative labels. Finally, within each group, we determined a single label for each comment using majority rule, ie the majority label selected by at least two of three coders within each group.

[1]Appen, a data annotation company, acquired Figure Eight (formerly CrowdFlower) in 2019, which was a platform widely used in early crowd-sourced text analysis.

[2]Prolific is an online platform primarily used to recruit survey participants, which also supports text annotation tasks.

[3]The NGO is a politically independent organization with a focus on women's rights. While exact demographics are unavailable, the coders were likely predominantly (though not exclusively) women, largely white, with an over-representation of young adults.

[4]Even though they independently agreed on many instances, prior to deliberation, our experts did not always agree on whether speech was targeted and against whom. See Appendix 5 for details.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

Table 1. Comparison of annotation quality across various coder groups

	Research assistants	Prolific	Citizen scientists	Appen	NGO activists
Hate/Toxic Speech vs. Non-harmful Speech (binary)					
F1-score	0.84	0.78	0.76	0.65	0.63
Precision	0.86	0.81	0.80	0.62	0.87
Recall	0.82	0.74	0.72	0.68	0.49
Hate Speech vs. Toxic Speech vs. Non-harmful Speech (multi-class)					
F1-score	0.71	0.48	0.39	0.47	0.56
Precision	0.77	0.54	0.41	0.48	0.65
Recall	0.70	0.53	0.40	0.49	0.55

Note: All annotation quality measures are calculated relative to the expert-based gold-standard hate speech labels.

Annotation quality of human coders

When quantifying the variation in annotation quality among our human coders, we use standard metrics to both measure inter-coder agreement and the quality of annotations relative to the expert-based gold-standard labels. Specifically, we rely on Krippendorff's alpha to evaluate the level of consistency among multiple annotators when assigning labels to the same data set (Barberá et al. 2021; Van Atteveldt et al. 2021). In assessing consistency, we consider the binary (harmful speech vs. non-harmful speech, *bin*) and the multi-class (hate speech vs. toxic speech vs. non-harmful speech, *multi*) distinctions. The two most consistent groups of coders are the trained research assistants ($\alpha_{bin} = 0.506$, $\alpha_{multi} = 0.435$) and NGO activists ($\alpha_{bin} = 0.458$, $\alpha_{multi} = 0.433$). Crowd workers from Prolific ($\alpha_{bin} = 0.391$, $\alpha_{multi} = 0.391$) and the citizen scientists ($\alpha_{bin} = 0.335$, $\alpha_{multi} = 0.308$) are both less consistent. The by far lowest level of inter-coder agreement is found for crowd workers from Appen ($\alpha_{bin} = 0.080$, $\alpha_{multi} = 0.129$). The overall relatively low inter-coder agreement within all groups is in line with existing literature, emphasizing the inherent challenges in achieving consensus among annotators for subjective tasks like hate speech detection (Akhtar et al. 2020; Kralj Novak et al. 2022).

Evaluating the annotation quality for each group relative to the expert-based gold standard, we rely on three standard measures commonly used for quantifying annotation and classification performance: precision, recall, and F1 scores.[5] Taken together, these metrics provide a comprehensive picture of annotation quality, reflecting the reliability in identifying true instances of hate speech (precision), the ability to capture all such instances (recall), and the respective trade-offs (F1). Coding conducted by research assistants has the highest quality, coming closest to the gold-standard labels for the binary and multi-class annotations (Table 1). Crowd workers from Prolific follow in quality, but their annotations are significantly less accurate when distinguishing between (targeted) hate speech, (non-targeted) toxic speech, and non-harmful content.

Interestingly, annotations by NGO activists achieve the highest precision in recovering the gold-standard labels for the harmful vs. non-harmful speech distinction, but also the lowest recall. That is, they most accurately identify instances of harmful speech but fail to capture many of them. This suggests a solid understanding of the nuances of coding harmful content, which is also evident in the hate speech vs. toxic speech distinction, where they outperform Prolific crowd workers in precision and recall.[6] Quality drops notably in the other two groups. Citizen scientists perform well in distinguishing harmful from non-harmful speech but struggle to separately distinguish (targeted) hate speech and (non-targeted) toxic speech from non-harmful speech. Meanwhile, Appen crowd workers perform significantly worse than research assistants on both tasks.

[5]The F1 score is the harmonic mean of precision and recall, $F1 = 2 \cdot precision \cdot recall / (precision + recall)$.

[6]This may reflect the activist group's larger size and diverse labeling patterns: participants with higher thresholds for hate speech may dominate majority votes, leaving some hateful comments unlabeled. Their infrequent participation may also limit calibration over time, yielding high precision but lower recall.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

Taken together, the comparisons among coders and the level of agreement with the gold-standard labels demonstrate that our different annotator groups produced labels of significantly different quality. This is crucial in evaluating the impact of data quality on the fine-tuning of LLMs.

Fine-tuning GPT-4o-mini

A key strength of LLMs is that by virtue of their pre-training on huge amounts of data, they already perform well on a wide range of tasks without additional training (Brown et al. 2020; Liu et al. 2023). At the same time, the broad scope of the models typically implies a lack of context- and domain-specificity (Au Yeung et al. 2023; Yang et al. 2023). Fine-tuning here offers a way to fully leverage the strength of a broad pre-trained model while enhancing performance with specific nuances or domain-specific knowledge (Ouyang et al. 2022; Wei et al. 2022). On a technical level, fine-tuning adjusts a subset of the model's weight parameters based on additional data with task-specific labels (Hu et al. 2023). In particular, for tasks such as hate speech detection, which are intrinsically complex and subjective, fine-tuning also explicitly brings human perspectives into the loop. As a result, fine-tuning not only improves model performance but also addresses concerns regarding model-induced biases, which can be especially critical for tasks such as hate speech detection (Kasneci et al. 2023).

We use GPT-4o-mini with the standard temperature setting (Gilardi et al. 2023).[7] The model is prompted to classify a text input without any task-specific fine-tuning. It uses its general understanding of language and context, acquired from pre-training, to decide on the classification category. For more details on the classification, see Appendix 1. Fine-tuning of the model was then implemented on OpenAI's GPT-4o-mini architecture by using the openly available API.[8] Next to the model's zero-shot performance, ie the base GPT-4o-mini model performance without fine-tuning, we tested two different fine-tuning regimes: one using 100, the other 250 comments. We used the same two random sets of comments across all groups, assigning to each comment the respective group-specific (majority) label obtained from the annotations.

To quantify the hate speech classification performance of GPT-4o-mini, we rely on the same gold standard and same standard metrics introduced above: precision, recall, and F1. We also report confidence intervals obtained through bootstrapping as a measure of the robustness of the model's classification performance.

Results

The main findings are displayed in two figures. Figure 1 shows the performance of the different annotator groups, zero-shot GPT-4o-mini, and fine-tuned GPT-4o-mini for the binary harmful vs. non-harmful classification task, while Figure 2 does the same for the multi-class distinction between hate speech, toxic speech, and non-harmful speech. Our discussion focuses first on the zero-shot performance of GPT-4o-mini, and then on the performance of the fine-tuned model.

Zero-shot classification

The baseline performance of zero-shot GPT-4o-mini varies significantly between the broader harmful speech detection task and the narrower hate speech detection task. Even without prior task-specific training, it performs quite well on the binary harmful vs. non-harmful speech

[7]LLMs can generate some variation in outputs for identical instructions because of their probabilistic nature. In the zero-shot GPT-4o-mini classification, we prompted the model three times to classify each comment to account for this variation and use the majority label. The variations in comment classifications are minor.

[8]Additional information on data preparation and analysis for chat model fine-tuning can be found here: https://cookbook.openai.com/examples/chat_finetuning_data_prep.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

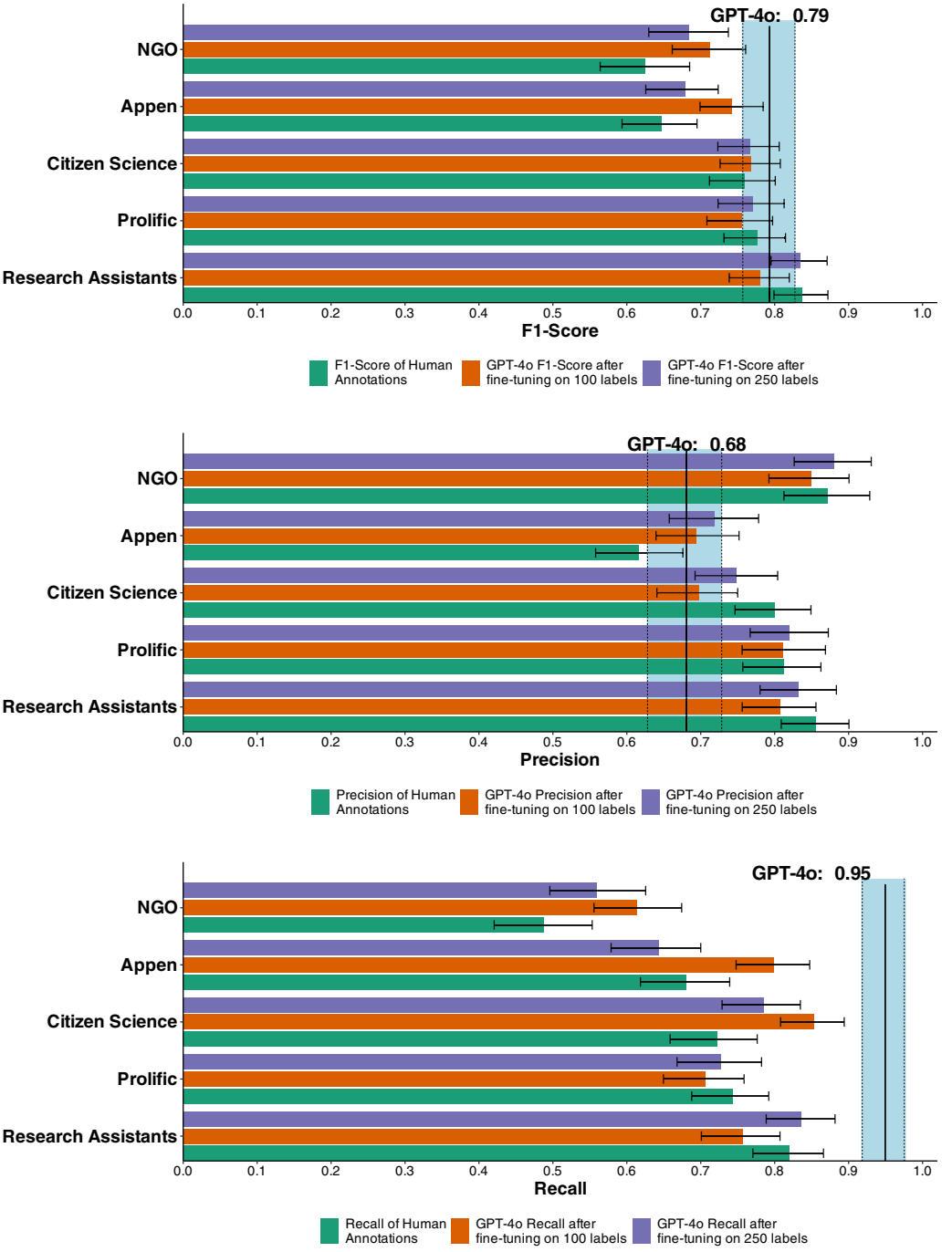


Figure 1. F1, precision, and recall of the fine-tuned GPT-4o-mini for binary classification of harmful vs. non-harmful speech across annotator groups, using 100 or 250 training labels. Performance is shown relative to the expert gold standard; vertical lines mark zero-shot performance. Green bars indicate label quality. Error bars show 95% bootstrapped confidence intervals.

distinction (vertical lines in Figure 1). With an F1 score of 0.79, it matches or outperforms all human annotator groups when compared to the expert-based gold standard, except for the trained research assistants, who achieve an F1 score of 0.84. However, its performance declines in the

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

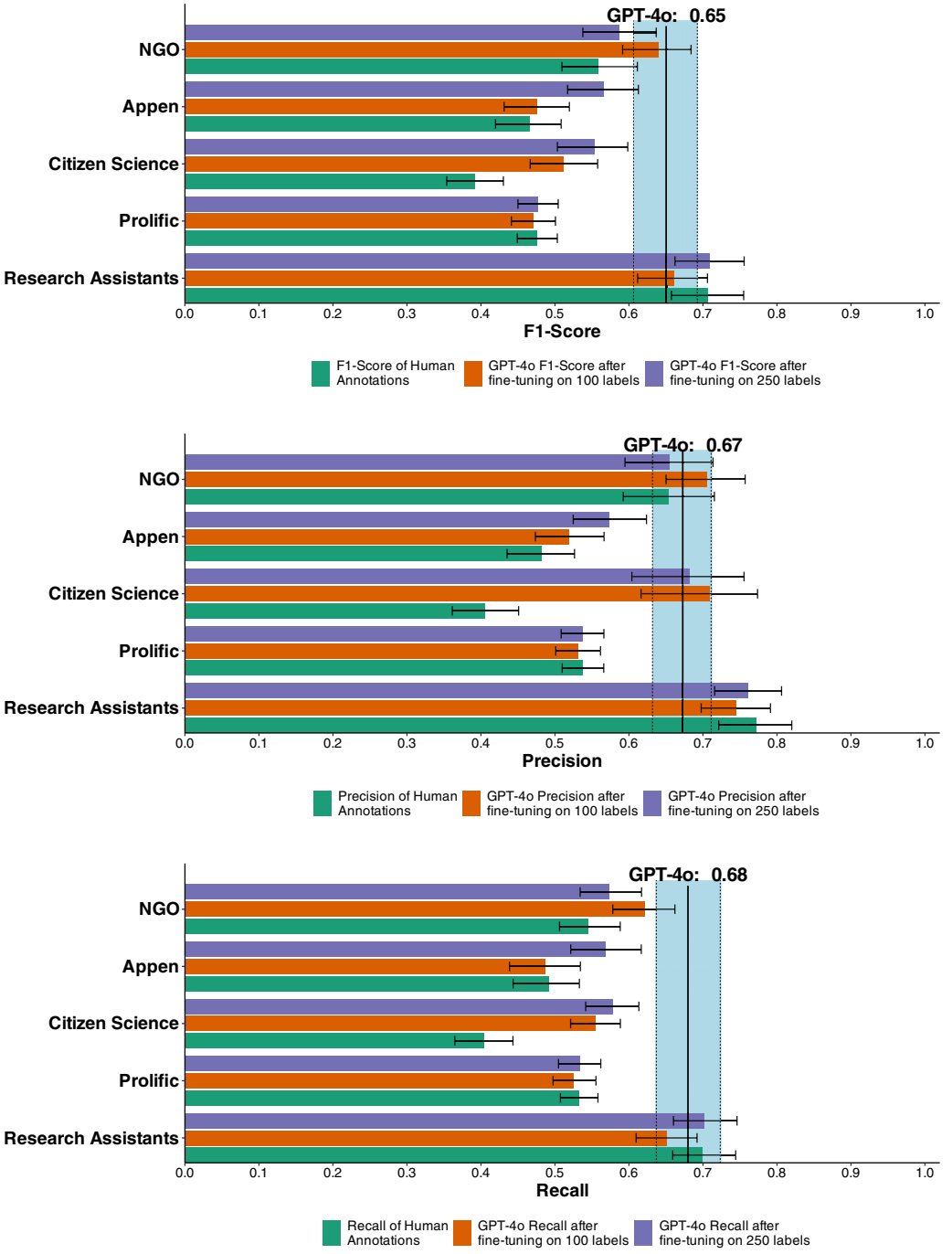


Figure 2. Average F1, precision, and recall of the fine-tuned GPT-4o-mini for classifying hate, toxic, and non-harmful speech across annotator groups, using 100 or 250 training labels. Performance is shown relative to the expert gold standard, alongside the label quality per group. Error bars indicate 95% bootstrapped confidence intervals.

multi-class classification task, which further distinguishes (targeted) hate speech from (non-targeted) toxic speech within harmful speech (vertical lines in Figure 2). Here, its F1 score of 0.65 still surpasses all human annotator groups except the trained research assistants (F1: 0.71).

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

The GPT-4o-mini zero-shot model exhibits a stark imbalance between precision (0.68) and recall (0.95) for the broader harmful vs. non-harmful speech distinction but not for the narrower hate speech detection task (precision: 0.67, recall: 0.68). In other words, its higher zero-shot performance in the broader task is driven entirely by its significantly higher recall. Specifically, the model correctly predicts harmful speech in 226 out of 238 instances, missing only 12 cases. However, this comes at the cost of a high number of false positives – 106 out of 332 predicted harmful speech instances are incorrect, which is consistent with its relatively low precision. In contrast, for the narrower hate speech detection task, the model captures true instances of hate speech (83 out of 148) and toxic speech (78 out of 94) reasonably well but frequently misclassifies non-harmful speech as either toxic or hate speech. This explains its relatively low precision and recall for this task (see Appendix 4.1).

Using classifiers with low precision can be problematic in real-world applications, as it may lead to a high rate of (over-)moderation, flagging comments that are not actually hate speech (or more broadly, harmful speech). Human annotator groups tend to have lower recall but higher precision, particularly in distinguishing between harmful and non-harmful speech, resulting in a more balanced annotation quality overall. This suggests that fine-tuning GPT-4o-mini with human labels should improve precision, ideally without a significant loss in recall. For the narrower hate speech detection task, we also expect fine-tuning to enhance performance, but only with annotations from research assistants – the only group whose labels are of sufficiently high quality to improve precision. In practice, this means a model fine-tuned on high-quality data should maintain its ability to identify relevant instances of hate (or harmful) speech while reducing instances of over-moderation.

Fine-tuned model

The fine-tuned GPT-4o-mini model demonstrates a clear improvement in performance compared to the zero-shot model for the broader harmful speech and the narrower hate speech classification tasks (horizontal bars in Figures 1 and 2).[9] The best-performing models enhance GPT-4o-mini's performance by up to five percentage points, particularly when fine-tuned with 250 labels (violet bars), achieving F1 scores of up to 0.83 for harmful vs. non-harmful speech and 0.71 for hate speech detection. These gains stem primarily from notable improvements in precision compared to the zero-shot model. The only consistently reliable annotation set for fine-tuning comes from the trained research assistants, who provide the highest annotation quality across groups. Notably, GPT-4o-mini does not seem to absorb much noise from the citizen science fine-tuning set, as indicated by minimal deviation from its zero-shot performance. This suggests that the model is relatively resistant to lower-quality annotations, likely because of its strong pre-trained decision boundaries.

A detailed analysis of model performance reveals that the overall improvement across both tasks is primarily driven by a reduction in false positive classifications in the fine-tuned model, without a corresponding increase in false negatives or a decrease in true positives[10]. The best-performing fine-tuned model closely matches the highest human annotation quality in both classification settings. For the broader harmful vs. non-harmful speech distinction, precision and recall are significantly more balanced in the fine-tuned model compared to the zero-shot GPT-4o-mini, indicating a more robust classifier. We also observe a notable precision gain when fine-tuning with lower-quality annotations, such as those from NGOs, which tend to have high

[9]The confusion matrices for all classifiers are provided in Appendix 4.

[10]Error analysis reveals a precision–recall trade-off: more training data reduces false positives but increases false negatives. Most misses involve nationality-based or toxic speech, while false positives often stem from non-hateful comments that mimic hateful structure. This highlights the need for high-quality annotations and careful category design when using LLMs to code hate speech. For more details, see Appendix 4.2.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

precision but low recall. However, this comes at the cost of reduced recall, leading to an overall decrease in F1-score and less robust performance than the zero-shot model.

While it may seem intuitive in hindsight that LLM fine-tuning is sensitive to differences in annotation quality, we did not anticipate the extent of this sensitivity, particularly given the strong performance of LLMs in zero-shot settings. Our study contributes to ongoing discussions on the usage of LLMs for text annotation by providing empirical evidence that annotation quality significantly influences fine-tuning outcomes. At the same time, we demonstrate that the differences in model performance are primarily attributable to the quality of the annotations provided by each group, particularly the balance between precision and recall. We expect these insights to hold in other language contexts. Taken together, the results underscore that in our context, with Swiss news sites that enforce comparatively strict ex-ante and ex-post moderation, fine-tuning with expert labels can make LLMs viable for assisted moderation pipelines, while scaling up with inexpensive crowd labels risks either over-blocking speech or under-detecting subtle hostility. The unique Swiss variation of German included in our study provides evidence of the fine-tuned model's ability to handle linguistic nuances, suggesting that similar performance might be observed in other languages with regional variations.

Conclusion

Efforts to address online hate speech and safeguard democratic discourse depend not only on scalable detection capacities but also on navigating the political and normative decisions that shape what is considered harmful, who decides, and how such speech is governed in pluralistic societies. Focusing on detection, this paper demonstrates the potential of fine-tuned, context- and task-specific LLMs to improve the identification of hate speech. However, as our study shows, the quality of fine-tuning data is crucial for the ability of LLMs like GPT-4o-mini to classify hate speech effectively. Simply fine-tuning with any data does not guarantee improved performance. In fact, we show that using low-quality annotations for fine-tuning can even degrade the model's zero-shot performance. Only high-quality data that appropriately balances the precision-recall trade-off leads to a meaningful boost in classification performance, regardless of coder group size.

A crucial aspect of NLP approaches, for hate speech detection and similar tasks, is the substantial number of annotations required to train context-specific classifiers (Barberá et al. 2021; Benoit et al. 2016; Boukes et al. 2020; Laurer et al. 2024; Van Atteveldt et al. 2021). The use of fine-tuned LLMs offers a shift away from relying solely on large quantities of labels, instead emphasizing fewer but higher-quality annotations. Our results suggest that high-quality human labels will remain essential for achieving the task- and context-sensitivity needed for LLMs to match high-quality human annotators' classification performance. As we demonstrate, this is reasonably achievable in practice: the amount of high-quality data required is small and can be generated at a reasonable cost and effort by trained research assistants to effectively enhance the model's performance.

We acknowledge that our approach, which relies on proprietary models, has certain limitations, particularly with respect to reproducibility and transparency, as well as data protection (Barrie et al. 2024; Ollion et al. 2024; Spirling 2023). However, recent work has demonstrated that open-source models can achieve similar or even superior performance, offering a way to address these limitations (Alizadeh et al. 2025). Therefore, although our study uses proprietary models, there are compelling reasons to believe that the findings generalize to LLMs that are better aligned with reproducibility, transparency, and data protection.

Fine-tuned LLMs have important real-world applications beyond hate speech detection. For instance, an emerging body of research explores the role of supervised fine-tuning in facilitating content moderation decisions more broadly (Ma et al. 2023). As with hate speech detection, this approach integrates LLMs with high-quality labels provided by human moderators. This hybrid

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

model leverages the strengths of both LLMs and human judgment, potentially enabling more accurate, fair, and nuanced content moderation at scale. However, further research is needed to establish clear guidelines and frameworks for the use of LLMs in content moderation.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S1475676525100546

Data availability statement. Replication materials are available at the Harvard Dataverse (https://doi.org/10.7910/DVN/U9ODTE).

Acknowledgments. We are grateful to the Public Discourse Foundation, whose StopHateSpeech campaign inspired this project. We would also like to thank Citizen Science Zurich and the different groups of volunteers who contributed their time to this project.

Funding statement. We gratefully acknowledge funding by InnoSuisse (Grant 46165.1 IP-SBM) and the Swiss Federal Office of Communications.

Competing interests. The author declares that they have no competing interests.

References

Akhtar, S., Basile, V. and Patti, V. (2020). 'Modeling annotator perspective and polarized opinions to improve hate speech detection', in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Volume 8, Washington, DC (pp. 151–154). The AAAI Press.

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J.D., Korobeynikova, M. and Gilardi, F. (2025). 'Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning'. *Journal of Computational Social Science*, **8**(1), 17.

Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R.J. and Teo, J.T. (2023). 'Ai chatbots not yet ready for clinical use'. *Frontiers in Digital Health*, **5**, 60.

Barberá, P., Boydstun, A.E., Linn, S., McMahon, R. and Nagler, J. (2021). 'Automated text classification of news articles: a practical guide'. *Political Analysis*, **29**(1), 19–42.

Barrie, C., Palmer, A. and Spirling, A. (2024). *Replication for Language Models: Problems, Principles, and Best Practice for Political Science*. Preprint. Princeton University.

Benoit, K., Conway, D., Lauderdale, B.E., Laver, M. and Mikhaylov, S. (2016). 'Crowd-sourced text analysis: reproducible and agile production of political data'. *American Political Science Review*, **110**(2), 278–295.

Boukes, M., Van de Velde, B., Araujo, T., and Vliegenthart, R. (2020). 'What's the tone? Easy doesn't do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools'. *Communication Methods & Measures*, **14**(2), 83–104.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). 'Language models are few-shot learners'. *Advances in Neural Information Processing Systems*, **33**, 1877–1901.

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., and Chen, L. (2023). 'Generative AI and ChatGPT: applications, challenges, and AI-human collaboration'. *Journal of Information Technology Case and Application Research*, **25**(3), 277–304.

Gagliardone, I., Gal, D., Alves, T., and Martínez, G. (2015). *Countering Online Hate Speech. UNESCO Series on Internet Freedom*. UNESCO.

Gilardi, F., Alizadeh, M. and Kubli, M. (2023). 'ChatGPT outperforms crowd workers for text-annotation tasks'. *Proceedings of the National Academy of Sciences*, **120**(30), e2305016120.

Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E.-P., Bing, L., Xu, X., Poria, S. and Lee, R. (2023). 'LLM-adapters: an adapter family for parameter-efficient fine-tuning of large language models', in H. Bouamor, J. Pino and K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5254–5276). Association for Computational Linguistics.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). 'Chatgpt for good? on opportunities and challenges of large language models for education'. *Learning and individual differences*, **103**, 102274.

Koco, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K. et al. (2023). 'ChatGPT: jack of all trades, master of none'. *Information Fusion*, **99**, 101861.

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press

Kotarcic, A., Hangartner, D., Gilardi, F., Kurer, S. and Donnay, K. (2022). Human-in-the-loop hate speech classification in a multilingual context. in Y. Goldberg, Z. Kozareva and Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 7414–7442). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozeti, I. and Zollo, F. (2022). 'Handling Disagreement in Hate Speech Modelling', in *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 19th International Conference, IPMU 2022, Milan, Italy, July 11–15, 2022, Proceedings, Part II* (pp. 681–695). Springer.

Langton, R. (1993). 'Speech acts and unspeakable acts'. *Philosophy & Public Affairs*, **22**(4), 293–330.

Laurer, M., van Atteveldt, W., Casas, A. and Welbers, K. (2024). Less annotating, more classifying: addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, **32**(1), 84–100.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, **55**(9), 1–35.

Ma, H., Zhang, C., Fu, H., Zhao, P. and Wu, B. (2023). Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning.

Maitra, I. and McGowan, M.K. (2012). 'Introduction and overview', in I. Maitra and M. K. McGowan (Eds.), *Speech and Harm: Controversies Over Free Speech* (pp. 1–23). Oxford: Oxford University Press.

Ollion, É., Shen, R., Macanovic, A. and Chatelain, A. (2024). 'The dangers of using proprietary LLMs for research'. *Nature Machine Intelligence*, **6**(1), 4–5.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J. and Lowe, R. (2022). 'Training language models to follow instructions with human feedback', in S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Volume **35**, pp. 27730–27744). Curran Associates, Inc.

Parker, S. and Ruths, D. (2023). 'Is hate speech detection the solution the world wants?' *Proceedings of the National Academy of Sciences*, **120**(10), e2209384120.

Ray, P.P. (2023). ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, **3**, 121–154.

Shaikh, O., Zhang, H., Held, W., Bernstein, M. and Yang, D. (2022). On second thought, let's not think step by step! Bias and toxicity in zero-shot reasoning. arXiv preprint *arXiv:2212.08061*.

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature* **616**(7957), 413–413.

United Nations (2020). *United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences*. United Nations.

Van Atteveldt, W., Van der Velden, M.A., and Boukes, M. (2021). 'The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms'. *Communication Methods and Measures*, **15**(2), 121–140.

Waldron, J. (2014). *The Harm in Hate Speech*. Harvard University Press.

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., and Le, Q.V. (2022). Finetuned Language Models Are Zero-Shot Learners. Preprint arXiv: 2109.01652.

Yang, R., Tan, T.F., Lu, W., Thirunavukarasu, A.J., Ting, D.S.W. and Liu, N. (2023). 'Large language models in health care: development, applications, and challenges'. *Health Care Science*, **2**(4), 255–263.

Cite this article: Umansky N, Kubli M, Kotarcic A, Bronner L, Kurer S, Grech P, Hangartner D, Gilardi F, and Donnay K (2026). Improving hate speech detection with large language models. *European Journal of Political Research*, 1–12. https://doi.org/10.1017/S1475676525100546

https://doi.org/10.1017/S1475676525100546 Published online by Cambridge University Press