

LLMs and Finetuning: Benchmarking Cross-Domain Performance for Hate Speech Detection

Ahmad Nasir[1], Aadish Sharma[1], Kokil Jaidka[2(✉)], and Saifuddin Ahmed[3]

[1] Indian Institute of Technology Delhi, Delhi, India
[2] Communications and New Media, National University of Singapore, Singapore, Singapore
[3] Wee Kim Wee School of Communcation and Information, Nanyang Technological University, Singapore, Singapore

Abstract. In the evolving landscape of online communication, hate speech detection remains a formidable challenge, further compounded by the diversity of digital platforms. This study investigates the effectiveness and adaptability of pre-trained and fine-tuned Large Language Models (LLMs) in identifying hate speech, to address two central questions: (1) To what extent does the model performance depend on the fine-tuning and training parameters?, (2) To what extent do models generalize to cross-domain hate speech detection? and (3) What are the specific features of the datasets or models that influence the generalization potential? The experiment shows that LLMs offer a huge advantage over the state-of-the-art even without pretraining. Ordinary least squares analyses suggest that the advantage of training with fine-grained hate speech labels is washed away with the increase in dataset size. While our research demonstrates the potential of large language models (LLMs) for hate speech detection, several limitations remain, particularly regarding the validity and the reproducibility of the results. We conclude with an exhaustive discussion of the challenges we faced in our experimentation and offer recommended best practices for future scholars designing benchmarking experiments of this kind.

Keywords: Hate speech detection · Large Language Models (LLMs) · fine-tuning · pretraining · cross-domain generalization

1 Introduction

The widespread use of the internet, especially social media, has amplified online hate speech, with varying regulations across regions like the US and the EU [14]. While platforms like Facebook employ automated systems to remove potential hate content, certain severe forms of hate speech, such as threats or racism, require legal action. As these platforms become increasingly prevalent, effectively identifying and mitigating such harmful content becomes paramount. Given the high costs and psychological toll of human moderation, there is a pressing need for automated solutions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15650, pp. 17–34, 2025.
https://doi.org/10.1007/978-3-031-96962-1_2

However, hate speech detection is a challenging problem where a single model may not generalize to different scenarios. For instance, differentiating between hateful, spam, abusive, and profane content requires a deep understanding of the underlying intent and the context in which the words are used [18,30,39]. Large language models (LLMs) may offer a promising solution for high-precision hate speech detection, as they are often trained on extensive, multilingual and multi-source text data for various text classification tasks [4,29,31]. However, in this case, benchmarking black-box large language models (LLMs) remains a significant challenge. This is because the opacity surrounding how LLMs are trained – specifically, what their training data comprised, and what specific tasks they were trained on – poses a considerable barrier to their effective utilization. Without this knowledge, determining the most appropriate LLM for a task such as hate speech detection becomes an increasingly difficult guessing game, in part due to the proliferation of LLMs that successively outperform each other on a standard set of tasks [4]. The lack of transparency also raises concerns about their generalizability to real-world scenarios.

From an application standpoint, a second research gap emerges regarding whether fine-tuning would improve the model performance for a popular task such as hate speech detection, as there is no prior evidence to suggest that it consistently enhances performance across various domains. A third research gap lies in whether models, if finetuned and/or pre-trained, would generalize easily to new domains and what factors are predictive of generalizability. Prior research has referred to the importance of dataset size, label similarity [15,28], coverage [22], and the number of authors [2], yet they do not offer an exhaustive analysis of the effects of dataset characteristics on model generalizability. Furthermore, although cross-domain classification studies of hate speech detection have been conducted in the past [10], similar benchmarks are unavailable in the context of LLMs.

To address these research gaps, in this study, we evaluate and benchmark various standard and fine-tuned large language models (LLMs) for detecting hate speech in different domains and contexts. Our analytical framework serves as a benchmarking tool, setting a standard for future research in hate speech detection using LLMs, ensuring consistent and comparable results. Our work addresses the following research questions:

1. To what extent does the model performance depend on the fine-tuning and training parameters?
2. To what extent do models generalize to cross-domain hate speech detection?
3. What are the specific features of the datasets or models that influence the generalization potential?

2 Background

Cross-domain classification is the generalization of a classification model trained on a particular type of dataset and used to classify datasets that the model has

Table 1. The datasets used in our experiments. 1 = "Hateful" and 0 = "Non-Hateful"

Dataset	Instances	Instances after preprocessing	Source	Positive (1)	Negative (0)
Gab	33,776	32,010	Gab	14,614	17,396
Reddit	22,334	16,982	Reddit	5,255	11,727
OLID	14,200	9,914	Twitter	1,074	8,840
SOLID$_{E}xtended$	3340	3,337	Twitter	347	2,990
SOLID$_{S}emEval$	2997	2,995	Twitter	188	2,807
HASOC (English)	5,983	4,734	Twitter and Facebook	1,143	3,591
ICWSM	5,143	3,221	YouTube and Facebook	2,364	857
Restricted	100,000	58,812	Twitter	4,963	53,849
Toraman (English)	100,000	89,236	Twitter	1,617	87,619

not seen while training. There has recently been a spurt in studies that offer a cross-domain setting for hate speech detection [1,3,6,9,13,15,26,28,33,34,36,37].

The studies by [33] and [35] collapsed the data into hate speech and not hate speech and tested their models on out-of-domain data, reporting poor generalizability. [3] applied a different approach to generalize models to new datasets. Given unlabeled data from multiple sources, they used a semantic distance measure to distinguish hate speech from others. The authors found that this approach outperformed models that were trained on in-domain labeled data and could generalize well. The study by [36] was able to generalize to new data by enriching their training data with a third dataset and then using multi-task learning. [15] used 9 different datasets - W&H [35], Waseem [33], TRAC [17], Kol [16], Gao [12], Kaggle [5], Wul1, Wul2 and Wul3 [38] in training their hate speech models. First, they categorized all datasets into positive (abusive) and negative (non-abusive), then trained each of the datasets using support vector machines and tested on 8 other datasets, they observed 50% reduction in F1 score on out-of-domain data. Next, they used transfer learning to augment the original domain with the information from a different domain. Their approach resulted in improvement in 6 out of 9 cross-domain datasets, and the authors concluded that achieving strong performance in classifying the target dataset requires including at least some training data from that specific dataset.

However, the findings from [28] and [10] offer a different perspective. Unlike [36] and [15], they showed generalization can be achieved without augmenting out-of-domain data. [28] merged datasets into binary categories and trained state-of-the-art BERT-based models. They reported that in-domain results were significantly better than other state-of-the-art models just by fine-tuning. They also extrapolated that the model would perform better on out-of-domain data if the datasets were of a similar type. Furthermore, their findings suggest that datasets with larger percentages of positive samples tend to generalize better than datasets with fewer positive samples, in particular when tested against dissimilar datasets. [10] used 9 publicly available datasets and collapsed them into binary categories and were able to train BERT, ALBERT, SVM, and

fastText. They observed that in most of the cases transformer models generalize better. For generalization, in-domain performance is also important.

In another study, [2] bring up an additional characteristic that should be taken into account in the context of cross-dataset hate speech classification, namely the number of authors of the material captured in a dataset; however, the authors report an F1 score of 0.54, suggesting that dataset size and author complexity may increase the difficulty of the task.

Recent work has explored similar concerns. For example, [1] provide a large-scale empirical evaluation across multiple datasets and show that cross-dataset performance is often overestimated due to insufficient methodological rigor. Likewise, [37] introduce a novel framework to audit model disagreement with humans on what counts as offensive, surfacing errors in both labeling and model behavior. Complementing these, [9] offer strategies for generalizing offensive language identification models across domains using low-resource fine-tuning. Other studies have introduced novel hate speech datasets drawn from specific sociocultural contexts, such as online chess communities [27], or redefined what counts as "extreme speech" in consultation with affected communities [20].

In this work, we address some of the pressing gaps that remain in prior work, such as the need to relate the advances in natural language processing research to the body of prior work on cross-domain studies of hate speech detection. Previous studies have achieved the best performance using different approaches, ranging from SVMs to deep learning models (e.g., [13]). However, in recent years, transformer models like BERT have outperformed most others. Transfer learning has also been shown to improve BERT performance [21]. Our work benchmarks the performance of LLMs on hate speech detection tasks for many standard datasets against these prior state-of-the-art results, while also setting an agenda for future research.

3 Approach

We aim to understand the robustness and adaptability of our models in identifying hate speech across varied contexts and sources. Therefore, we have reported the efficacy of LLMs in detecting hate speech as in-domain performance, as well as across different domains (cross-domain), both pre- and post-fine-tuning.

3.1 Model Setup

We have illustrated our model pipeline in Fig. 1 and described it in the following paragraphs.

As our goal is to train large language models to classify an input text into two classes, i.e. Neutral or Hateful, we attempted a few approaches in building the classification pipeline. First, we used the LLM to generate textual output and used techniques like keyword extraction to classify it into two classes. Second, we used the LLM to generate textual output, get the BERT embeddings of the generated text, and build a classifier using the BERT embeddings to classify it

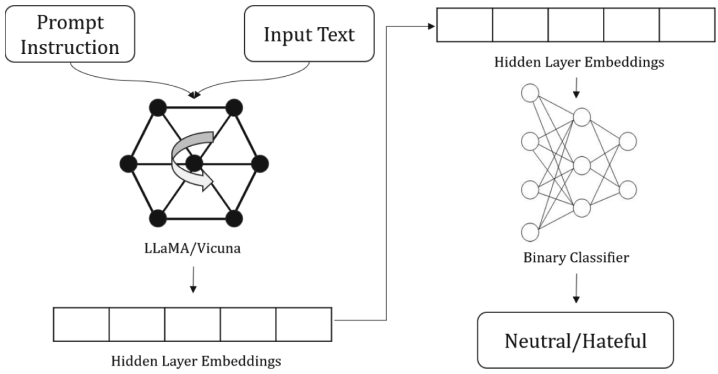


Fig. 1. Model Pipeline

into the two classes. Finally, we used the LLM to get the last hidden layer's embeddings and build a classifier using the hidden embeddings to classify it into the two classes.

We had very limited success with the first two techniques, perhaps because we were trying to fit the model's textual output into the two classes, which contains many extra output words apart from the correct classification. Thus, we chose the third technique to build our framework. The framework comprises a classification pipeline using a Backbone LLM to get text embeddings. The LLM may or may not be fine-tuned on a sub-sample of the hate speech dataset, depending on the designated parameters of the pipeline. Subsequently, we use the embeddings to classify it as Neutral or Hateful with the help of a Binary Classifier. First, we provide an input text, converted into the following instruction format <*Instruction, Input, Output*>. Next, the instruction format input is fed to a Backbone LLM model, and the LLM's last hidden layer is extracted. Finally, the last hidden layer is fed to a Binary Classification model, which outputs the predicted class.

Backbone Model. The LLM Backbone can be a Large Language Model, such as LLaMA [31], Alpaca [29], Vicuna [4] etc. In our paper, we have used the following 4 Backbone Models:

- LLaMA-7B[1]: LLaMA is open source foundation model released by Meta AI. It has been trained on publicly accessible data. It is based on the transformer architecture proposed by [32] and comprises of 7 billion parameters.
- LLaMA-7B-finetuned: We instruction fine-tune the base LLaMA-7B model on a sub-sample of our hate-speech data using QLoRA [7] and PEFT [19] which is a cost-effective technique for low resource training.
- Vicuna-7B:[2] Vicuna-7B is obtained by fine-tuning LLaMA-7B base model using approximately 70K user-shared conversations gathered from ShareGPT.com

[1] https://huggingface.co/decapoda-research/llama-7b-hf.
[2] https://huggingface.co/lmsys/vicuna-7b-v1.5.

– Vicuna-7B-finetuned: Similar to the fine-tuning of base LLaMA-7B, we have used instruction-based fine-tuning to fine-tune the Vicuna-7B base model on a sub-sample of our hate-speech dataset.

The reason for our choice of the above two base models, viz. LLaMA [31] and Vicuna [4] is that while LLaMA is a foundation model, which has not been fine-tuned on any specific task, Vicuna is built on top of Alpaca [29], and Alpaca has been fine-tuned to follow instructions. Thus, we wanted to compare the performance of foundation and instruction-tuned models.

Binary Classifier. The Binary Classifier is trained on the last hidden layer embeddings of the Backbone LLMs. We convert all the training samples to the *Alpaca Instruction Format*, and pass them through the Backbone LLMs, and dump the last hidden layer weights. Then, we use these 4096-dimensional embeddings as input to the classifier and 0/1 as labels for Neutral/Hateful cases, respectively, to train the classifier.

3.2 LLM Fine-Tuning and Classifier Training

LLM Fine-Tuning. We fine-tuned both of the base models, i.e. LLaMA and Vicuna on a sub-sample of our hate-speech dataset, using QLoRA [7], an extremely cost-effective and affordable training technique. By fine-tuning the base models on our hate-speech dataset, we hope to help the model learn various nuances of the actual examples of hateful texts.

We used 1000 examples from each of the 9 datasets and converted them in the Alpaca Instruction Format *<Instruction, Input, Output>* to do the instruction fine-tuning. We used *"Detect if the text is hateful or neutral"* as the instruction, the content of the input text as the input, and *HATEFUL or NEUTRAL* as the output. Making use of fine-tuning scripts[3,4], we ran the fine-tuning for 1 epoch with batch size 128. This provided us with two more Backbone LLMs, i.e. *LLaMA-7B-finetuned*, and *Vicuna-7B-finetuned*.

Classifier Training. For a given dataset, the Binary Classifier is trained on the last hidden layer embeddings of the Backbone LLMs. We take 2000 training examples from the given dataset, and convert them to *Alpaca Instruction Format*, and pass them through one of the four Backbone LLMs, and dump the last hidden layer weights. Then, we use these dumped embeddings as input to the classifier and 0/1 as labels for Neutral/Hateful cases, respectively, to train the classifier. Thus, for each dataset, we end up with 4 Binary Classifiers, one for each Backbone LLM. The classifier is trained for 50 epochs, with early stopping, and in most cases, training ends at around 15–20 epochs.

As a result of Binary Classification training, we have 4 classifiers for each of the 9 datasets, and we use these classifiers to evaluate the self-domain and cross-domain performance on the test sets (end-domains).

[3] https://github.com/tloen/alpaca-lora.
[4] https://github.com/artidoro/qlora.

Also, when we are training the classifiers on the fine-tuned models, we observe that using a model fine-tuned on an enriched dataset (9000 samples, 1000 from each dataset) performs significantly better than non-enriched dataset (only 1000 samples from the respective dataset), thus we decide to use the models fine-tuned on the enriched datasets to get the weight dumps to train the classifiers.

4 Experimental Setup

In the *In-Domain Performance* section, we evaluated various Backbone LLMs on their performance on hate speech detection across nine datasets reported in Table 1. Therefore, each LLM is paired with 9 classifiers tailored to a specific dataset. These classifiers are trained on 2,000 samples from their respective datasets and subsequently tested on the entirety of the same dataset. The results comparing the best model performance on the F1 scores are reported in Table 2.

In the *Cross-Domain Performance* evaluation, we explored the adaptability of these models. Each backbone LLM is associated with 8 out-of-domain classifiers from different domains for every end-domain dataset. Though trained on 2,000 samples from their datasets, these classifiers are evaluated on all samples from a different end-domain dataset to obtain cross-domain F1 scores. The results are detailed in Table 3.

Finally, for the evaluation of factors predicting *Model Generalizability*, we consider the predictive performance as a multivariate optimization problem dependent on the source similarity, source of origin, label distribution, dataset size, base model, and other considerations.

Therefore, our experiments denote a $5(source) \times 2(basemodel) \times 2(finetuning)$ ablation design, with additional covariates, such as training dataset size and in-domain performance. We then conducted a multiple linear regression over these parameters to identify which of these factors played a significant role in explaining model performance in a cross-domain setting.

4.1 Datasets

The Hate Speech detection task was formulated as a binary classification problem of whether a text is hateful or non-hateful, on nine publicly available datasets, with label distributions reported in Table 1.

- Reddit and Gab [23]: These datasets comprise of 22,334 and 33,776 points respectively and the texts are categorised as hateful and non-hateful. Reddit dataset is more imbalanced than Gab. For testing and training we have kept the same ratios as present in the data.
- HASOC [18]: This dataset has 5 different columns HateSpeech, Offensive, Profanity, Non-Hate-Offensive and Hate-Offensive. Our study has used HateSpeech and Non-Hate-Offensive columns from the English dataset.
- Toraman [30]: This dataset has 100,000 tweets; 20k tweets are taken from each different domain - religion, sports, gender, race and politics, with columns as hate, offensive and normal.

- ICWSM [25]: This is comments data on videos posted by social media news on YouTube and Facebook. The comments are divided into hateful and normal and further divided into 13 main categories and 16 subcategories.
- OLID [39], SOLID SemEval and SOLID Extended [24]: The datasets are hierarchically divided into three tasks - Offensive Language Detection, Categorization of Offensive Language, and Offensive Language Target Identification. We have used the first task to get the neutral examples and the third task to get the hateful examples.
- Restricted [11]: Comprising of 100,000 annotated data points categorized into four distinct classes: normal, hateful, spam, and abusive. We have segregated normal and hateful classes as per the need of our study.

5 Results and Analysis

The following paragraphs summarize our key results, while the detailed model performance is reported in the online supplement.[5]

5.1 In-Domain Performance

In answer to RQ1, we observe that using LLMs, even without fine-tuning, significantly improves the self-domain as well as the cross-domain performance on the datasets, as compared to the previously available best models, as reported in Table 2

Table 2 provides the best results per dataset per model variant. For Gab, the F1-score of the best model (CNN) was 0.896, while that of our best model is 0.9942, which is an almost 10% increase (0.0982). For Reddit, the F1-score of the best model (RNN) was 0.77, which increased to 0.9601, resulting in a substantive 19% (0.1901) increase. For the Toraman, the F1-score of the best model (Megatron) was 0.830, which increased to 0.9851 for our best model, resulting in 15% (0.1551) increase in the F1-score. For ICWSM, the F1-score of the best model (SVM) was 0.96, which increased to 0.9978 for our best model.

For datasets like OLID, SOLID_SemEval, SOLID_Extended and HASOC, there isn't a direct comparison, as their dataset is in a hierarchical format while ours is in a binary classification format, but even for the first task, our best model's performance significantly outperforms the baselines, as illustrated in the Self-Domain Performance.

As per Table 2, in about half of the cases the base model for LLaMA outperformed the fine-tuned model. Note the very high scores on the three Solid datasets, which leads us to suspect that they are pre-included in the models' pretraining corpora.

By instruction fine-tuning the base LLaMA-7B model, we observe that the average F1-score of LLaMA-7B-finetuned model slightly decreases (0.00304) as compared to the LLaMA-7B model while at the same time, by instruction

[5] https://github.com/kj2013/hate-finetune.

fine-tuning Vicuna-7B, the average F1-score of Vicuna-7B-finetuned increases (0.001319) as compared to the Vicuna-7B model. There is a significant improvement in the F1-score of Toraman, Reddit and OLID.

Table 2. The best-performing LLaMA and Vicuna variants for in-domain hate speech detection.

Dataset	LLaMA Variants		Vicuna Variants	
	Best-Performing Variant (F1)	F1	Best-Performing Variant (F1)	F1
Gab	Training	0.994	Training	0.994
HASOC	Finetuning + Training	0.997	Finetuning + Training	0.997
ICWSM	Finetuning + Training	0.998	Training	0.998
Reddit	Finetuning + Training	0.959	Finetuning + Training	0.960
Restricted	Training	0.994	Training	0.998
Toraman	Training	0.978	Finetuning + Training	0.985
Solid Extended	Training	1	Training	1
OLID	Training	0.998	Finetuning + Training	0.999
Solid SemEval	Training	1	Finetuning + Training	0.997

Table 3. The best-performing models for cross-domain hate speech detection.

Test Dataset	Best Performing Model	Training Dataset	F1
Gab	Vicuna	Reddit	0.992
Reddit	LLaMA	Gab	0.952
HASOC	Vicuna	Gab	0.997
ICWSM	LLaMA	HASOC	0.995
Restricted	LLaMA	Gab	0.998
Toraman	Vicuna	Gab, OLID	0.993
OLID	LLaMA	Gab, Reddit	0.999
Solid Extended	LLaMA	Gab, HASOC	1
Solid SemEval	LLaMA	Gab, HASOC	1

5.2 Cross-Domain Performance

In answer to RQ2, Table 3 and Fig. 2 reports the best-performing fine-tuned models for cross-domain hate speech detection, with the second column denoting the best or fine-tuned models for each validation dataset in the third column. In general, the Table suggests that models fine-tuned on the Gab dataset had the best cross-domain generalizability, except in the case of ICWSM (sourced from Twitter), where a model fine-tuned on HASOC, also sourced from Twitter, performed the best. Once again, the OLID, SOLID-Extended,and SOLID SemEval

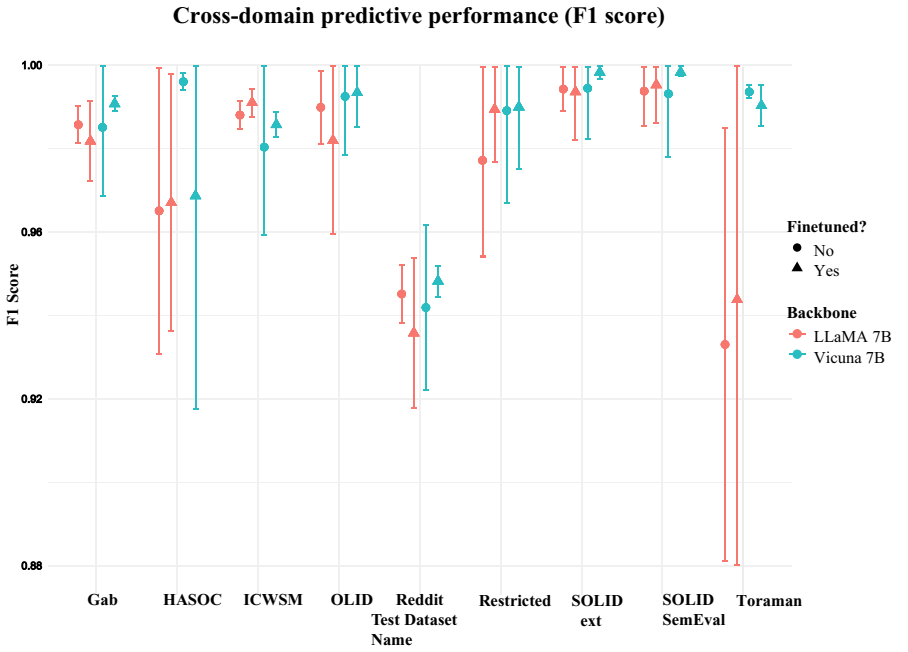


Fig. 2. Cross-domain predictive performance for hate speech detection

datasets proved to be less useful benchmarks as suggested by the extremely high model accuracies. We expect that this occurred because they may have been used for pretraining the LLMs we chose, and have little to do with the dataset we fine-tuned on.

The findings suggest that instruction fine-tuning Vicuna-7B model has resulted in the increase of the cross-domain performance for 7 out of the 9 datasets (Gab, ICWSM, OLID, Reddit, Restricted, SOLID_Extended and SOLID_SemEval) as compared to the base Vicuna-7B model. Similarly, instruction fine-tuning LLaMA-7B has resulted in the increase of the cross-domain performance for 5 out of 9 datasets (HASOC, ICWSM, Restricted, SOLID_SemEval and Toraman) as compared to the base LLaMA model. Instances of false negatives in the fine-tuned Vicuna backbone trained on Gab (typically the best-performing cross-domain setup in our findings) and tested on Reddit, and vice-versa, are reported in the Appendix in Table 4. Many of the instances of the Gab-trained model tested on Reddit fail on the subtle instances of hate speech, while the Reddit-trained model appear to fail on both subtle and explicitly hateful instances.

5.3 Error Analysis

Table 4 reports some of the examples where some of the best-performing cross-domain finetuned Vicuna models (as per Table 3) failed to correctly classify the input (false negatives).

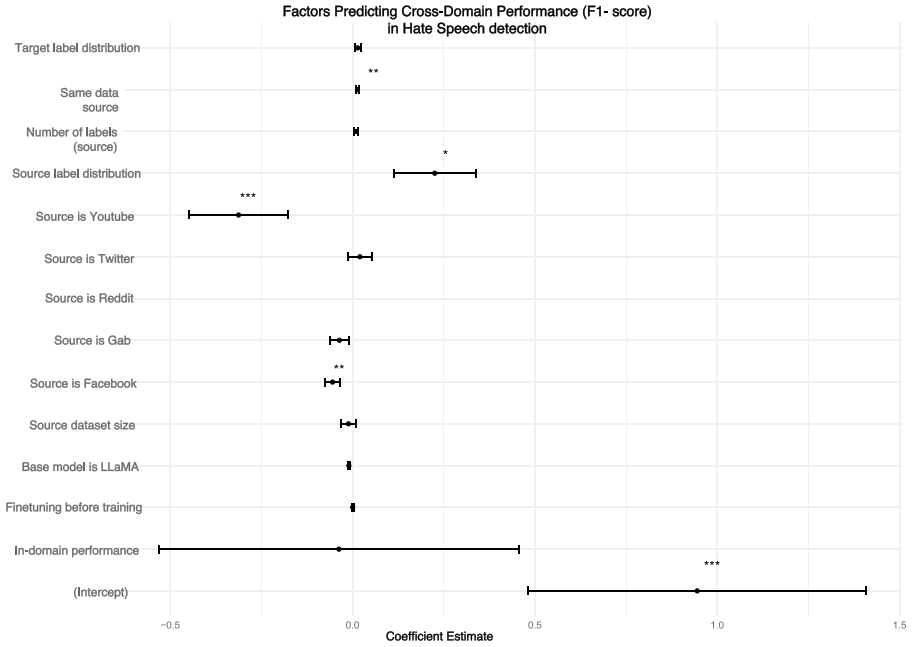


Fig. 3. Model generalizability measured through regressing all model performance on training and testing parameters

In general, we observed that false negatives occur when the model struggles to detect implicit toxicity or subtle offensive language. This is especially evident in cases where the hate speech is masked by sarcasm or indirect phrasing (e.g., "A chinaman and a commie teaming up?"). In particular, long and convoluted sentences that bury toxic elements are often missed (e.g., "The flag is helpless to stop you..."). Similarly, the classifier tends to overlook offensive slurs that are casually integrated into discourse (e.g., "You argue like a child, although..."). Moreover, slurs and offensive terms that are normalized in certain online communities seem to evade detection more frequently (e.g., "Get on the Front line C***!!"). On platforms like Gab or Reddit, where language norms differ, the model often fails to adapt (e.g., "Rick, the couple people I know..."). For instance, terms used casually in one platform might not trigger the model's filters on another (e.g., "C*** hair away from achieving the same..."). The model also struggles with offensive terms that appear at the end of otherwise neutral statements (e.g., "It only shows you're a..."). While certain phrases may seem contextually benign, their deeper connotations may go unnoticed (e.g., "Kanye saw through Taylor Tw** Swift..."). Lastly, the classifier's reliance on explicit markers of toxicity may limit its ability to capture more nuanced expressions of hate.

5.4 Factors Determining Model Generalizability

The ordinary least squares analyses reported in Fig. 3 suggest that to generalize well, training dataset size and in-domain performance are secondary to label imbalance. Simple, small-sample, pre-training of Vicuna-based language models

Table 4. Examples of false negatives in cross-domain hate speech detection.

Classifier with a fine-tuned Vicuna backbone LLM, trained on Reddit, tested on Gab
– A chinaman and a commie teaming up? The great proletarian cultural revolution is upon us!
– I think Kanye saw through Taylor Twat Swift before any of us saw who she really is. I question no more about why Kanye started cracking on her by taking the microphone away.
– Politicians misspeak all the time Yes I'm sure Kasich, a RINO and Trump hater, conspired to put McCain to death and knew about it advance. When McCain was a RINO and loved by other RINOs. ????? You are dumber than a black African.
– Twat, he loves them because they finance him. Book deals are a dead give away.
– Give it to a soldier who has defend it, then try to burn it. Show me what a badass you really are. The flag is helpless to stop you or fight back. Completely unnecessary act, its proves nothing and accomplishes nothing. It only shows your a ignorant weak faggot. Like kicking a puppy, it shows how weak and worthless you are
– Trump may not be our guy but he proved in the primaries that it's possible to mount a successful primary challenge against multiple deep state insiders and media favorites. Ron Paul was a cunt hair away from achieving the same, himself.
– As I said previously, you argue like a child, although with additional feedback I'd argue a retarded one at best.
– Get on the Front line Cunt !! Quit mouthin off & Get out there !!
– Astronaughts should stick to what they know best... getting out there in space...perhaps the silly cunt should take his helmet off next time he's having a spacewalk. Fuck me these lefty scum give me the literal shits...and some of these useless cunts think that having fucked the good ole USA up the ass with their liberal crapola they can move to New Zealand
Classifier with a fine-tuned Vicuna backbone LLM, trained on Gab, tested on Reddit
– I'll agree on Rick, the couple people I know who say he's one of their favorite characters are insufferable twats.
– Look a historical revisionist, how cute. Go back to your gender studies echo chamber and plot on how the world would be perfect if men simply didn't exist.
– Oh wow we got a white knight here that goes on PPD and starts talking shit about commenters. Why don't you fly to England, put on your white knight armor, and defend all the fat worthless alimony money stealing fatties.
– An American will start yelling or disparaging. Verbal sparring is fun and really interesting to watch... but a bumbling piece of white trash is neither.

rather than fine-tuning significantly impact model performance. The effects of having fine-grained labels matter primarily when training datasets are small in size; however, these effects are washed away for larger training datasets. We see that fine-tuning LLMs matters little, and researchers may simply pre-train their chosen language models on a small labeled sample, preferably from the same source, in order to get the best possible results. Fine-tuning or training on Facebook or YouTube data appears to adversely impact model performance, more so in the case of YouTube than Facebook.

6 Discussion

Our findings suggest that cross-domain fine-tuning, even on a single epoch, is nearly always beneficial for greater precision on a known target dataset. Although newer models have emerged, these findings remain highly relevant, as the principles of enhancing model precision through targeted fine-tuning and leveraging label diversity apply to current and future models. Therefore, our results raise questions about the need to fine-tune large language models (LLMs) for text classification tasks, such as hate speech detection. While models like Vicuna-7B, when fine-tuned, consistently outperform their base versions across various datasets, this is not the case with LLaMA-7B.

Furthermore, our findings provide a complementary perspective to the insights from [28] and [10]. While their findings suggested that training data from a similar platform or a model with high in-domain performance was a critical factor for achieving effective generalization in hate speech detection, our analysis suggests that label distribution offers a key indicator of effective fine-tuning for LLMs.

6.1 Limitations

While our research demonstrates the potential of large language models (LLMs) for hate speech detection, several limitations remain, particularly regarding the validity and the reproducibility of the results. We offer these in the interest of transparency, as points of discussion, and as recommended best practices for future scholars designing benchmarking experiments of this kind.

First, hate speech detection is highly context-dependent, and the same model may not generalize well across different sociocultural settings or types of hate speech. Differentiating between hateful, abusive, spam, or profane content requires a deep understanding of the language and intent behind it, which current LLMs may not fully capture [30].

Second, we acknowledge that aspects of our experimental design and implementation setup introduced severe limitations, several of which were raised by reviewers. Our evaluation framework—where each LLM is paired with nine classifiers, each trained on 2,000 examples from a single dataset and tested on the same dataset—may inadvertently reinforce dataset-specific artifacts. While this approach standardizes training size across datasets, it also risks overfitting to

dataset idiosyncrasies rather than surfacing generalization capabilities. Reviewers were concerned about the use of overlapping training and test sets, which risked test-train contamination, and the implausibly high F1 scores reported across all datasets—including subjective hate speech datasets known to suffer from annotation noise.

Thirdly, reviewers were concerned that the results looked "too good to be true" in some cases for a task like offensive language detection. Many approaches achieved almost perfect F1 scores. The opacity of LLMs regarding their training data pose a significant challenge for us to diagnose whether this was an issue in our benchmarking experiments or in selecting models for specific tasks like hate speech detection [4]. Without transparency regarding the datasets and tasks used for pre-training LLMs, it is challenging to determine a model's suitability for benchmarking experiments as well as real-world applications.

Lastly, while we observed cross-domain fine-tuning consistently improving performance, the lack of labeled data for certain hate speech subtypes and label imbalances across datasets may limit the generalizability of our findings. This is particularly concerning for underrepresented or emergent forms of hate speech, which may not be adequately captured by models trained on canonical datasets like SOLID or Semeval.

Despite these valid concerns, we would like to defend our design, which was chosen based on several principles. Our goal was to simulate low-resource, high-variance environments typical of real-world hate speech detection settings, where only a small amount of labeled data is available per domain. Fixing the training size to 2,000 samples allowed us to control for dataset size effects while testing on the entire dataset helped us capture model behavior on rare or ambiguous cases. We prioritized consistency and comparability across all nine datasets, striking a balance between practical constraints and analytical scope. In practice, downstream applications often involve inference on previously unseen data distributions, and such per-dataset classifiers may underperform in such settings.

Nonetheless, a major concern we would like to highlight is that our replication attempts were hindered by model versioning and package incompatibilities, which prevented us from reproducing our own results within a short span of three months. At the time of our experiments, the model and tokenizer versions we used were stable and well-supported; it was only with subsequent updates to LLaMA, Vicuna, and core HuggingFace libraries that incompatibilities arose, which we regret not anticipating or isolating earlier.

7 Conclusion

Our contribution is a research design to benchmark cross-domain performance of hate speech detection models. We demonstrate substantive improvements over the state-of-the-art hate speech detection models, often with a single epoch of fine-tuning. The first key takeaway of our work is that generalization varies from model to model and that, in general, label imbalance affects model performance in cross-dataset applications. The second takeaway is that for out-of-domain

prediction on a given end-domain dataset, there are specific advantages for fine-tuned models trained on small datasets with fine-grained hate speech labels.

In light of the limitations and community feedback, we propose the following recommendations for future work:

- Use clearly separated train, validation, and test sets with no content overlap; document all sampling and split procedures.
- Avoid evaluating on full datasets after training on partial subsets; instead, use dedicated held-out sets or k-fold cross-validation.
- Track and version all dependencies, including model weights, tokenizers, and environments, using tools such as Docker or conda-lock.
- Include comparisons with standard baselines (e.g., CNNs, RNNs, or instruction-tuned LLMs) for interpretability and external validity.
- Report the full range of performance metrics across all tested configurations, not only the top-performing model.
- Use the term cross-dataset unless there is demonstrable variation in domain characteristics such as platform, language variety, or sociocultural context.
- Acknowledge and analyze the definitional and annotation inconsistencies across hate speech datasets, particularly when interpreting cross-dataset generalization results.

These best practices are necessary for improving the interpretability, reproducibility, and credibility of LLM-based hate speech detection studies.

Acknowledgments. This research was supported by the Ministry of Education, Singapore, through its MOE AcRF Tier 3 Grant (MOE- MOET32022-0001) and MOE Tier 1 Grant (WBS A-8000231-01-00). We also extend our gratitude to the anonymous reviewers of the Conference on Language Modeling (COLM) 2024 and the International Conference on Computational Linguistics (COLING) 2025 for their invaluable feedback.

Ethical Considerations. We caution against relying exclusively on AI-generated predictions of hate speech for high-stakes or real-world applications. Hate speech is inherently subjective and highly context-dependent; even accurate models at the aggregate level may fail to account for cultural nuance, sarcasm, or reclaimed language at the individual level. These risks are amplified when models trained in one sociocultural context are applied to another, where norms and linguistic cues differ [8]. Furthermore, the opacity of large language models and the lack of transparency in their pretraining data make it difficult to assess how sociocultural priors are encoded or reinforced. To mitigate these risks, we emphasize the need for continual in-context evaluation, supported by human oversight and participatory annotation practices that reflect the values and expectations of affected communities. Transparency, context-awareness, and sustained monitoring are essential for the responsible deployment of LLMs in hate speech detection.

A Baseline results

Table 5 reports the state-of-the-art in prior work for each of the nine datasets under study, for the task of hate speech detection.

Table 5. Baseline results of the best model for each dataset

Dataset	Best Performing Model	Metric	Score
Gab	CNN	F1	0.896
Reddit	RNN	F1	0.775
Toraman	Megatron	F1	0.830
ICWSM	SVM	F1	0.960
OLID (Task A)	CNN	F1	0.900
OLID (Task C)	CNN	F1	0.670
SOLID SemEval (Task A)	BERT	Macro-F1	0.923
SOLID SemEval (Task C)	BERT	Macro-F1	0.645
HASOC (Task A)	LSTM	Weighted F1	0.839
HASOC (Task C)	LSTM	Weighted F1	0.820

B Additional results

The comprehensive set of results for the 36 classifiers evaluated for RQ1 and the 288 classifiers evaluated for RQ2 are reported in https://github.com/kj2013/hate-finetune.

References

1. Antypas, D., Camacho-Collados, J.: Robust hate speech detection in social media: a cross-dataset empirical evaluation. In: The 7th Workshop on Online Abuse and Harms (WOAH), pp. 231–242 (July 2023)
2. Arango, A., Pérez, J., Poblete, B.: Hate speech detection is not as easy as you may think: a closer look at model validation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 45–54 (2019)
3. Chandrasekharan, E., Samory, M., Srinivasan, A., Gilbert, E.: The bag of communities: identifying abusive behavior online with preexisting internet data. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3175–3187 (2017)
4. Chiang, W.L., et al.: Vicuna: an open-source chatbot impressing GPT-4 with 90%* chatgpt quality. See https://vicuna.lmsys.org. Accessed 14 April 2023 (2023)
5. Conversation, A.: Toxic comment classification challenge: Identify and classify toxic online comments (2017)

6. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 512–515 (2017)
7. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLORA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314 (2023)
8. Diener, E., Diener, M., Diener, C.: Factors predicting the subjective well-being of nations. In: Culture and well-being, pp. 43–70. Springer (2009)
9. Dmonte, A., Arya, T., Ranasinghe, T., Zampieri, M.: Towards generalized offensive language identification. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (2024)
10. Fortuna, P., Soler-Company, J., Wanner, L.: How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Inform. Process. Manage. **58**(3), 102524 (2021)
11. Founta, A., et al.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
12. Gao, L., Huang, R.: Detecting online hate speech using context aware models. arXiv preprint arXiv:1710.07395 (2017)
13. Gröndahl, T., Pajola, L., Juuti, M., Conti, M., Asokan, N.: All you need is" love" evading hate speech detection. In: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, pp. 2–12 (2018)
14. Herz, M., Molnár, P.: The content and context of hate speech: Rethinking regulation and responses. Cambridge University Press (2012)
15. Karan, M., Šnajder, J.: Cross-domain detection of abusive language online. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 132–137 (2018)
16. Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., Taboada, M.: The SFU opinion and comments corpus: A corpus for the analysis of online news comments. Corpus Pragmatics **4**, 155–190 (2020)
17. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 1–11 (2018)
18. Mandl, T., et al.: Overview of the HASOC track at fire 2019: Hate speech and offensive content identification in Indo-European languages. In: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 14–17 (2019)
19. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S.: PEFT: State-of-the-art parameter-efficient fine-tuning methods (2022). https://github.com/huggingface/peft
20. Maronikolakis, A., Wisiorek, A., Nann, L., Jabbar, H., Udupa, S., Schütze, H.: Listening to affected communities to define extreme speech: dataset and experiments. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 1089–1104 (May 2022)
21. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-based transfer learning approach for hate speech detection in online social media. In: Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8, pp. 928–940. Springer (2020)
22. Pamungkas, E.W., Patti, V.: Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In:

Proceedings of the 57th Annual Meeting Of the Association For Computational Linguistics: Student Research Workshop, pp. 363–370 (2019)
23. Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251 (2019)
24. Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., Nakov, P.: Solid: A large-scale semi-supervised dataset for offensive language identification. arXiv preprint arXiv:2004.14454 (2020)
25. Salminen, J., et al.: Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
26. Salminen, J., Hopf, M., Chowdhury, S.A., Jung, S.g., Almerekhi, H., Jansen, B.J.: Developing an online hate classifier for multiple social media platforms. Human-centric Comput. Inform. Sci. **10**, 1–34 (2020)
27. Sarkar, R., KhudaBukhsh, A.: Are chess discussions racist? an adversarial hate speech data set (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15881–15882 (May 2021)
28. Swamy, S.D., Jamatia, A., Gambäck, B.: Studying generalisability across abusive language detection datasets. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 940–950 (2019)
29. Taori, R., et al.: Alpaca: A strong, replicable instruction-following model. Stanford Center Res. Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html**3**(6), 7 (2023)
30. Toraman, C., Şahinuç, F., Yilmaz, E.H.: Large-scale hate speech detection with cross-domain transfer. arXiv preprint arXiv:2203.01111 (2022)
31. Touvron, H., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
32. Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Waseem, Z.: Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 138–142 (2016)
34. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: a typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899 (2017)
35. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93 (2016)
36. Waseem, Z., Thorne, J., Bingel, J.: Bridging the gaps: Multi task learning for domain transfer of hate speech detection. Online harassment, pp. 29–55 (2018)
37. Weerasooriya, T., Dutta, S., Ranasinghe, T., Zampieri, M., Homan, C., Khudabukhsh, A.: Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 11648–11668 (2023)
38. Wulczyn, E., Thain, N., Dixon, L.: Ex Machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399 (2017)
39. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666 (2019)