From Rules to Reasoning: Empowering LLMs with Contrastive Learning for Effective Hate Speech Moderation

Anirudh Ravi Kumar, Avantika Singh, Arihant Banthia,
Sharanya Kumari Shivakumar, Vishnu Shetty Belanje

Group 36
Viterbi School of Engineering, Department of Computer Science, USC
{ar16847,singhava,abanthia,ss82937,shettybe}@usc.edu

Abstract

This research paper introduces a novel framework designed to enhance the effectiveness of Large Language Models (LLMs) in moderating hate speech online. While LLMs excel at understanding and generating human-like text, their ability to follow explicit rules through prompting for content moderation is hindered by the complexities of in-context retrieval within extensive exemplars. Our proposed solution employs a method of exemplar retrieval that specifically addresses these challenges, refining LLMs' ability to accurately identify and moderate hate speech. By enhancing the precision of hate speech detection, this approach aims to improve digital safety and foster a more inclusive online environment.

1 Introduction

In the rapidly evolving digital landscape, Large Language Models (LLMs) are increasingly pivotal in automating content moderation, particularly in detecting and mitigating hate speech. However, despite their sophisticated capabilities, LLMs encounter substantial hurdles. Primarily, these models often generate unsafe responses when confronted with hate speech due to their limited ability to apply logical reasoning beyond the confines of their training data. This fundamental limitation underscores a critical challenge: LLMs' struggle to accurately retrieve and apply the contextual nuances essential for interpreting complex language patterns typically associated with hate speech.

One of the significant impediments to effective moderation by LLMs is their inherent difficulty in managing the extensive sets of rules needed for accurate hate speech detection. This issue is compounded by the models' typically small context windows, which hinder their ability to maintain a comprehensive view of conversational threads. Consequently, this restricted perspective often leads to errors in judgment, where the context is either misinterpreted or completely overlooked, thereby affecting the overall performance of hate speech detection mechanisms.

Furthermore, the intricacies of integrating multiple rules within these limited context windows pose a substantial challenge. The need for an enhanced approach that can simplify rule retrieval and improve the efficiency of hate speech detection is therefore evident. Traditional methods that rely heavily on vast rule-based systems require reevaluation to foster more nuanced and reasoning-driven moderation processes.

To address these challenges, we propose an innovative architecture that leverages few-shot learning techniques to refine hate speech detection capabilities of LLMs. Our approach aims to minimize the dependency on extensive rule-based systems and instead enhance the models' ability to reason and contextualize with minimal supervision. By focusing on few-shot retrieval, the proposed architecture is designed to improve the precision of hate speech detection, thereby enhancing online safety and promoting inclusivity across digital platforms.

This paper will delve into the technical underpinnings of our proposed solution, demonstrating its potential to transform LLMs from merely rule-following entities to sophisticated reasoning machines. In doing so, we aspire to set a new standard for content moderation technologies, ensuring they are not only effective but also equitable and sensitive to the diverse nuances of human communication.

2 Related Work

The field of using Large Language Models (LLMs) for hate speech detection, as explored in recent literature, exhibits considerable advancements and concurrently unveils key areas for improvement. Studies like Probing LLMs for Hate Speech Detection (Roy et al., 2023) and LLMs for Real-World

Hate Speech Detection (Guo et al., 2024) have enhanced the capabilities of LLMs by integrating additional context, victim community information, and advanced prompting strategies. However, they reveal limitations in dynamically adapting to the evolving nature of hate speech, particularly in capturing new slurs, coded language, and implicit or nuanced forms of hate speech. This gap highlights the necessity for models that can adapt more fluidly to changes in language use and societal norms, and the need for more nuanced prompting strategies that can effectively handle the complexities of hate speech in various forms, including multimodal content.

In LLMs in Implicit Hate Speech Detection (Zhang et al., 2024) the focus on the challenges of detecting implicit hate speech brings to light the difficulties LLMs face in accurately identifying such speech without generating false positives. This underscores the need for advanced natural language understanding techniques that can better grasp context and subtext. Moreover, the Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection (Clarke et al., 2023) study introduces an innovative approach, yet it confronts challenges in maintaining the accuracy and relevance of rule-based systems as language and social norms evolve. These insights suggest that future research should concentrate on enhancing the scalability and adaptability of these systems, ensuring they remain effective as language use continues to change.

Large Language Models Can Learn Rules (Zhu et al., 2023) demonstrates the significant improvements in accuracy and reasoning capabilities of LLMs when equipped to learn and apply rules. Similarly, the paper Large Language Models as Analogical Reasoners (Yasunaga et al., 2024) presents a new prompting method called analogical prompting, which enhances the reasoning capabilities of large language models by allowing them to autonomously generate relevant examples tailored to each task.

However, the above works point to the challenges in in-context retrieval, indicating limitations in how LLMs integrate and apply learned rules in diverse scenarios. This revelation points to a critical need for improving the mechanisms through which LLMs retrieve and apply rules, especially in varied and complex contexts. Together, these studies outline a landscape of rapid advancement with substantial achievements, yet they also clearly delineate the need for continuous innovation and adaptation in LLMs to effectively tackle the multifaceted and dynamic challenges of hate speech detection.

3 Methodology

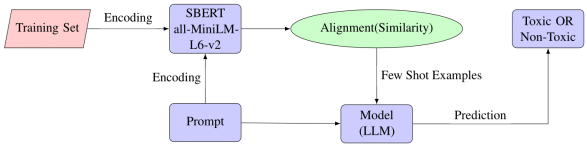

Figure 1: Flowchart of our proposed methodology - Few Shot Retrieval based on Similarity

Our research is focused on exploring how well large language models (LLMs) can identify hate speech using case-based reasoning. Figure 1 gives an overview of our approach. Our proposed architecture consists of the following components:

3.1 Training Set and Encoding

The foundation of our model's training begins with collecting a diverse training set that includes instances of both toxic and non-toxic content. This training set serves as the foundation for the LLMs to draw upon during the prediction process. Utilizing the SBERT (all-MiniLM-L6-v2), a derivative of the BERT model optimized for sentence embeddings, we encode the training set to transform natural language into vector representations.

3.2 Similarity Alignment

Post encoding, our methodology employs an alignment process to evaluate the similarity between the training data and the prompts. This step is crucial for ensuring that the LLMs can discern the relevancy between given examples and the prediction tasks. The similarity alignment not only streamlines the decision-making process but also refines the prompt's context to be in sync with the trained model's understanding.

3.3 Prompting and Few-Shot Examples

The prompts act as a gateway through which encoded inputs are fed into the model, steering the LLM towards the prediction task. Our methodology capitalizes on few-shot examples to fine-tune the model's decision-making process.

3.4 Model Prediction

At the end of the alignment and few-shot learning process, the LLM is tasked with predicting whether the content is toxic or non-toxic. This predictive output encapsulates the essence of our methodology, highlighting the model's capability to discern content based on its toxicity levels effectively.

4 Experiments

In this section, we discuss the experiments conducted to investigate the ability of LLMs to perform case based reasoning in Hate speech detection.

4.1 Experimental Design

Our research is methodically organized into a series of experiments to evaluate the efficacy of Large Language Models (LLMs) in identifying hate speech, highlighting the introduction of a novel framework aimed at augmenting their detection capabilities. (Please refer to Appendix A for the prompts.)

Experiment 1: Zero-Shot Reasoning. This experiment assesses the baseline capability of LLMs to tackle hate speech detection without any tailored instruction, relying solely on their pre-trained knowledge base. We plan to reveal the limitations of LLMs in addressing complex reasoning tasks like hate speech detection in a zero-shot learning context.

Experiment 2: Few-Shot CoT Reasoning. This experiment contrasts with the zero-shot learning model by introducing a few-shot learning strategy. A predefined collection of reasoning exemplars is provided to LLMs for each test case within a dataset. The purpose is to evaluate LLMs' capacity for leveraging specific examples during their decision-making process. Our goal is to highlight how our proposed method outperforms traditional few-shot CoT reasoning by offering more precise guidance, thereby enhancing the accuracy of LLMs in identifying hate speech.

Experiment 3: Analogical Prompting This prompting strategy (Yasunaga et al., 2024) is similar to zero-shot prompting and fewshot prompting. In this method, the LLM itself is asked to generate relevant example similar to the given prompt based on its pretrained knowledge. Using the generated examples the model makes its predictions. The goal of this experiment is to evaluate the ability of model to retrieve relevant exemplars and make the prediction to the given prompt.

Experiment 4: Few Shot With Retrieval Prompting Our proposed method focuses on retrieving highly relevant and similar exemplars based on the query at hand. Utilizing a dual encoder architecture for efficient exemplar matching, this strategy involves enhancing LLM prompts with pertinent examples from the training set, thus providing a scaffold for the models to draw upon external knowledge in responding to hate speech queries. The objective of this benchmark is to evaluate the improved precision of LLMs in detecting and moderating hate speech when augmented with our retrieval-based approach, positioning it as a scalable solution in the ongoing effort to refine content moderation systems.

4.2 Models

To evaluate the effectiveness of our approach in identifying and classifying hate speech, we employed three state-of-the-art language models.

Phi-1.5(Li et al., 2023) with 1.5 billion parameters, is adept at capturing complex language patterns and nuances, making it suitable for tasks requiring fine-grained language understanding, such as hate speech detection. **Phi-2** extends the capabilities of Phi-1_5 by incorporating an enhanced training dataset and doubling the number of parameters to 3 billion containing attributes that are crucial for hate speech detection. **Mistral-7B** (Jiang et al., 2023) represents a significant leap in model complexity and understanding, featuring 7 billion parameters.

4.3 Datasets

In our study, we adapt three datasets—**HateXplain** (Mathew et al., 2020), **Jigsaw** (Clarke et al., 2023), and **CAD** (Vidgen et al., 2021)—to a binary classification of toxic and non-toxic content. HateXplain merges "hateful" and "offensive" labels into a toxic category, with dataset splits of 8,000/1,000/1,000 (toxic) and 6,000/781/782 (non-toxic). Jigsaw classifies "identity hate" as toxic, resulting in 1,405/100/712 (toxic) and 158,000/1,000/63,000 (non-toxic) splits. Similarly, CAD's "identity-directed" entries are deemed toxic, with splits of 1,353/513/428 (toxic) and 12,000/4,000/4,000 (non-toxic). These adaptations allow for focused binary classification analysis across diverse online contexts.

The training set functions as an exemplar database to guide the model with similar instances for predictions, while the test set measures the

		Zero Shot		Few Shot		Few Shot Retrieval		Analogical Prompting	
Dataset	**Model**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
HateXplain	Phi-1_5	0.47	0.64	0.49	0.64	0.58	**0.69**	0.28	0.42
	Phi-2	0.48	0.43	0.39	0.40	0.52	**0.56**	0.51	0.29
	Mistral-7B	0.49	0.63	0.51	0.67	0.53	**0.67**	0.57	0.58
Jigsaw	Phi-1_5	0.47	0.64	0.46	0.62	0.49	**0.65**	0.27	0.40
	Phi-2	0.49	0.48	0.42	0.48	0.53	**0.59**	0.59	0.43
	Mistral-7B	0.57	0.69	0.55	0.68	0.55	**0.71**	0.61	0.64
CAD	Phi-1_5	0.44	0.61	0.46	0.60	0.58	**0.66**	0.23	0.33
	Phi-2	0.47	0.41	0.34	0.33	0.49	**0.49**	0.50	0.20
	Mistral-7B	0.55	0.66	0.52	**0.67**	0.49	0.62	0.55	0.52

Table 1: Performance of the models on Hatespeech datasets - HateXplain, Jigsaw and CAD using various methodologies

model's precision in separating toxic from non-toxic messages.

4.4 Evaluation metrics

In this study, two key evaluation metrics are given precedence: **Accuracy** and **F1 Score**. These metrics collectively offer a comprehensive assessment of our model's predictive accuracy and robustness.

5 Results

The evaluation of various models on the HateXplain, Jigsaw, and CAD datasets underscores the robust performance of Few-Shot Retrieval methodology and the inherent strength of the models in accurately classifying content across different contexts. The findings are summarized in Table 1. Our proposed Few-Shot Retrieval method consistently emerged as the top-performing methodology. Particularly on the Jigsaw dataset, Mistral-7B attained an impressive F1 score of 71%, demonstrating the potential of advanced models to adapt and perform with similar examples as input.

Zero-Shot prompting showcased the models' remarkable innate capability for content interpretation. Despite Phi-2's larger parameter count compared to Phi-1.5, its performance doesn't fully meet expectations, due to its lack of exposure to hate speech during training. Nevertheless, Phi-2's few-shot retrieval scores—an F1 of 59.23% on Jigsaw—are impressively equivalent to Phi-1.5's zero-shot outcomes, demonstrating Phi-2's remarkable adaptability.

Across all datasets, the Mistral-7B model demonstrated superior F1 scores, highlighting its effectiveness in understanding and classifying complex content. The model's large-scale architecture can be attributed to its high performance, reflecting the advantages of scale in language moderation and hate speech detection.

The experimental results highlight that the few-shot retrieval approach and the inherent capabilities of LLMs like Mistral-7B and Phi variants offer promising avenues for enhancing hate speech detection.

Conclusion

We have introduced an innovative few-shot retrieval architecture that substantially elevates the performance of large language models (LLMs) for hate speech detection. By selectively employing exemplars pertinent to the prompts, our approach adeptly mitigates the challenges associated with in-context learning, such as the constraints of context window size and the complexities involved in handling diverse rules within LLMs.

Looking forward, we are pragmatic about the application of retrieval based mechanism combined with analogical prompting to domains such code generation, code refactoring and mathematical problem solving.

6 Individual Contributions

The project was a joint effort, with all members contributing equally. Avantika worked on evaluation script used to determine the metrics of the given methodology and prompting methodologies. Sharanya worked on different prompting methodlogies, research of papers and developing the retrieval script for Few Shot with retrieval. Vishnu worked on prompting and developing the retrieval script

for Few Shot with retrieval. Anirudh worked on the model inference script and running the experiments. Arihant worked on dataset preprocessing and running the experiments. The entire team participated in report creation.

References

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376, Toronto, Canada. Association for Computational Linguistics.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. 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*, pages 2289–2303, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023. Large language models can learn rules.

A Hate Speech Classification Prompts

Model Prompt:
Determine if the provided sentence constitutes hate speech. Hate speech is defined as any communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group based on attributes such as their religion, ethnicity, nationality, race, color, descent, gender, or other identity factors. Classify the sentence as "Toxic" if it constitutes hate speech, or "Non-Toxic" if it does not.
Input: *and this is why i end up with ****** trainee doctors who can not speak properly lack basic knowledge of biology it truly scary if the public only knew.*
Response: toxic

Figure 2: Zero Shot Prompting

Model Prompt:
Based on the examples given, determine if the provided sentence constitutes hate speech. Hate speech is defined as any communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group based on attributes such as their religion, ethnicity, nationality, race, color, descent, gender, or other identity factors. Classify the sentence as "Toxic" if it constitutes hate speech, or "Non-Toxic" if it does not:
Examples
- "Agree but races reversed. Africans are literally dependent on white everywhere on earth." -> Toxic
- "Black Margaret Sanger, is that you" -> Non-Toxic
- "my cat is more intelligent and capable of feeling human emotions than a nigger" -> Toxic
- "yeah it a sad for in america when white supremacists start a violence protest" -> Non-Toxic
- "Your country sucks" -> Toxic
- "Please stop. If you continue to vandalize Wikipedia, as you did to Homosexuality, you will be blocked from editing." -> Non-Toxic
Please use the above knowledge to classify the below sentence as "Toxic" or "Non-Toxic". Do not hallucinate.
Input: *and this is why i end up with ****** trainee doctors who can not speak properly lack basic knowledge of biology it truly scary if the public only knew.*
Response: toxic

Figure 3: Few-Shot COT Prompting

Model Prompt:
Given the task of identifying hate speech, where hate speech is defined as any communication that attacks or uses pejorative or discriminatory language with reference to a person or a group based on attributes such as their religion, ethnicity, nationality, race, color, descent, gender, or other identity factors:
1. Recall relevant examples or knowledge about hate speech. Consider different contexts in which hate speech may occur and how it can be identified.
2. Using the insights from step 1, analyze the following sentence and determine if it constitutes hate speech. Consider the language used, the context, and the intent behind the sentence.
3. Classify the sentence as "Toxic" if it constitutes hate speech, or "Non-Toxic" if it does not.
Input: *and this is why i end up with ****** trainee doctors who can not speak properly lack basic knowledge of biology it truly scary if the public only knew.*
Response: toxic

Figure 4: Analogical Prompting

Model Prompt:
Based on the examples given, determine if the provided sentence constitutes hate speech. Hate speech is defined as any communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group based on attributes such as their religion, ethnicity, nationality, race, color, descent, gender, or other identity factors. Classify the sentence as "Toxic" if it constitutes hate speech, or "Non-Toxic" if it does not:
Examples
- Muh doctors and engineers bullshit. -> Toxic, Similarity Score: 0.44
- Isnt the M in STEM meant to be M*****. -> Toxic, Similarity Score: 0.44
- Text: Same here it seems cringey when they do it In my mind Male and female are similar to **** and ***** Slightly more scientific wording. -> Toxic, Similarity Score: 0.35
- Text: But cant be a bachelors degree So itll be a spinsters of science. -> Toxic, Similarity Score: 0.35
- Text: It always makes me laugh whenever someone complains a scifi is getting political. -> Toxic, Similarity Score: 0.33
Please use the above knowledge to classify the below sentence as "Toxic" or "Non-Toxic". Do not hallucinate.
Input: *and this is why i end up with ****** trainee doctors who can not speak properly lack basic knowledge of biology it truly scary if the public only knew.*
Response: toxic

Figure 5: Few Shot with Retrieval Prompting