computation

MDPI

Article

Beyond Traditional Classifiers: Evaluating Large Language Models for Robust Hate Speech Detection

Basel Barakat [1,*] and Sardar Jaf [2]

[1] School of Computing, Goldsmiths University of London, London SE14 6NW, UK
[2] School of Engineering and Computer Science, University of Sunderland, Sunderland SR1 3SD, UK; sardar.jaf@sunderland.ac.uk
* Correspondence: b.barakat@gold.ac.uk

Abstract

Hate speech detection remains a significant challenge due to the nuanced and context-dependent nature of hateful language. Traditional classifiers, trained on specialized corpora, often struggle to accurately identify subtle or manipulated hate speech. This paper explores the potential of utilizing large language models (LLMs) to address these limitations. By leveraging their extensive training on diverse texts, LLMs demonstrate a superior ability to understand context, which is crucial for effective hate speech detection. We conduct a comprehensive evaluation of various LLMs on both binary and multi-label hate speech datasets to assess their performance. Our findings aim to clarify the extent to which LLMs can enhance hate speech classification accuracy, particularly in complex and challenging cases.

Keywords: hate speech detection; large language models (LLMs); context understanding; binary hate speech datasets; multi-label hate speech datasets; classification accuracy

check for updates

Academic Editors: Khaled Shaalan and Filippo Palombi

Received: 29 May 2025
Revised: 29 July 2025
Accepted: 5 August 2025
Published: 10 August 2025

Citation: Barakat, B.; Jaf, S. Beyond Traditional Classifiers: Evaluating Large Language Models for Robust Hate Speech Detection. *Computation* **2025**, *13*, 196. https://doi.org/10.3390/computation13080196

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Hate speech (HS) can be defined as content that expresses or incites harm toward an individual or a group of people based on one or more of their personal characteristics, such as gender, race, religion, sexuality, etc. Hate speech detection is critical for protecting online users from abuse and for enabling service providers to offer a safe and trusted environment for their users. It is a challenging task in the domain of Natural Language Processing (NLP) to accurately distinguish between hate and non-hate text, as hateful content has significant implications for the moderation of online content and the prevention of harm. Traditionally, HS classifiers have been trained on specialized corpora, often leading to models that demonstrate high performance within the constraints of their training data. However, these dedicated HS classifiers frequently struggle to understand contextual information when processing hateful/abusive content. For example, the fictitious sentence "I hate seeing children being upset about their race" requires HS classifiers to understand the context in which the terms "hate" and "children" are used. The sentence does not convey hate toward any children, despite containing the term "hate" and expressing that the speaker hates seeing children in a certain state, as children are not targeted explicitly or implicitly. Furthermore, manipulated hate speech refers to content intentionally altered to evade detection by traditional classifiers, such as through misspellings, coded language, or emojis. For example, users may replace letters with numbers (e.g., 'fr33 sp33ch' for 'free speech') to bypass keyword-based filters, as observed in studies such as [1]. Without a deep grasp of the nuanced contexts in which language is used, these models are prone to

misclassifications, especially in cases where hate speech is subtle, implicit, or manipulated to evade detection.

In contrast, large language models (LLMs) have been trained on vast amounts of diverse text, equipping them with a more comprehensive understanding of language. This extensive training allows LLMs to capture complex relationships between words and to recognize context in ways that are not as straightforward for models trained solely on HS-specific data. The ability of LLMs to "understand" the context makes them promising candidates to improve the accuracy of HS classification.

In this paper, we provide a systematic evaluation of four commonly used LLMs for HS classification, using eight widely used hate speech datasets. We examine the effectiveness of those LLMs in both binary and multi-label HS classification tasks. Our study aims to evaluate the ability of LLMs to improve hate speech detection, with a particular focus on instances where context significantly influences classification outcomes.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of relevant prior work in this domain. Section 3 presents the evaluation methodology and experimental design employed in this study. Section 4 discusses the experimental results and provides a detailed analysis of the findings. Section 5 examines the limitations of the current study and their potential implications. Section 6 addresses practical deployment considerations and operational requirements for real-world implementation. Section 7 outlines promising directions for future research, and Section 8 concludes this paper with a summary of key contributions and implications.

2. Related Works

The problem of HS detection has garnered significant attention within the NLP community, leading to the development of various methodologies and approaches. Early efforts in this domain predominantly relied on traditional machine learning techniques, such as Support Vector Machines (SVMs) and logistic regression, trained on manually curated hate speech datasets. These approaches often utilized handcrafted features, such as word n-grams, sentiment analysis, and lexicon-based methods [2], to identify hate speech in online content. However, while they were effective to some extent, these models struggled with generalization, particularly when exposed to variations in language use or context manipulation.

The emergence of deep learning techniques marked a significant shift in hate speech detection. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) were leveraged to automatically learn features from text data, leading to improvements in model performance. Notably, the use of Long Short-Term Memory (LSTM) networks allowed models to better capture sequential dependencies in text, aiding in the identification of more complex patterns indicative of hate speech [3]. However, these models still faced limitations, particularly in understanding the broader context of statements, which is crucial for accurately detecting nuanced or implicit hate speech.

More recent work has explored the use of transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT), for hate speech detection [4]. BERT, with its bidirectional attention mechanism, allows for a deeper understanding of context by considering both the left and right contexts of a given word. This approach has shown promise in improving the detection of subtle forms of hate speech and reducing false positives. However, while BERT and similar models have demonstrated strong performance, they are still constrained by the domain-specific nature of the training data and may not fully generalize to diverse or manipulated content.

The advent of LLMs like Generative Pre-trained Transformer 3 (GPT-3) and similar models has opened new avenues for hate speech detection. These models, trained on

extensive and diverse datasets, possess a broader understanding of language and context, enabling them to potentially outperform traditional and even transformer-based HS classifiers. Recent research has begun to explore the application of LLMs in hate speech detection, with initial findings suggesting that these models can capture complex linguistic nuances and reduce misclassifications of non-hateful content.

Using LLMs has several advantages for hate speech detection, as shown in a recent study by Plaza-del Arco et al. [5], which explores the potential of zero-shot learning for HS detection. The researchers investigated the effectiveness of this approach for hate speech detection in three languages with limited labeled datasets. By experimenting with various LLMs across eight benchmark datasets, they revealed the critical role of prompt selection in determining LLM performance. Their findings suggest that prompting, especially when using state-of-the-art LLMs, can achieve performance on par with or even exceeding that of fine-tuned models. This approach offers a promising alternative for hate speech detection in under-resourced languages, underscoring the importance of both prompt design and model choice in enhancing detection accuracy.

The effectiveness of LLMs in detecting offensive and harmful online behavior, particularly sexist and hateful content, was studied in [6]. They examined various LLMs, including zero-shot, few-shot, and fine-tuning approaches, to assess their ability to identify hate speech without model training. They reported that LLMs can successfully detect hate speech, with the encoder–decoder model achieving the highest performance. Specifically, the Zephyr model [7] scored 86.811% on the Explainable Detection of Online Sexism (EDOS) test set and 57.453% on the Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) test set, surpassing the previous best results on the HatEval leader-board. However, their study also highlighted challenges in contextual learning, particularly in distinguishing between different types of hate speech and figurative language, and noted that the fine-tuned approach may lead to a higher rate of false positives.

The work by Saha et al. [8] investigated the performance of various LLMs in zero-shot settings for the task of counter-speech generation, a critical area in combating hate speech online. They focused on four prominent LLMs—GPT-2, DialoGPT, ChatGPT V3.5, and FlanT5—making it the first comprehensive analysis of these LLMs' effectiveness in this context without fine-tuning. For GPT-2 and DialoGPT, they further examined the effect of model size (small, medium, and large) on performance. Additionally, they proposed three different prompting strategies to generate various types of counter-speech and evaluated their impact on model performance. Their findings revealed that while generation quality improved by 17% for two datasets, there was a 25% increase in toxicity with larger models. GPT-2 and FlanT5 produced higher-quality counter-speech but exhibited greater toxicity compared to DialoGPT. Notably, ChatGPT consistently outperformed the other models across all metrics. The proposed prompting strategies significantly enhanced counter-speech generation across all models, highlighting their importance in this task.

Nirmal et al.'s study [9] addressed the critical need for interpretable hate speech detection methods on social media platforms, where users often exploit anonymity to spread offensive content. While existing detection methods largely operate as black-box models with little transparency, the authors proposed a novel approach that leverages state-of-the-art LLMs to extract interpretive features, or rationales, from input text. These rationales are then used for training a baseline hate speech classifier, ensuring that the model remains interpretable by design. The proposed framework effectively integrates the advanced textual understanding capabilities of LLMs with the discriminative power of modern hate speech classifiers, resulting in a system that is both accurate and transparent. Comprehensive evaluations on multiple English-language social media datasets demonstrated the effectiveness of LLM-extracted rationales and showed that the model's

performance was largely retained even after incorporating interpretability. This approach offers a promising direction for creating more transparent and explainable hate speech detection systems.

In summary, while significant progress has been made in the development of hate speech classifiers, the challenge of accurately detecting hate speech in varied and context-dependent scenarios persists. Moreover, current HS classifiers lack comprehensive evaluations of LLMs specifically in the context of hate speech classification, particularly across different types of hate speech datasets (e.g., binary vs. multi-class). This paper contributes to the growing body of literature by evaluating the performance of publicly available LLMs in hate speech classification, comparing their effectiveness across different datasets and classification tasks. By doing so, we aim to advance the understanding of how LLMs can be leveraged to address the limitations of current hate speech detection methods.

3. Methodology

Our methodology involved several key steps: (i) selecting and pre-processing datasets, (ii) selecting the most open-access LLMs, (iii) designing and conducting experiments for binary and multi-class classification, and (iv) assessing the performance of the LLMs.

3.1. Dataset Selection and Preparation

To comprehensively evaluate the performance of LLMs in hate speech classification, we utilized several commonly used HS datasets that cover both binary and multi-class classification tasks. The datasets used in this study are presented in Table 1.

Table 1. Datasets used for binary classification.

Dataset	Ref.	Dataset Size
Suryawanish	[10]	743
Salminen	[11]	3222
Davidson	[12]	2860
Gibert	[13]	10,944
Waseem	[14]	10,458
Qian	[15]	27,546
Vidgen	[16]	8186

For the binary HS classification task, we selected widely used HS datasets from the literature, in which most examples are labeled as either "hate" or "not-hate". These datasets encompass diverse forms of hate speech across different contexts. For the multi-class datasets, we considered only the "hate" and "not-hate" classes. For example, the Vidgen dataset [16] originally contained more nuanced classifications. The objective of this binarization task was to ensure that the datasets contained consistent labels and were thus comparable. The label binarization process was as follows:

- Fine-grained HS labels (such as "racist" or "sexiest") were consolidated into broader categories. Content labeled with any form of HS was reclassified under the general label "hate", and those labeled "neutral" or "not-hateful" were converted to "not-hate".
- Ambiguous labels were removed by excluding any content with unclear categorization, such as "abusive", since such content may not necessarily be considered hateful.
- Content labeled "neutral", "not-hate", or "not abusive" was reclassified as "not-hate".

For the multi-class HS classification task, we utilized two datasets: Vidgen [16] and Kennedy [17]. In these datasets, each instance is assigned a label corresponding to different categories of hate speech, such as targeting an affiliation or identity. The different classes are shown in Table 2.

Table 2. Multi-class datasets and their different classes.

Dataset	Hate Speech Classes	Ref.
Vidgen	Affiliation, Person, Identity	[16]
Kennedy	Race, Religion, Origin, Gender, Sexuality, Age, Disability	[17]

3.2. LLM Selection and Rationale

The explosive growth of large language models has created a complex landscape of competing architectures, training approaches, and optimization techniques, each tailored to distinct computational challenges and application needs.

3.2.1. LLM Selection

In this paper, we analyze four commonly used instruction-tuned language models that exemplify different strategies for optimizing the critical trade-offs between performance, computational efficiency, and task specialization. The tested models are as follows:

- Meta-Llama-3-8B-Instruct.Q4_0 (Meta) [18]: This model is part of the Meta-Llama series, designed to excel in instruction-following tasks. With 8 billion parameters, it has been fine-tuned on diverse instructional data, making it well-suited for tasks requiring nuanced understanding and contextual interpretation.
- Phi-3-Mini-4k-instruct.Q4_0 (Phi) [19]: This model is a smaller yet efficient member of the Phi series, with 3 billion parameters. Despite its size, it is designed for instruction-based tasks and is optimized for quick inference, making it an effective choice for scenarios where computational resources are limited.
- Nous-Hermes-2-Mistral-7B-DPO.Q4_0 (Hermes) [20]: This model, part of the Nous-Hermes series, incorporates the Mistral architecture and has 7 billion parameters. It has been fine-tuned with a focus on dialogue and contextual understanding, which are crucial for accurately identifying hate speech in conversational contexts.
- WizardLM-13B-v1.2.Q4_0 (WizardLM) [21]: This is a large 13-billion-parameter model from the WizardLM series, designed to perform well on a variety of NLP tasks, including text generation and comprehension. Its architecture is optimized for both speed and accuracy, providing a balance of performance and resource efficiency.

These models span a range of parameter counts from 3.8 billion to 13 billion, employ varying context window sizes from 4K to 32K tokens, and utilize distinct training paradigms, including direct preference optimization (DPO), instruction-following enhancement, and compact efficiency optimization. All models examined in this comparison utilize Q4_0 quantization, making them suitable for deployment in resource-constrained environments while maintaining competitive performance. A comparison of the models is presented in Table 3.

Table 3. Comparison of large language models.

Characteristic	Llama-3-8B	Phi-3-Mini-4k	Hermes-2-Mistral-7B	WizardLM-13B
Base Model	Llama 3	Phi-3	Mistral-7B	Llama 2
Parameters	8B	3.8B	7B	13B
Quantization	Q4_0	Q4_0	Q4_0	Q4_0
Context Length	8192	4096	32,768	4096
Developer	Meta	Microsoft	Nous Research	WizardLM Team
Training Focus	General Instruct	Compact Efficiency	DPO Fine-Tuning	Instruction-Following
Specialization	Balanced	Mobile/Edge	Conversational	Complex Reasoning
Memory Usage	Medium	Low	Medium	High
Performance	High	Good	High	Very High

3.2.2. LLM Selection Rationale

Given the novel nature of applying LLMs specifically to hate speech detection, established benchmarks for model performance in this domain are currently limited. This necessitates a strategic approach to model selection based on architectural diversity, practical deployment considerations, and proven performance across related NLP tasks. Our comparative study employs four distinct LLMs: Llama-3-8B, Phi-3-Mini-4k, Hermes-2-Mistral-7B, and WizardLM-13B. To ensure a comprehensive evaluation while maintaining practical relevance, the selection was guided by several key criteria.

Architectural Diversity: The selected models showcase a range of modern transformer architectures. Llama 3 and WizardLM are based on the Llama architecture, known for its efficiency and strong performance. In contrast, Hermes-2 is built upon the Mistral architecture, which notably employs Grouped-Query Attention (GQA) to accelerate inference speed. Phi-3 represents a distinct architectural design from Microsoft, optimized to balance performance and computational cost. This diversity allows for an examination of how different architectural choices impact downstream task performance and efficiency.

Parameter Scale Coverage: Our selection spans a practical range of model sizes, crucial for understanding the trade-offs between performance and resource requirements. At the smaller end, we include Phi-3-Mini-4k with approximately 3.8 billion parameters. The 7 to 8 billion parameter range is represented by Hermes-2-Mistral-7B and Llama-3-8B, reflecting a popular balance for many applications. At the higher end, WizardLM-13B, with its 13 billion parameters, allows us to investigate the benefits of a larger model size on more complex reasoning and generation tasks.

Edge-Device Compatibility: A key motivation for this study is the increasing need for capable models that can operate in resource-constrained environments. Models like Phi-3-Mini-4k and Llama-3-8B are specifically designed with a smaller memory and computational footprint, making them strong candidates for deployment on edge devices such as mobile phones and laptops. The inclusion of the 7B models also allows for an evaluation of the upper limits of what is currently feasible on high-end edge hardware.

Training Specializations: The chosen models exhibit varied training and fine-tuning methodologies, which influence their capabilities. Llama-3-8B is a base model pre-trained on a massive and diverse dataset. In contrast, Hermes-2-Mistral-7B has been fine-tuned on a large, curated dataset of open-source conversational data, enhancing its performance in dialogue and instruction-following. WizardLM-13B employs an innovative "Evol-Instruct" method, where instruction data is automatically generated and progressively complexified to improve the model's ability to follow intricate commands. Phi-3-Mini-4k was trained on a heavily filtered, "textbook-quality" dataset, aiming to achieve high performance with a smaller model size.

Context Length Variation: The ability to process long sequences of text is a critical differentiator for modern LLMs. Our selection includes models with different context window sizes to assess this capability. Phi-3-Mini-4k has a default 4096-token context length. Both Llama-3-8B and Hermes-2-Mistral-7B (based on the original Mistral-7B) feature a standard 8192-token context window. While WizardLM-13B was initially based on a model with a shorter context, its more recent versions have been adapted for longer contexts, providing another point of comparison. This variation is essential for evaluating performance on tasks requiring the assimilation of extensive information.

LLM Quantization: All large language models used in this study were quantized using the Q4_0 format from the GGUF framework. This decision was based on a pragmatic trade-off between maintaining high model fidelity and ensuring computational feasibility. The Q4_0 quantization level reduces the model's memory footprint by approximately 75% compared to its native FP16 precision, enabling comprehensive experimentation across

multiple models and datasets on consumer-grade and standard research GPUs (e.g., with 16–24 GB of VRAM). This strategy is widely adopted, as it typically preserves the vast majority of a model's performance on downstream tasks while making the experimental setup more accessible and reproducible for the broader research community [22]. More aggressive quantization schemes were avoided due to the higher risk of performance degradation on a nuanced task such as hate speech classification.

3.3. Experimental Design

To evaluate the performance of LLMs in hate speech detection, we designed experiments for both binary and multi-class classification tasks. For the **binary classification task**, the LLMs were prompted as follows:

> *"Write 1 if the text is hate speech and 0 if it is not. Do not include any comments; your response must be either 0 or 1. Only provide the numeric value. The text is:"*

For the **multi-class classification task**, the LLMs were provided with a more detailed prompt to classify the text into specific categories of hate speech.

The prompt for the Vidgen dataset [16] was as follows:

> *"Please respond with only a single numeric value based on the following criteria:*
> *0 if the text is not hate speech,*
> *1 if it is hate speech targeting an affiliation,*
> *2 if it is hate speech targeting a person,*
> *3 if it is hate speech targeting an identity,*
> *Only provide the numeric value.*
> *The text is:"*

The prompt for the Kennedy dataset [17] was as follows:

> *"Please respond with only a single numeric value based on the following criteria:*
> *0 if the text is not hate speech,*
> *1 if it is hate speech targeting race,*
> *2 if it is hate speech targeting religion,*
> *3 if it is hate speech targeting origin,*
> *4 if it is hate speech targeting gender,*
> *5 if it is hate speech targeting sexuality,*
> *6 if it is hate speech targeting age,*
> *7 if it is hate speech targeting disability. Only provide the numeric value.*
> *The text is:"*

The classification tasks were conducted by instructing various LLMs to predict the labels of entries in multiple datasets. The experimental process is detailed in Algorithm 1. A critical step in this pipeline is the parsing and validation of the models' responses. To ensure a robust evaluation, we implemented a strict two-stage process. First, to prevent the models from generating extraneous text beyond the desired label, the max_new_tokens parameter was set to 1 during inference. Second, a validation script analyzed the single-token response, r_{ij}, to confirm it was a digit corresponding to a valid class label. If the output was a valid digit, it was recorded as the predicted label, p_{ij}. In cases where the output was non-numeric or otherwise invalid, the prediction was conservatively counted

as an incorrect classification. This final, validated prediction was then compared against the ground-truth label to compute the performance metrics.

Algorithm 1: Hate speech detection methodology

Input: Datasets $D = D_1, D_2, \ldots, D_n$, LLMs $L = L_1, L_2, \ldots, L_m$
Output: Accuracy results for each LLM on each dataset
for *dataset* $D_i \in D$ **do**
  for *text instance* $t \in D_i$ **do**
    for *LLM* $L_j \in L$ **do**
      Preprocess text if necessary
    end
    Create prompt for LLM L_j
    Send prompt to LLM L_j and obtain response r_{ij}
    Parse and validate r_{ij} to get predicted label p_{ij}
  end
end
Calculate overall performance for LLM L_j on dataset D_i

The algorithm systematically processed each sentence in the datasets, constructed the appropriate prompt for the given LLM, and recorded the model's response. This response was then used to compute the accuracy of each model's predictions. The dataset labels served as the ground truth.

4. Classification Results and Analysis

4.1. Binary Classification

Table 4 presents a comparison of the classification performance of four models—Phi, Meta, WizardLM, and Hermes—using standard metrics: Cohen's Kappa, accuracy, F1 score (micro, macro, and weighted), recall, and precision. The results reveal varying levels of effectiveness across datasets, highlighting which models consistently outperformed others and which models struggled to deliver reliable classifications. This comparison provides critical insights into the models' strengths and weaknesses.

Table 4. Performance metrics of large language models on the binary classification task.

Dataset	Model	Kappa	Acc.	Micro F1	Macro F1	Wtd. F1	Recall	Prec.
Davison	Phi	0.6622	0.8311	0.8311	0.8301	0.8301	0.8311	0.8395
	Meta	0.7385	0.8692	0.8692	0.8688	0.8688	0.8692	0.8747
	WizardLM	−0.0518	0.4741	0.4741	0.4741	0.4741	0.4741	0.4741
	Hermes	**0.7790**	**0.8895**	**0.8895**	**0.8895**	**0.8895**	**0.8895**	**0.8895**
Salmenin	Phi	0.5163	0.7582	0.7582	0.7487	0.7487	0.7582	0.8039
	Meta	**0.6725**	**0.8363**	**0.8363**	**0.8362**	**0.8362**	**0.8363**	**0.8364**
	WizardLM	0.0221	0.5111	0.5111	0.4876	0.4876	0.5111	0.5136
	Hermes	0.5431	0.7716	0.7716	0.7634	0.7634	0.7716	0.8148
Suryawanish	Phi	0.1617	0.5809	0.5809	0.5159	0.5159	0.5809	**0.6746**
	Meta	**0.2904**	**0.6452**	**0.6452**	**0.6324**	**0.6324**	**0.6452**	0.6687
	WizardLM	0.2178	0.6089	0.6089	0.6088	0.6088	0.6089	0.6091
	Hermes	0.2112	0.6056	0.6056	0.5688	0.5688	0.6056	0.6603
Gibert	Phi	0.3591	0.6795	0.6795	0.6588	0.6588	0.6795	0.7373
	Meta	**0.5630**	**0.7815**	**0.7815**	**0.7812**	**0.7812**	**0.7815**	**0.7829**
	WizardLM	−0.0056	0.4972	0.4972	0.4967	0.4967	0.4972	0.4972
	Hermes	0.4941	0.7470	0.7470	0.7419	0.7419	0.7470	0.7684

Table 4. *Cont.*

Dataset	Model	Kappa	Acc.	Micro F1	Macro F1	Wtd. F1	Recall	Prec.
Waseem	Phi	0.0873	0.5436	0.5436	0.4677	0.4677	0.5436	0.6016
	Meta	**0.3695**	**0.6847**	**0.6847**	**0.6696**	**0.6696**	**0.6847**	**0.7261**
	WizardLM	−0.0650	0.4675	0.4675	0.4654	0.4654	0.4675	0.4670
	Hermes	0.1749	0.5875	0.5875	0.5240	0.5240	0.5875	0.6872
Qian	Phi	0.4302	0.7151	0.7151	0.7064	0.7064	0.7151	0.7438
	Meta	**0.5967**	**0.7983**	**0.7983**	**0.7982**	**0.7982**	**0.7983**	**0.7992**
	WizardLM	0.0503	0.5251	0.5251	0.5163	0.5163	0.5251	0.5271
	Hermes	0.4911	0.7455	0.7455	0.7396	0.7396	0.7455	0.7702
Vidgen	Phi	0.2448	0.6224	0.6224	0.5880	0.5880	0.6224	0.6838
	Meta	**0.4283**	**0.7142**	**0.7142**	**0.7093**	**0.7093**	**0.7142**	**0.7294**
	WizardLM	0.0259	0.5130	0.5130	0.5025	0.5025	0.5130	0.5141
	Hermes	0.2800	0.6400	0.6400	0.6067	0.6067	0.6400	0.7118

Note: Acc. = accuracy; Prec. = precision; Wtd. F1 = weighted F1. Bold values indicate the best performance on each dataset.

Table 5 summarizes the average performance of the LLMs across the datasets used in this study. Figure 1 further illustrates the distribution of these metrics, highlighting the variability and comparative performance of each model across the different datasets.

Table 5. Average model performance across all datasets.

Metric	Phi	Meta	WizardLM	Hermes
Kappa	0.35165	**0.52269**	0.02769	0.42477
Accuracy	0.67583	**0.76134**	0.51385	0.71239
Micro F1	0.67583	**0.76134**	0.51385	0.71239
Macro F1	0.64508	**0.75654**	0.50734	0.69057
Weighted F1	0.64508	**0.75654**	0.50734	0.69057
Recall	0.67583	**0.76134**	0.51385	0.71239
Precision	0.72635	**0.77391**	0.51459	0.75746

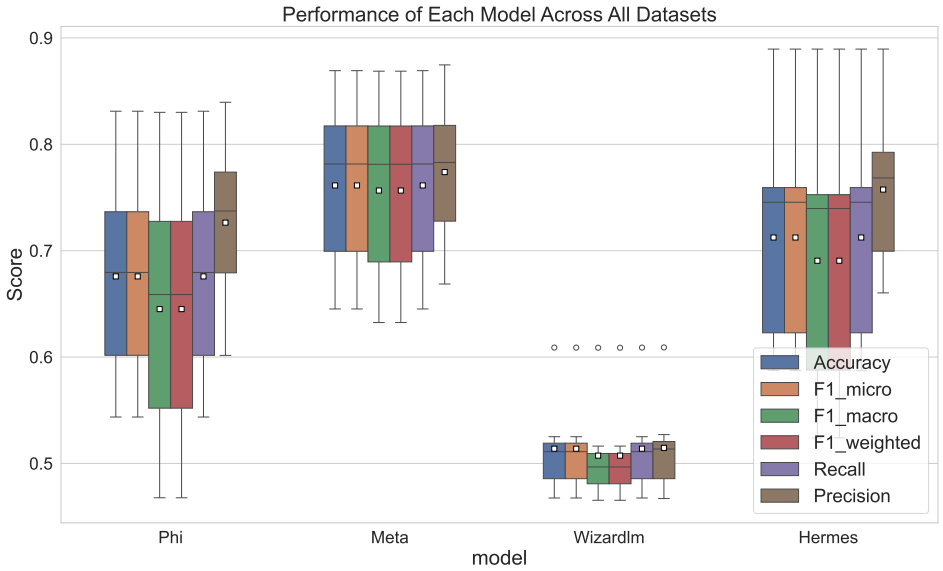


Figure 1. Performance of each model across all datasets, highlighting both the mean (represented by a white square) and median performance.

The results indicate that the Meta model [18] outperformed the other models across most of the tested datasets and evaluation metrics. It achieved the highest average accuracy (0.76134), micro F1 (0.76134), and macro F1 (0.75654), indicating that it is not only reliable

in correctly classifying the data but also balanced in its ability to handle both majority and minority classes effectively. Furthermore, its high precision (0.77391) suggests that it made fewer false-positive errors compared to the other models. This performance might be attributed to the model's architecture or training strategies that are particularly well suited for binary classification tasks across diverse datasets. The confusion matrices for the Meta model are presented in Figure 2; 0 represents "not-hate" and 1 represents "hate". It appears the model confused "not-hate" with "hate" in five out of seven datasets.

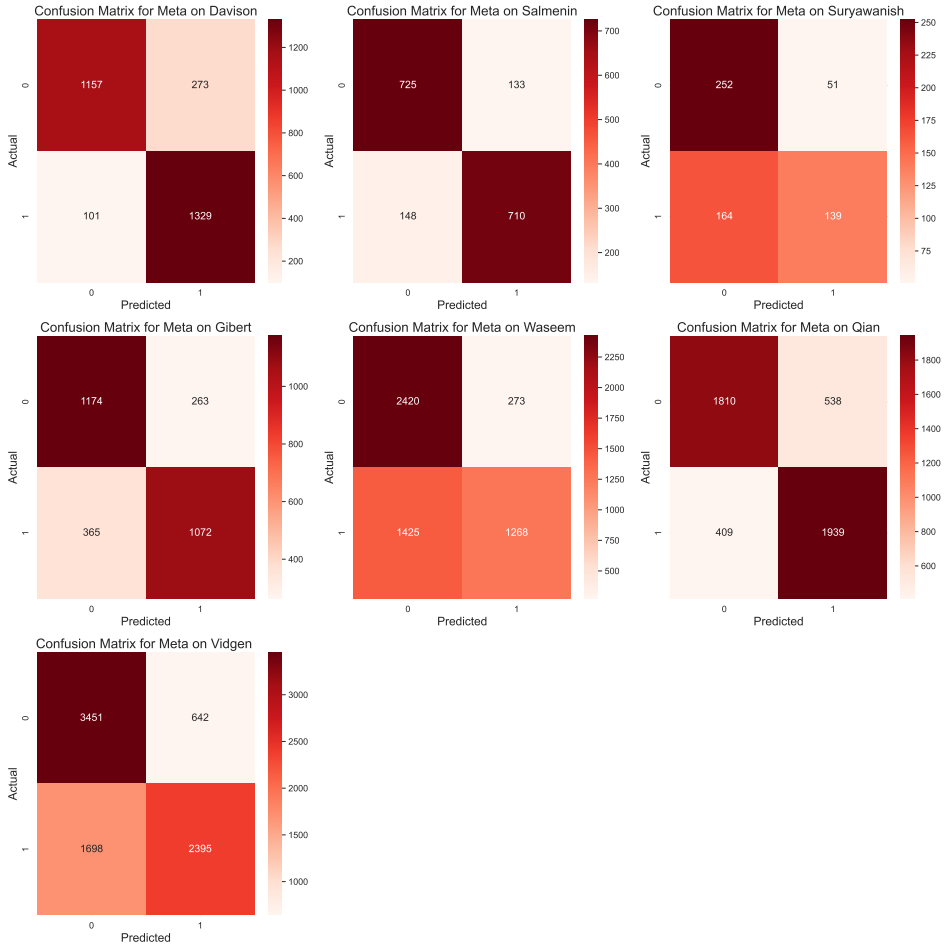


Figure 2. Confusion matrices of the Meta model.

The Hermes model [20] also performed competitively, particularly in terms of recall (0.71239) and precision (0.75746). The model's recall indicates that it has a strong ability to correctly identify positive instances, which is critical in applications where missing a positive case has a high cost. However, its slightly lower Kappa (0.42477) compared to Meta suggests that its agreement with the true labels was not as high, which may indicate some variability in performance across different datasets. Nonetheless, the Hermes model's balance between recall and precision shows that it is a robust choice when the goal is to maintain a good trade-off between catching all relevant cases and minimizing false positives.

The Phi model [19] exhibited moderate performance, with an average accuracy of 0.67583 and a precision of 0.72635. Although these results were lower than those of the Meta and Hermes models, the Phi model still showed consistent performance across the different metrics. Its F1 scores (both micro and macro) were relatively balanced, suggesting that the model did not significantly favor one class over another. This model might be suitable for scenarios in which interpretability or simplicity is favored over top-tier performance.

The WizardLM model [21] exhibited the weakest performance across all metrics, with an average Kappa of 0.02769 and an accuracy of 0.51385. Its performance was close to random guessing, particularly evident in the low F1 scores and precision (0.51459). This suggests that the WizardLM model struggled with these binary classification tasks, potentially due to inadequate model capacity, suboptimal hyperparameter tuning, or a lack of task-specific training. The model's low Kappa score further highlights its poor agreement with the true labels, indicating that its predictions were not consistently reliable. This result suggests that the WizardLM model may not be suitable for binary classification tasks, especially in contexts requiring high accuracy and reliability.

4.2. Multi-Class Classification

The results for the multi-class classification task on the Vidgen [16] and Kennedy [17] datasets are presented in Table 6, showing varying levels of performance among the four models—Wizard, Phi, Hermes, and Meta—across the evaluation metrics. The Vidgen dataset consists of three classes of HS, while the Kennedy dataset contains seven classes.

Table 6. Performance metrics of LLMs on the multi-class classification task.

Dataset	Model	Kappa	Accuracy	Micro F1	Macro F1	Weighted F1	Recall	Precision
Vidgen	Phi	0.08226	0.580545	0.580545	0.112344	0.6914	0.580545	0.889401
	Meta	0.125767	0.742737	0.742737	0.248307	0.806462	0.742737	**0.889668**
	Hermes	**0.156025**	0.833877	0.833877	**0.280837**	0.851569	0.833877	0.87374
	WizardLM	0.145076	0.864848	**0.864848**	0.276903	**0.861384**	**0.864848**	0.867348
Kennedy	Phi	0.045342	0.210755	0.210755	0.099052	0.278826	0.210755	**0.723553**
	Meta	0.066452	0.259884	0.259884	0.087799	0.356399	0.259884	0.651437
	Hermes	0.08737	0.309698	0.309698	0.099408	0.423658	0.309698	0.678297
	WizardLM	**0.13222**	**0.594667**	**0.594667**	**0.13243**	**0.548781**	**0.594667**	0.514553

The results reveal notable differences in model performance across the two datasets, reflecting the impact of class distribution and complexity on the models' ability to generalize effectively. For the Vidgen dataset [16], which involved a simpler three-class classification problem, the models generally exhibited strong performance. WizardLM achieved the highest accuracy (86.48%) and weighted F1 score (86.13%), closely followed by Hermes, with an accuracy of 83.39%. Despite their competitive accuracy, both models showed relatively low macro F1 scores, indicating some difficulty in balancing predictions across classes. Phi, with the lowest accuracy (58.05%), struggled to provide robust classification, particularly in terms of the macro F1 score (11.23%), suggesting poor performance on underrepresented or more challenging classes.

In contrast, the Kennedy dataset, which involved a more complex seven-class classification task, posed greater challenges for the models. WizardLM again outperformed the other models with an accuracy of 59.47%, although this was notably lower compared to its performance on Vidgen, reflecting the increased difficulty of multi-class classification with a larger number of categories. Meta and Hermes demonstrated moderate accuracies of 25.99% and 30.97%, respectively, while Phi performed the worst, achieving an accuracy of only 21.07%. The relatively low macro F1 scores across all models highlight the difficulties in providing consistent predictions across the broader range of classes in the Kennedy dataset.

Overall, the results suggest that model performance declines as the number of classes increases, with significant variations in how well the models balance class-specific predictions. WizardLM consistently performed well across datasets, while Phi showed limitations in handling multi-class tasks.

The WizardLM model achieved the highest accuracy (0.8648) and micro F1 score (0.8648) among all models, indicating its strong ability to correctly classify the majority of instances. However, its low macro F1 score (0.2769) suggests that the model struggled to maintain balanced performance across all classes, particularly the minority classes. The relatively low Kappa score (0.1451) also suggests that, while the model appeared effective overall, its agreement with the gold standard labels was limited, possibly due to its high dependence on the dominant class.

The Phi model showed a lower accuracy (0.5805) and micro F1 score (0.5805) compared to the other models, indicating weaker overall accuracy in classifying instances. Despite this, the Phi model achieved the highest precision (0.8894), suggesting that it made fewer false-positive errors than the other models. However, its low macro F1 score (0.1123) and Kappa (0.0823) further emphasize its difficulty in dealing with class imbalance, showing a tendency to overestimate certain classes while missing others.

The Hermes model achieved the highest Kappa score (0.1560) among all models, indicating stronger agreement with the actual labels compared to the other models. It also demonstrated a high accuracy (0.8339) and micro F1 score (0.8339), indicating robust performance in correctly classifying instances. However, similar to the WizardLM model, Hermes showed a relatively low macro F1 score (0.2808). Despite this, Hermes's relatively high weighted F1 score (0.8516) and precision (0.8737) indicate that it still performed well when considering the distribution of class instances.

The Meta model achieved balanced performance with moderate scores across most metrics. With an accuracy of 0.7427 and a micro F1 score of 0.7427, it ranked lower than the WizardLM and Hermes models in terms of overall correctness. However, its macro F1 score (0.2483) was higher than that of the Phi model, suggesting a better ability to handle class imbalance. The Meta model also exhibited a strong weighted F1 score (0.8065) and high precision (0.8897), indicating that it effectively minimized false positives, even though its recall was comparatively lower.

Overall, the results indicate that while the WizardLM model achieved the highest overall accuracy, it, along with the other models, struggled with class-level recognition, as evidenced by the low macro F1 scores. The Hermes model stood out slightly with the highest Kappa score, suggesting a better match with the gold standard labels, but it also faced challenges with minority class performance. The Phi and Meta models provided trade-offs between precision and recall, with the Phi model particularly excelling in minimizing false positives.

5. Limitations

5.1. Technical Limitations

Despite the promising results, several limitations should be acknowledged in this study. First, the performance of LLMs was evaluated on a relatively small number of datasets (seven for binary and two for multi-class), which may limit the generalizability of the findings across broader and more diverse datasets. The specific characteristics of the datasets used, such as the number of classes (three in Vidgen and seven in Kennedy), may have influenced the results, making it difficult to predict how these models would perform on datasets with different class structures.

Second, the models' performance in multi-class classification tasks, particularly on the Kennedy dataset, revealed significant weaknesses in balancing class-specific predictions. The low macro F1 scores suggest that the models struggled with underrepresented classes or those that were more complex to classify. This highlights a limitation in their ability to handle highly imbalanced datasets, which is a common scenario in real-world applications.

Additionally, this study did not explore model-level fine-tuning or hyperparameter optimization, which could potentially improve performance, especially for models like Phi that consistently underperformed. Lastly, computational constraints and limited resources restricted a deeper exploration into the reasons behind the variation in performance across the models and datasets, such as how each model's architecture or training may have influenced its multi-class capabilities. Further research could address these limitations by incorporating more datasets, optimizing models, and examining broader classification tasks.

5.2. Ethical Risks and Societal Implications of LLM-Based Hate Speech Detection

The deployment of large language models for hate speech detection, while technically promising, raises significant ethical concerns that extend beyond traditional performance metrics. As these systems increasingly influence online discourse moderation, it is crucial to examine their potential societal implications and associated risks.

5.2.1. Censorship and Freedom of Expression

LLM-based hate speech detection systems operate at the intersection of content moderation and free speech, creating inherent tensions that require careful consideration. These systems may exhibit conservative bias when encountering ambiguous content, potentially flagging legitimate discourse as hate speech to minimize false negatives. This over-censorship phenomenon poses particular risks to marginalized communities whose linguistic expressions, cultural references, or political viewpoints may deviate from the mainstream patterns represented in training data. The automated nature of LLM-based moderation can lead to the suppression of legitimate political dissent, minority perspectives, or culturally specific expressions that algorithms misinterpret as hateful content. For instance, discussions of historical injustices, critiques of systemic discrimination, or reclaimed language use within affected communities may be inappropriately flagged, effectively silencing voices that contribute to important social discourse.

5.2.2. Cultural Sensitivity and Representational Bias

The predominant training of LLMs on Western, English-language datasets creates significant risks for cross-cultural content moderation. These models may inadequately understand cultural nuances, idiomatic expressions, humor, or context-dependent language use prevalent in non-Western communities. Consequently, content moderation systems may exhibit discriminatory enforcement patterns, disproportionately flagging content from linguistically or culturally diverse users. This cultural insensitivity can manifest as systematic bias against certain linguistic patterns, dialectal variations, or cultural communication styles. Such bias not only undermines the fairness of content moderation but also risks cultural imperialism in digital spaces, in which dominant cultural norms embedded in AI systems suppress diverse forms of expression and communication.

5.2.3. Transparency and Algorithmic Accountability

The opacity of LLM decision-making processes presents significant challenges for accountability in content moderation. Users whose content is flagged or removed by LLM-based systems often receive limited explanation for these decisions, undermining their ability to understand, learn from, or contest moderation actions. This lack of transparency raises due process concerns, particularly when moderation decisions significantly impact users' ability to participate in digital discourse. The complexity of LLM architectures makes it difficult even for platform operators to fully understand why specific content was flagged, complicating appeals processes and error-correction mechanisms. This opacity

can erode user trust in platform moderation and create power imbalances between users and automated systems that govern their online participation.

5.2.4. Amplification of Societal Biases

LLMs inevitably reflect biases present in their training data, which often mirror existing societal prejudices and power structures. When deployed for hate speech detection, these systems may perpetuate discriminatory patterns by exhibiting differential sensitivity to hate speech targeting different demographic groups. For example, models might show higher sensitivity to certain types of hate speech while being less effective at detecting harassment targeting marginalized communities. This bias amplification can create feedback loops where existing inequalities in digital spaces are reinforced and institutionalized through automated moderation systems. Communities already facing discrimination may experience inadequate protection from hate speech while simultaneously facing higher rates of content removal due to biased algorithmic interpretations of their expressions.

5.2.5. Scale and Human Oversight Challenges

The deployment of LLM-based moderation systems at scale presents unique ethical challenges related to the magnitude of potential impact. Systematic errors or biases in these systems can affect millions of users simultaneously, creating widespread consequences that may be difficult to detect and correct in real time. The sheer volume of content processed by these systems makes comprehensive human oversight practically challenging, potentially allowing biased or erroneous decisions to persist without adequate review. The economic pressures to reduce human moderation costs may lead to over-reliance on automated systems, undermining the human judgment necessary for nuanced content evaluation. This shift toward automation raises questions about the appropriate balance between efficiency and the careful consideration required for decisions that significantly impact users' rights and participation in digital discourse.

5.2.6. Democratic Discourse and Self-Censorship

LLM-based hate speech detection systems may inadvertently influence the quality and diversity of democratic discourse online. Users' awareness of automated moderation can lead to anticipatory self-censorship, where individuals avoid discussing controversial but legitimate topics due to fear of algorithmic misinterpretation. This chilling effect can impoverish public discourse by reducing the range of perspectives and topics considered acceptable for discussion. The potential for creating echo chambers through biased content filtering poses additional risks to democratic participation. If moderation systems consistently remove certain viewpoints or perspectives due to algorithmic bias, they may contribute to political polarization and reduce exposure to the diverse opinions necessary for informed democratic participation.

5.2.7. Stakeholder Impact and Justice Considerations

The deployment of LLM-based hate speech detection systems affects various stakeholders differently, raising important questions about distributive and procedural justice. Content creators, particularly those from marginalized communities, may bear disproportionate costs from false-positive detections, including lost revenue, reduced visibility, and platform penalties. Meanwhile, the benefits of improved hate speech detection may not be equally distributed, potentially creating scenarios where some communities receive better protection than others. Platform moderators, often members of marginalized communities themselves, may face psychological harm from reviewing content flagged by LLM systems, particularly if these systems exhibit bias that requires human moderators to repeatedly review discriminatory content. The broader society faces risks from

the potential normalization of automated decision-making in contexts that significantly impact fundamental rights like freedom of expression and equal participation in digital public spheres.

5.2.8. Recommendations for Ethical Deployment

Given these substantial ethical risks, the deployment of LLM-based hate speech detection systems requires careful consideration of safeguards and oversight mechanisms. Key recommendations include implementing robust bias auditing and fairness-testing protocols; ensuring meaningful human oversight and appeals processes; developing culturally sensitive evaluation frameworks; maintaining transparency about system limitations and decision-making processes; and engaging diverse stakeholders in system design and evaluation. The technical capabilities demonstrated in this study must be balanced against these ethical considerations to ensure that advances in hate speech detection contribute to more equitable and inclusive digital environments rather than perpetuating or amplifying existing forms of discrimination and bias.

6. Practical Deployment and Operational Considerations

6.1. Real-World Applications and Implementation

The practical implications of our findings extend across multiple domains of content moderation and digital safety. For social media platforms, our results suggest that implementing LLM-based hate speech detection could significantly reduce the burden on human moderators while improving detection accuracy. The demonstrated edge-device compatibility of our evaluated models, achieved through Q4_0 quantization, makes real-time content moderation feasible for platforms with diverse computational constraints. This capability enables both large-scale platforms and smaller community-driven websites to deploy sophisticated hate speech detection without requiring extensive cloud infrastructure.

Educational institutions can leverage these findings to develop more effective online safety systems for learning management platforms and student communication channels. The superior contextual understanding exhibited by models like Meta-Llama-3-8B makes them particularly suitable for educational environments where nuanced language use and diverse cultural expressions are common. Similarly, corporate environments can implement these systems to maintain respectful workplace communication in digital collaboration platforms while minimizing false positives that could inhibit legitimate professional discourse.

The practical deployment of these models in real-world environments, such as for real-time chat moderation in educational platforms or high-volume content filtering in corporate settings, introduces critical operational challenges. For instance, the latency (prediction speed) and throughput (predictions per second) of LLMs must be sufficient to handle content at scale. While our use of quantized models helps improve performance on standard hardware, deploying these systems for instantaneous feedback requires further engineering and optimization. Moreover, these environments often rely on human-in-the-loop workflows, where interpretability becomes paramount. A human moderator must be able to understand the model's reasoning to effectively review flagged content and handle user appeals. Therefore, future work should not only focus on improving accuracy but also on enhancing the efficiency and transparency of these models to ensure they are both effective and trustworthy in practice.

6.2. Policy and Regulatory Implications

Our research provides empirical evidence that can inform emerging policy frameworks for automated content moderation. The demonstrated variability in model performance across different types of hate speech suggests that regulatory approaches should account

for the technical limitations and capabilities of different AI systems. Policymakers can use our benchmarking framework to establish minimum performance standards for hate speech detection systems deployed in regulated environments.

The edge-device deployment capabilities we demonstrated have particular relevance for privacy-focused regulatory frameworks, such as GDPR, where on-device processing can reduce data transfer and storage requirements while maintaining detection effectiveness. This technical capability enables compliance with data localization requirements and supports the development of privacy-preserving content moderation systems. Furthermore, our findings on the ethical risks and bias considerations provide a foundation for developing responsible AI guidelines specific to hate speech detection. Regulatory bodies can reference our analysis of fairness metrics and bias assessment methodologies when establishing auditing requirements for automated content moderation systems.

6.3. Industry Standards and Best Practices

The comprehensive evaluation framework developed in this study can serve as a template for industry-wide standardization of hate speech detection benchmarking. Our methodology provides a reproducible approach that technology companies can adopt to evaluate and compare different LLM-based solutions before deployment. Our analysis of the trade-offs between model performance, computational efficiency, and deployment constraints offers practical guidance for system architects designing content moderation pipelines. The demonstrated effectiveness of smaller models like Phi-3-Mini-4k in resource-constrained environments provides viable alternatives for organizations with limited computational budgets while maintaining acceptable detection performance.

6.4. Societal Impact and Digital Rights

The enhanced hate speech detection capabilities demonstrated in this study have significant implications for protecting vulnerable communities in digital spaces while preserving freedom of expression. Our findings suggest that LLM-based systems can reduce both false positives that suppress legitimate discourse and false negatives that allow harmful content to persist. This improvement in detection accuracy supports the creation of safer online environments without disproportionately censoring marginalized voices.

The scalability of our approach, demonstrated through edge-device deployment, democratizes access to sophisticated hate speech detection capabilities. Smaller platforms and community organizations can now implement effective content moderation systems that were previously accessible only to major technology companies, promoting more equitable safety standards across diverse digital communities.

The integration of LLMs into broader content moderation frameworks, as demonstrated in this study, represents a paradigm shift toward more intelligent and adaptable safety systems. As these technologies mature, they will enable the development of personalized moderation systems that can adapt to different community standards while maintaining consistent protection against harmful content.

Ultimately, this work contributes to the broader goal of creating digital environments that are both safe and inclusive, where technology serves to protect and empower users rather than constrain legitimate expression. The methodologies and insights presented here provide a roadmap for achieving this vision through continued research, responsible deployment, and ongoing collaboration between technologists, policymakers, and communities affected by these systems.

7. Future Work

This study establishes foundational benchmarks for LLM performance in text-based hate speech detection, opening several promising avenues for future research that address current limitations and extend the practical applicability of our findings:

1. **Cross-Platform and Multimodal Extensions:** A critical direction for future work could involve expanding the evaluation framework to encompass diverse social media platforms and their unique contextual characteristics. While our current study focuses on text-based detection, future research should investigate how hate speech manifestations vary across platforms like YouTube, Reddit, TikTok, and Instagram, each with distinct user demographics, communication norms, and content formats. This expansion would require developing platform-specific datasets and evaluation metrics that capture the nuanced ways hate speech adapts to different social environments. Building upon our text-based foundation, multimodal hate speech detection [23] represents a natural and necessary evolution. Future work should explore how LLMs can be integrated with computer vision and audio processing models to detect hate speech in videos, memes, and multimedia content. This multimodal approach would significantly enhance real-world deployment readiness by addressing the increasingly visual and interactive nature of online hate speech.
2. **Context-Aware Evaluation Frameworks:** The findings highlight the need for more sophisticated evaluation frameworks that incorporate the complex dynamics of public discourse behavior. Future research should develop benchmarking methodologies (such as [4]) that account for sentiment polarity, toxicity gradients, and user engagement patterns, as they influence hate speech detection effectiveness [24]. This could involve creating sentiment-aware evaluation metrics that assess model performance across different emotional contexts and toxicity-gradient benchmarks that evaluate detection accuracy at varying levels of content harmfulness. Additionally, investigating temporal dynamics and evolving hate speech patterns would provide valuable insights into model robustness over time. Future work should examine how LLM performance degrades or adapts as hate speech evolves linguistically and contextually, potentially incorporating continual learning approaches to maintain detection effectiveness.
3. **Hybrid Classification Frameworks:** A particularly promising direction could involve utilizing LLMs as components within broader classification frameworks rather than as standalone detection systems. Future research should explore hybrid architectures where LLMs serve specific roles such as feature extraction, contextual understanding, or ensemble voting within multi-stage classification pipelines. This approach could combine the semantic understanding capabilities of LLMs with the efficiency and specialization of traditional machine learning models, potentially achieving superior performance while maintaining computational feasibility for large-scale deployment. Such hybrid frameworks could leverage LLMs for tasks like generating contextual embeddings [25], performing semantic similarity analysis, or providing interpretable explanations for classification decisions, while relying on lighter models for initial filtering or real-time processing components.
4. **Robustness and Adversarial Evaluation:** Future work should address the robustness of LLM-based hate speech detection against adversarial attacks and evasion techniques. This could include evaluating model performance against character-level perturbations, linguistic obfuscation, and emerging evasion strategies employed by users attempting to circumvent detection systems. Developing adversarially robust models and evaluation protocols would enhance the practical reliability of hate speech detection systems.

5. **Ethical and Bias Considerations:** Expanding our evaluation framework to include comprehensive bias analysis represents another critical future direction [26]. This should encompass investigating demographic biases, cultural sensitivity across different communities, and fairness metrics that ensure equitable detection performance across diverse user populations. Future research should also explore how different prompting strategies and model fine-tuning approaches can mitigate inherent biases while maintaining detection effectiveness.
6. **Real-World Deployment Studies:** Finally, bridging the gap between academic evaluation and practical deployment requires longitudinal studies of LLM-based hate speech detection systems in real-world environments. Future work should investigate how model performance translates to actual content moderation effectiveness, user satisfaction, and platform safety improvements. This could include studying the interaction between automated detection systems and human moderators, developing effective human-in-the-loop workflows, and measuring the broader societal impact of improved hate speech detection capabilities. These future research directions collectively address the limitations identified in our current study while building upon the foundational benchmarks we have established. By pursuing these avenues, the research community can develop more comprehensive, contextually aware, and practically deployable hate speech detection systems that effectively serve the complex needs of modern digital communication platforms.

8. Conclusions

In this study, we evaluated the effectiveness of various state-of-the-art LLMs for hate speech classification across multiple datasets. Our findings demonstrate that while traditional hate speech classifiers have historically faced challenges due to their limited understanding of context, LLMs show significant promise in addressing these limitations. Specifically, models such as Meta-Llama-3-8B exhibited strong performance, particularly on datasets where a nuanced understanding of language and context is crucial. In addition, our findings highlight the potential of LLMs in improving the accuracy, precision, and recall of hate speech detection, particularly compared to traditional models. However, the varying performance among LLMs also signals that there is no one-size-fits-all solution, and further fine-tuning and model selection are required depending on the use case.

Author Contributions: Conceptualization, B.B. and S.J.; methodology, B.B. and S.J.; software, B.B.; validation, B.B.; formal analysis, B.B.; investigation, B.B.; resources, B.B.; data curation, B.B.; writing—original draft preparation, B.B.; writing—review and editing, B.B.; visualization, B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. 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, Toronto, ON, Canada, 15–19 October 2018; pp. 2–12.
2. Wang, C.C.; Day, M.Y.; Wu, C.L. Political Hate Speech Detection and Lexicon Building: A Study in Taiwan. *IEEE Access* **2022**, *10*, 44337–44346. [CrossRef]
3. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Republic and Canton of Geneva, CHE, Perth, Australia, 3–7 April 2017; WWW '17 Companion, pp. 759–760. [CrossRef]

4. Jaf, S.; Barakat, B. Empirical Evaluation of Public HateSpeech Datasets. *arXiv* **2024**, arXiv:2407.12018. [CrossRef]
5. Plaza-del Arco, F.M.; Nozza, D.; Hovy, D. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH), Toronto, ON, Canada, 13 July 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023.
6. Pan, R.; García-Díaz, J.A.; Valencia-García, R. Comparing Fine-Tuning, Zero and Few-Shot Strategies with Large Language Models in Hate Speech Detection in English. *CMES-Comput. Model. Eng. Sci.* **2024**, *140*, 2849–2868. [CrossRef]
7. Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; Von Werra, L.; Fourrier, C.; Habib, N.; et al. Zephyr: Direct distillation of lm alignment. *arXiv* **2023**, arXiv:2310.16944. [CrossRef]
8. Saha, P.; Agrawal, A.; Jana, A.; Biemann, C.; Mukherjee, A. On Zero-Shot Counterspeech Generation by LLMs. *arXiv* **2024**, arXiv:2403.14938. [CrossRef]
9. Nirmal, A.; Bhattacharjee, A.; Sheth, P.; Liu, H. Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales. *arXiv* **2024**, arXiv:2403.12403. [CrossRef]
10. Suryawanshi, S.; Chakravarthi, B.R.; Arcan, M.; Buitelaar, P. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 11–16 May 2020; pp. 32–41.
11. Salminen, J.; Almerekhi, H.; Milenković, M.; Jung, S.g.; An, J.; Kwak, H.; Jansen, B.J. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In Proceedings of the Twelfth International AAAI Conference on Web and Social Media, Palo Alto, CA, USA, 25–28 June 2018.
12. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. *Proc. Int. AAAI Conf. Web Soc. Media* **2017**, *11*, 512–515. [CrossRef]
13. De Gibert, O.; Perez, N.; García-Pablos, A.; Cuadros, M. Hate speech dataset from a white supremacy forum. *arXiv* **2018**, arXiv:1809.04444. [CrossRef]
14. 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, San Diego, CA, USA, 13–15 June 2016; pp. 88–93. [CrossRef]
15. Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; Wang, W.Y. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. *arXiv* **2019**, arXiv:1909.04251. [CrossRef]
16. Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; Tromble, R.; Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; et al. Introducing CAD: The contextual abuse dataset. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2289–2303.
17. Kennedy, C.J.; Bacon, G.; Sahn, A.; von Vacano, C. Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv* **2020**, arXiv:2009.10277. [CrossRef]
18. AI@Meta. Llama 3 Model Card 2024. Available online: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (accessed on 4 August 2025).
19. Microsoft. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219. [CrossRef]
20. Teknium; Theemozilla; Karan4d; Huemin_art. Nous Hermes 2 Mistral 7B DPO , 2024. Available online: https://huggingface.co/NousResearch/Nous-Hermes-2-Mistral-7B-DPO (accessed on 4 August 2025).
21. Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; Jiang, D. WizardLM: Empowering Large Language Models to Follow Complex Instructions. *arXiv* **2023**, arXiv:2304.12244. [CrossRef]
22. Song, Q.; Liao, P.; Zhao, W.; Wang, Y.; Hu, S.; Zhen, H.L.; Jiang, N.; Yuan, M. Harnessing On-Device Large Language Model: Empirical Results and Implications for AI PC. *arXiv* **2025**, arXiv:2505.15030. [CrossRef]
23. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2611–2624.
24. Carvallo, A.; Mendoza, M.; Fernandez, M.; Ojeda, M.; Guevara, L.; Varela, D.; Borquez, M.; Buzeta, N.; Ayala, F. Hate Explained: Evaluating NER-Enriched Text in Human and Machine Moderation of Hate Speech. In Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH), Vienna, Austria, 1 August 2025; pp. 458–467.
25. Tao, C.; Shen, T.; Gao, S.; Zhang, J.; Li, Z.; Tao, Z.; Ma, S. Llms are also effective embedding models: An in-depth overview. *arXiv* **2024**, arXiv:2412.12591. [CrossRef]
26. Lin, L.; Wang, L.; Guo, J.; Wong, K.F. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv* **2024**, arXiv:2403.14896. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.