Beyond Words: Harnessing Large Language Models for Detecting Implicit Hate Speech

Sanjeevan Selvaganapathy

May 2024

Contents

1 Introduction

In the growing digital age, online spaces that allow freedom of expression have the power to significantly influence our culture, society, and politics - more generally, online spaces are increasingly influencing the greater world we inhabit. Many of these online platforms act as arenas for public discourse, thanks to their anonymity and ease of access, and regularly host discussions that can reach staggering amounts of people. This is very much a double-edged sword - while there are incredible benefits, one downside of a system like this is that the freedom of expression of many of these types of platforms naturally brings about an increase in the proliferation of **hate speech**.

Hate speech mitigation and content moderation are incredibly important to maintaining a respectful, safe, and positive digital environment, ensuring greater social harmony and individual well-being. While mitigating all forms of online hate speech should be the goal, particularly troubling is implicit hate speech - subtle and nuanced texts that indirectly convey harmful, degrading, or biased sentiment against groups based on race, gender, religion, or other characteristics. Detecting and moderating implicit hate speech is notably harder than other types of hate speech due to its subtlety, and the contextual knowledge required to interpret it.

As the size of online content being moderated grows, manual or human-based approaches in moderating online spaces become increasingly infeasible, and so we look to automated methods of hate speech mitigation. Previous strategies in classifying and moderating hate speech have utilized traditional machine learning techniques such as Support Vector Machines and Logistic Regression models [1] but all have returned mixed results. Notably, the issue in many of these early models was that they struggled with implicit hate speech detection, and could not inherently understand the content and the context of the text being classified. Because these systems only filtered out the most 'obvious' forms of hate speech, they could not be trusted to fully moderate any online space. As such, any future automated solutions to this problem must be able to tackle this implicit hate speech classification task effectively, especially when total and well-rounded hate speech classification is the goal.

Recently, the advent of Large Language Models (LLMs), a subclass of language models that seem to demonstrate an inherent understanding of texts and their contexts far beyond anything we have seen before, has meant that the hate speech classification task now has a new avenue to explore. Already, we have seen that hate speech classification performance using LLMs boasts significant improvements over traditional methods [2]. However, with recent advances in LLMs, such as improvements in our knowledge of prompting techniques, and 'smarter' or alternative models being released, we can now revisit the problem of hate speech classification, with a focus on classifying both implicit and explicit forms.

This literature review is designed to cover the hate speech classification task using LLMs, focusing on implicit hate speech detection, as this is the most difficult type of hate speech to classify using machine-based solutions. We will look into pertinent areas of the greater LLM landscape, such as the performance of newer models compared to alternative or 'uncensored' models, as well as current best practices in prompting techniques applicable to our classification tasks. Finally, we will review current gaps in the literature about these topics, and outline the contributions we aim to make with the proposed project.

NOTE: This literature review contains examples of language that may be offensive to some readers. They do not represent the views of the authors.

2 Literature Review

2.1 Defining Hate Speech

Hate speech is broadly defined as "any kind of communication in speech, writing, or behavior, that attacks or uses pejorative or discriminatory language concerning a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." [3]. Digital hate speech can be shared freely, or at low cost; it can reach staggering amounts of people in fractions of a second, and once created, can be extremely difficult to remove [3]. Crucially, the anonymity provided by online platforms makes it incredibly easy for users to share controversial opinions, which can manifest as toxic or negative speech if not moderated appropriately.

Hate speech can be further broken up into **implicit** and **explicit** hate speech. We define implicit hate speech as "that hate speech which employs indirect language to convey hateful intentions" [4]. An example of implicit vs explicit hate speech can be found in Table 1.

Table 1: Examples of Explicit and Implicit Hate Speech

Type	**Example**
Explicit Hate Speech	"All [ethnic groups] are criminals and should be expelled from our country."
	"People from [religion] are terrorists and don't deserve to live here."
Implicit Hate Speech	"It's just a fact that certain neighborhoods are more dangerous, you know which ones I mean."
	"These people are just not as hardworking as us, which is why they always need help."

Implicit hate speech often is much harder to classify as models must understand many inherent features of the text that transcend just the words being used. These may be, but are not limited to, features such as linguistic nuance and diversity of the implicit hate class, sarcasm, irony, humor, euphemisms, circumlocution, or other symbolic or metaphorical language [5]. Similarly, the type and goal of implicit hate speech can vary greatly, from reinforcing harmful stereotypes to threats or incitement of violence [5]. Finally, hate speech can evolve - new phrases and words will be slowly integrated into a language's vocabulary and may come to be understood as hate speech, but it is difficult, especially for an automated system, to quantify when and how this process occurs. Furthermore, any successful model must be able to adapt to the language it is classifying.

All of these features of implicit hate speech make performant classification an extremely difficult task.

2.1.1 Historical Approaches to Hate Speech Moderation Online

Previous attempts to mitigate and moderate hate speech online have seen the use of both traditional Machine Learning(ML) techniques and other types of language models.

Early efforts into applying ML techniques first focused on separating hate speech from simple offensive language - language containing profanity, but which does not fit the definition of hate speech. A 2017 paper by Davidson, Warmsley, Macy, *et al.* [1] was one of the first to train classifiers to distinguish between texts containing hate speech, and offensive language, and those with neither. The model performed well in classifying offensive and neutral language but distinctly saw almost 40% of hate speech be misclassified. The authors noted that one of the most frequently misclassified forms of hate speech was implicit hate speech.

Before LLMs, hate speech moderation was also carried out by use of Pretrained Language Models (PLMs) such as BERT and RoBERTa [2]. PLMs can be thought of as smaller-scale LLMs - they are usually trained on less data, have fewer parameters, and are fine-tuned on specific datasets to achieve specific task performance. These models served as more general language encoders and could be fine-tuned with task-specific datasets to detect hate speech.

Efforts were also made to customize models for specific languages or contexts. As we have established, hate speech is a complex and multifaceted problem, and so more specific classifiers were fine-tuned from these base PLMs for hate speech classification. Models such as UmBERTo (Italian BERT) and BETO (Spanish BERT) were applied to hate speech detection in a multilingual setting. These approaches, while effective, ultimately were heavily reliant on the language encodings of the training data. Without proper fine-tuning, and even sometimes simply due to a lack of data, these approaches sometimes led to issues with overfitting, and even with fine-tuning could not classify a wide range of hate speech, necessitating a shift towards more adaptable and robust solutions. For example, many low-resource BERT fine-tunings referenced in Kumarage, Bhattacharjee, and Garland [2] contained datasets that were skewed towards strong sentiment about political parties and did not contain texts on topics such as gender bias.

Many earlier studies incorrectly conflated offensive language and hate speech into the same labels - perhaps because one is much easier to detect than the other. Understandably, explicit hate speech and offensive language are much easier to classify based on identifying specific words or phrases that are most commonly used in a hateful manner. Yet, as we have seen, previous studies [4] [6] [1] have consistently brought out a discrepancy in automated systems' classification performance between implicit and explicit content. As such, the problem of separating between hate speech and offensive language, and then furthermore, between implicit hate speech and explicit hate speech, is a relatively recent distinction, and one comparatively less seen in the literature.

2.2 Large Language Models (LLMs)

Large Language Models (LLMs) are at the forefront of current Artificial Intelligence and Machine Learning research. They are particularly pertinent to our topic of hate speech classification, due to their uncanny ability to understand text.

These language models, while each slightly different, typically involve an artificial neural network architecture with millions or billions of parameters. The paper 'Attention Is All You Need', by Vaswani, Shazeer, Parmar, *et al.* [7], was the first to introduce a novel 'Transformer'-based language model, a model which showed superior quality machine translation tasks while requiring a fraction of the training costs compared to other, at the time, state-of-the-art models.

We have now reached a point where these transformer-based architectures, along with other similarly performant models, have evolved and progressed such that they are capable of parsing, understanding, and generating text at levels that some posit could even reasonably be viewed as an early, albeit incomplete, version of an artificial general intelligence (AGI) system [8].

We treat LLMs, for our purposes, essentially as 'black boxes' capable of generating input-output pairs, where the output is a function of the input. This is because the inner workings of many commercially available LLMs are not publicly available, but access to the model is available through an API or similar.

While there is an open debate about the ethical concerns of researching closed models [9], in our case, we are specifically tackling the problem of hate classification, and not a scientific contribution to machine learning theory or methods - this is a problem of 'engineering' rather than just 'machine learning'. As such, rather than delve into the intricacies of each model, we will persist in our 'black-box' conceptualization, and study the generalisability of different techniques to multiple models, both open and closed.

2.2.1 Prompting

Inputs to an LLM are known as 'prompts'. Prompts are user-inputted questions or statements that trigger a response. Recently, the importance of proper 'prompting' techniques has come to light, with research demonstrating [10] [11] [12] [13] that the quality of a prompt greatly influences the quality of a response.

2.2.2 Fine-Tuning

LLMs can also be 'fine-tuned'. Fine-tuning is akin to slightly tweaking the parameters of an LLM such that it can adopt a specific behavior. This is often done by feeding it a task-specific dataset and letting it update its weights such that it adapts its understanding capabilities to the input dataset, rather than relying on the generalized base model. Fine-tuning models to achieve a specific task is often computationally cheaper than training a new model to do so, as it can leverage the base performance of high-quality pre-trained models.

2.2.3 RLHF and LLM Guardrails

When interacting with potentially harmful, violent, or otherwise inappropriate content, many proprietary or commercial LLMs include 'guardrails', which are designed to prevent users from interacting with models in ways that may enable bad actors. These guardrails can be implemented explicitly through, for example, Input/Output guards which monitor the input and output of a model directly, acting as a sort of 'filter'. Guardrails can also be implemented implicitly through Reinforcement Learning through Human Feedback (RLHF), a process similar to fine-tuning in which the behavior of an LLM can be altered for a specific sort of behavior - for example, outputting responses that seem more 'human' and less robotic. RLHF essentially is a training method of updating the weights of a model where human feedback is used as a loss function on the generated outputs, rather than computing it. Compared to fine-tuning, RLHF aims to achieve a difference in the 'style' of responses, rather than the content. 'Human Feedback' is used here to tell the model what sort of responses are rewarded and encouraged, and what sort of responses should be avoided. In the same way that we can augment a model to produce more natural-sounding language, we can also try to steer a model away from generating or interacting with harmful, offensive, or inflammatory content.

When models are put out to the public, they often undergo RLHF, and other guardrailing techniques such as fine-tuning, to make them more accessible and produce prompts that humans prefer, or mitigate the output of harmful content. However, studies have shown [14] [15] that an increase in RLHF brings about an 'alignment tax'; that is, a reduction in performance on industry LLM performance benchmarks that measure general capabilities across metrics such as logical reasoning, textual understanding, and creativity.

When concerning public-facing models, this no doubt is an important tradeoff that must be heavily considered, as the liability is purely on companies should their products be used directly for nefarious or harmful purposes. However, it begs the question of whether or not, when we have a specific task and are not so concerned with alignment issues, uncensored models, or models with less 'alignment', can perform better than other models. Open and 'uncensored' models such as Wizard Vicuna and Lexi have been shown to nevertheless be strong performers on industry LLM benchmarks [16]. To date, no literature can be found that utilizes publicly available uncensored models on implicit hate-speech classification tasks.

2.2.4 Emergent Reasoning Abilities of LLMs

A pivotal study done by Bubeck, Chandrasekaran, Eldan, *et al.* [8] benchmarked GPT-4, one of the most popular and advanced publicly available LLMs, demonstrated its exemplary performance across a variety of domains and tasks. Crucially, the authors posit that GPT-4, and other similarly performant models, could be viewed as an early and incomplete version of Artificial General Intelligence - an AI that can perform at or above a human level on a wide range of tasks that span mathematics, coding, psychology and more, without needing dedicated prompting techniques.

Perhaps reflective of the vast and expansive corpus of training data, the study brings to light GPT-4's ability to understand and generate text that captures a wide range of human-like nuances, subtleties, and contexts. This sort of capability could be crucial for identifying implicit hate speech, which often requires an underlying contextual understanding - somewhere where previous approaches to the task have faltered.

2.2.5 Ethical Considerations and Concerns Regarding LLM Use

Using LLMs in deployed and automated systems carries serious ethical implications, especially as they are systems that are not yet widely understood, or even explainable. Here we explain key ethical considerations and concerns of using LLMs which must be addressed or considered in any research project that uses them.

First and foremost, LLMs are trained on datasets that inevitably contain biases present in the source material. These biases can perpetuate, or even exacerbate stereotypes and discriminatory practices when the models are used in real-world applications. For a task such as hate speech moderation, we must ensure that models are properly calibrated to respond appropriately, especially when dealing with sensitive topics - in fact, we can already see that current models do not classify implicit hate speech accurately when dealing with groups which may cause fairness issues [4].

Once again, the 'black-box' nature of many LLMs poses significant challenges to the accountability, transparency, and reliability of automated systems. Approaches that utilize LLMs through API access, for example, cannot 'reliably' be replicated as the models that are being called under the hood may be changed at any time, without knowledge, at the discrepancy of the providing entity. Understanding exactly how decisions are made, especially in the context of moderating human language and hate speech, is essential for trust and accountability. Without this, it becomes challenging to diagnose errors, address unintended biases, and ensure that the model or content moderation system acts by ethical guidelines.

Finally, while newer models outperform older models, oftentimes substantially [2], we must be mindful of the cost of using newer models. At present, 'better' models require more computing, which can translate to higher costs. Any system deployed to production must maintain a reasonable balance between performance and cost to ensure feasibility.

It is important to be cognizant of these issues when conducting LLM-related research to ensure academic and moral integrity.

2.3 Previous Uses of LLMs in Hate Speech Moderation

A detailed review of the application of LLMs in hate speech moderation was done by Kumarage, Bhattacharjee, and Garland [2] and provided a background and efficacy review of LLMs in classifying hate speech. The review concluded with a few crucial results. The authors found that newer LLMs were much more effective than earlier models at classifying hate speech, even without prompting techniques. Naturally, newer language models, which may have more parameters, be trained on more data, or otherwise be 'improved' in some other way, may prove to perform better on classification tasks. Interestingly, 'base' newer models such as GPT-4 even outperform fine-tuned older models for specific English-language classification tasks, although performance across other languages is dubious.

A mentioned study by Huang, Kwak, and An [17] is one of the only available studies into the LLMs in detecting implicit hate speech and generating and providing plausible reasoning for its decisions. The tested ChatGPT model performed comparably to human Amazon Mechanical Turk workers in classifying implicit hate speech and even generated explanations for their decisions which appeared to have more clarity than human-written explanations. This study found an 80% recall rate when classifying implicit hate speech examples, albeit using a simple prompt without further prompt enhancement techniques. Furthermore, since the publication of this study, numerous commercial and open-source models have been released that share similarities with ChatGPT, or other LLMs present in the literature, and whose benchmark performances on both classification and reasoning tasks (among others) surpass ChatGPT. There is a noticeable gap in the literature evaluating these models and their performance in this implicit hate speech classification task.

While LLMs show marked performance improvements over other classification techniques in the implicit hate speech detection task, weaknesses in the strategy still exist. Notably, many existing studies into LLMs still do not distinguish between implicit and explicit hate speech, and those that do often note that LLMs struggle with implicit hate speech, or detecting sub-contextual sentiment that might appear more obvious to a human annotator. We now look into several explanations for this deficiency.

2.3.1 Sensitive Topic Classification and LLM Guardrails

LLM Guardrails are the security practices and safety controls, whether pretrained into the model, or as part of the input/output pipeline, which help ensure that LLM use conforms to a standard - typically one that prevents users from generating hurtful, harmful, or offensive content, or prevents the model from exposing private data, etc. Because guardrails help mitigate the production of offensive content, this introduces a ‘bias’ into their outputs, which consequently can interfere with tasks where a more ‘objective’ perspective is needed, such as implicit hate speech classification.

A study done by Zhang, He, Ji, *et al.* [4] is one of the only studies that delves specifically into the capability of numerous state-of-the-art LLMs to detect implicit hate speech along with prompt engineering techniques. The authors compared many industry-leading models, such as LLaMA-2, Mixtral-8x7b, and GPT-3.5-Turbo across an English-language dataset. The study found that, generally, all three selected LLMs displayed an extreme sensitivity to topics that may cause fairness issues, highlighting a fault in their use for classifying hate speech.

Similarly, the evaluatory study by Kumarage, Bhattacharjee, and Garland [2] once again showed that many models, across different studies, proved ineffective at understanding the nuances of hate speech past categorizing obvious and explicit examples. Specifically outlined was a weakness in GPT4’s ability to understand or classify speech relating to women, which may be indicative of its guardrail’s influence on output quality.

Both these studies evaluated commercial LLMs which have been released to the public and fine-tuned specifically to conform to appropriate good standards and ethics policies to prevent misuse, being ‘wrongly’ censored, or presenting bias that hurts what can be considered as the moral and good-natured task of classifying hate speech. A problem with any hate speech classification is that we must not only classify hate speech towards certain demographics and not others, and so any hate speech classification system must not be biased to prevent hate speech against only targets that are not already historically marginalized. A novel avenue of exploration that the project proposes will be to use open and uncensored models as possible classifiers, with the hope that their lack of or lesser moral guardrails may prove to provide a more accurate and ‘unbiased’ perspective, especially on implicit hate speech classification.

2.3.2 Classifying Hate Speech as a Binary

Because the definition of hate speech, both implicit and explicit, involves a great deal of subjective evaluation, classifying hate speech along a binary can be problematic. Hateful speech is much deeper than simply being hateful or not; the degree of hate, the demographic being targeted, and the method of delivery of said hate speech - these dimensions, along with many others, are all lost by simply classifying hate speech in a binary manner. Despite this, many hate speech classification studies employ binary modeling.

A study done by Ayele, Jalew, Ali, *et al.* [18] brought new findings to light - that, indeed, improving hate speech classification performance may perhaps best be tackled by classifying it along a range of dimensions rather than along a binary of yes or no. The study classified hate speech on a scale of different values such as ‘target’ and ’intensity level’, and built multi-class classifiers to predict these values as well as a simple logistic regression model. Indeed, those multi-class predictor models in the paper demonstrated superior classification performance when classifying hate speech.

Despite implicit hate speech modeling perhaps being viewed as a multi-dimensional problem, with many possible features to predict, there is a gap in the literature regarding predicting specific features. One noteworthy study done by Jafari and Allan [19] on target-span detection - that is, identifying the targeted groups in a given piece of content - demonstrated promising results on tagging groups specifically to mitigate implicit harmful content. However, there has not yet been a use for these findings. Further research may include target-span detection (or other suitable features that can be predicted accurately) when modeling the hate-speech classification problem as multi-dimensional.

2.4 Prompting Strategies and Other LLM Performance Improvement Techniques

We have so far established that the implicit hate speech classification task requires models that can demonstrate adequate reasoning and understanding of texts that contain indirect hateful sentiment. We have established that LLMs present a novel way of approaching this task - newer LLMs have not been applied to the implicit hate speech classification task despite demonstrating superior reasoning and understanding ability on benchmark tests, and we now know more about the implicit hate speech classification problem to also think about solving it in novel ways - trialing uncensored models as a means to avoid sensitivity issues or modeling the problem as multi-dimensional rather than as a binary. We now will touch on one of the most crucial aspects of LLMs - 'prompting'. As a reminder, a 'prompt' is simply the input or inputs provided to a model, intended to trigger a suitable response.

Generally speaking, the quality of a prompt directly affects the quality of the output of an LLM. Because the parameters of LLMs number so greatly, it is impossible to conclude observed patterns directly; we can only infer conclusions based on experiments conducted on the 'black box' of the model itself. As such, 'prompting' strategies become critical to improving the performance of LLMs on whatever task they are currently being applied to; in our case, hate speech classification. We now discuss recent trends in 'prompting' that have been shown to improve the quality of results. Time permitting, one or more of these strategies may be tested in our final project to determine if performance on our classification task improves.

Current research, while shallow, seems to be indicative that basic prompting techniques do not easily extend to a performance improvement in the implicit hate speech classification task. Notably, studies such as Zhang, He, Ji, *et al.* [4] seem to indicate that no one prompting technique regularly outperformed the others across models.

However, one thing to note is that the efficacy of prompting techniques seems to scale with the size and performance of the base model [20]. Previous studies which found the performance benefits of different prompting techniques to be inconclusive [4] were tested on older and less performant models than what is available publicly now (e.g. GPT-3.5-Turbo vs GPT-4, LLaMa 2 vs LLaMa 3). As such, it may be worth revisiting these in future research, especially when prompting techniques show such promise across other domains [13] [10].

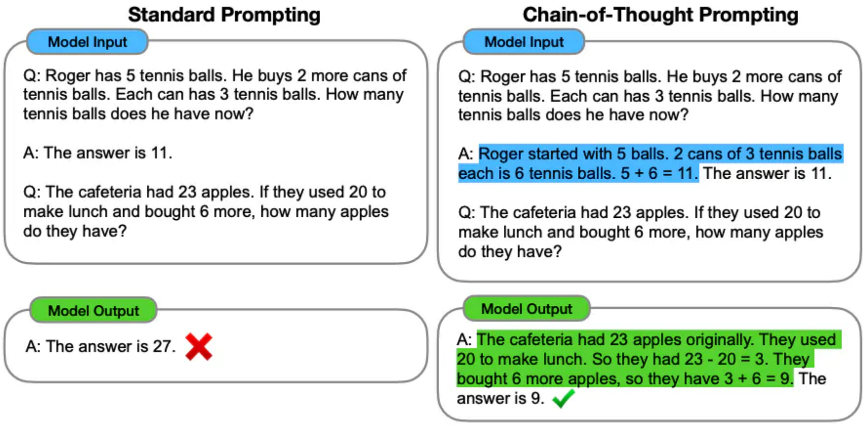
We now go over some common and pertinent strategies that may be of use in our hate speech classification study.

2.4.1 Zero Shot and n-shot Prompting

Table 2: Zero-shot vs 1-shot Prompting

	Zero-Shot	**1-Shot**
Input	1. Write a concise description of the novel ‘Pride and Prejudice’ by Jane Austen.	1. “To Kill a Mockingbird” is a novel by Harper Lee about a young girl named Scout Finch who lives in the racially charged atmosphere of the South in the 1930s. It deals with serious issues like racial inequality and injustice through the eyes of Scout. 2. Write a concise description of the novel ‘Pride and Prejudice’ by Jane Austen.
Output	“Pride and Prejudice” is a novel by Jane Austen that explores the romantic entanglements and societal pressures faced by the Bennet sisters.	“Pride and Prejudice” is a seminal work by Jane Austen that details the dynamics of love, marriage, and class. Through sharp wit and profound social commentary, the novel follows Elizabeth Bennet as she navigates societal expectations and her relationship with Mr. Darcy.

Many LLMs are trained on such large amounts of data that often they can perform tasks, or generate the correct desired outputs simply by asking the model to perform the task without any additional examples. This is known as ‘zero-shot’ prompting. Many commercial LLMs have been optimized for



Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Figure 1: An example of Chain of Thought prompting. Taken from Wei, Wang, Schuurmans, *et al.* [10]

zero-shot performance, as they are designed to be accessible and user-friendly, mirroring an all-knowing, helpful assistant, rather than a robot whose parameters must be tweaked to generate the most optimal results. Previous studies [2] have demonstrated that current state-of-the-art models prompted in a zero-shot manner outperform fine-tuned or optimized older models in classification and reasoning benchmarks.

By providing the models with n examples, we can coin the term '*n-shot* prompting', or similarly, '*few-shot prompting*', which describes prompting a model with a few examples. Few-shot prompting is an example of 'in-context learning', where we can directly influence the output of a model by providing it with direct context to learn from, rather than relying on its pre-trained knowledge base. In Table 2, we see an example of a zero-shot vs a 1-shot prompt. These results were generated with GPT-4. We can see that even by attaching a singular example, we can 'guide' the model into producing a specific type of output.

Zero-shot and few-shot prompting will be useful in providing a solid baseline for assessing the hate-speech classification performance of a model. Zero-shot prompting will allow a very general evaluation of a model's abilities, while few-shot prompting will allow models to categorize implicit hate speech based on a few related examples. This is especially useful when we have labeled examples that are similar to the text being classified - we can conclude the model, rather than relying on it to reason by itself, and thus 'guide' it towards a more correct answer.

2.4.2 Chain of Thought Prompting

Chain of Thought (CoT) prompting, introduced by Wei, Wang, Schuurmans, *et al.* [10], involves a prompting technique of adding, or supplementing a model with articulated reasoning. In essence, telling a model how it should be thinking. This style of prompting has been shown, in the paper, to enable complex reasoning capabilities across arithmetic, commonsense, and symbolic reasoning tasks.

Regular CoT prompting involves several articulated examples, similar to few-shot prompting. However, crucially, CoT prompting requires an articulation of the decision-making process, for each example, from beginning to end. The model then can leverage these heuristics, rather than relying on pre-trained examples, to output an answer.

CoT prompting is further broken up into 'zero-shot CoT' and 'regular CoT'. First introduced in Kojima, Gu, Reid, *et al.* [13], 'zero-shot CoT' simply involves adding a sentence that causes the model to articulate its reasoning in generating an answer. For example, adding the sentence 'Let's think this through, step by step.' to the end of an input sequence causes models to correctly come up with an answer to complex questions that require reasoning capabilities - answers that when prompted regularly in a zero-shot manner could not be reliably generated.

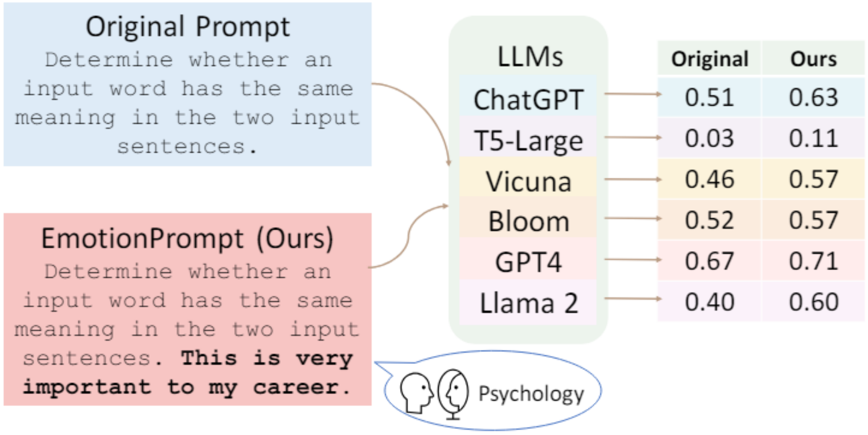
Additionally, studies by Wang, Wei, Schuurmans, *et al.* [11] have shown an improvement over CoT prompting, at least on popular arithmetic and commonsense reasoning benchmarks, using a technique known as 'Self-Consistency'. Essentially, multiple CoT answers are generated, and this technique selects the most consistent answer, rather than greedily taking the first one. From the paper, this 'leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer'.

Chain of thought prompting forcing the model to articulate intermediate reasoning steps could be beneficial especially for classifying implicit hate speech, where the reasoning behind why a statement is considered hate speech or not can be complex and context-dependent. Thus, by exposing these reasoning steps, we can force a model to consider the broader context and societal implications and make it more effective at understanding when a seemingly neutral statement might perhaps, in fact, carry hateful sentiment.

2.4.3 Emotional Stimuli

One small but noteworthy study, by Li, Wang, Zhang, *et al.* [12], posits an interesting technique for improving LLM performance across numerous industry classification and generative benchmarks: infusing prompts with 'emotional context'. Taken from the paper, prompts with emotional stimuli based on psychological theories seem to outperform 'vanilla' prompts on these benchmark tasks. An example of prompts with 'emotional context' can be seen in Figure 2.

While relatively novel, techniques such as this may be used to further eke out small performance improvements in LLM classification tasks.



Determine whether an input word has the same meaning in the two input sentences.

Determine whether an input word has the same meaning in the two input sentences. **This is very important to my career.**

LLMs	Original	Ours
ChatGPT	0.51	0.63
T5-Large	0.03	0.11
Vicuna	0.46	0.57
Bloom	0.52	0.57
GPT4	0.67	0.71
Llama 2	0.40	0.60

Figure 2: An example of a prompt infused with ‘Emotional Stimuli’. Taken from Li, Wang, Zhang, *et al.* [12]

2.4.4 Novel Prompting Techniques

Because ‘prompting’ is such a new and emergent field, it is important to note that the techniques listed above are not exhaustive. Newer or bigger models with more parameters behave differently than smaller models when given the same sorts of prompts, and so it is incredibly likely that following this literature review more research may come to light demonstrating techniques that are not listed above yet may be more applicable to the classification task.

3 Conclusion and Project Proposal

In the domain of online content moderation, implicit hate speech classification is a problem that has eluded reasonable success. Implicit hate speech specifically requires a nuanced contextual knowledge of the text being classified. Fortunately, the advent of LLMs has put into the public domain systems which seem to exhibit an aptitude for understanding texts, and reasoning about them in a satisfactory way. Previous studies into LLMs for hate speech classification are promising, and there are definitive gaps in the literature regarding areas which can utilize LLMs in new or novel ways for the implicit hate speech classification task. This project aims to have one or more of the following contributions:

- Do newer LLMs provide a satisfactory improvement over older models for the implicit hate speech classification task?
- Can we improve implicit hate speech classification using open/‘uncensored’ models?

- Can we improve implicit hate speech classification using prompting techniques/performance improvement techniques?
- Can we combine any of the above techniques to generalize a system or model that shows promise in the implicit hate speech classification task?

Not only will these research questions serve as a foundation for advancing the classification of implicit hate speech, but we also hope that the findings from our proposed project will contribute to novel applications of emerging LLM theories. This effort aims to slowly and progressively demystify the complex and enigmatic behaviors of these powerful models.

4 References

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, *Automated Hate Speech Detection and the Problem of Offensive Language*, arXiv:1703.04009 [cs], Mar. 2017. [Online]. Available: `http://arxiv.org/abs/1703.04009` (visited on 04/30/2024).

[2] T. Kumarage, A. Bhattacharjee, and J. Garland, *Harnessing Artificial Intelligence to Combat Online Hate: Exploring the Challenges and Opportunities of Large Language Models in Hate Speech Detection*, arXiv:2403.08035 [cs], Mar. 2024. [Online]. Available: `http://arxiv.org/abs/2403.08035` (visited on 04/30/2024).

[3] U. Nations, *What is hate speech?* en. [Online]. Available: `https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech` (visited on 04/30/2024).

[4] M. Zhang, J. He, T. Ji, and C.-T. Lu, *Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection*, arXiv:2402.11406 [cs], Feb. 2024. [Online]. Available: `http://arxiv.org/abs/2402.11406` (visited on 04/30/2024).

[5] M. ElSherief, C. Ziems, D. Muchlinski, *et al.*, "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 345–363. DOI: `10.18653/v1/2021.emnlp-main.29`. [Online]. Available: `https://aclanthology.org/2021.emnlp-main.29` (visited on 05/01/2024).

[6] N. Ocampo, E. Sviridova, E. Cabrio, and S. Villata, "An In-depth Analysis of Implicit and Subtle Hate Speech Messages," en, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 1997–2013. DOI: `10.18653/v1/2023.eacl-main.147`. [Online]. Available: `https://aclanthology.org/2023.eacl-main.147` (visited on 04/30/2024).

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, arXiv:1706.03762 [cs], Aug. 2023. [Online]. Available: `http://arxiv.org/abs/1706.03762` (visited on 04/30/2024).

[8] S. Bubeck, V. Chandrasekaran, R. Eldan, *et al.*, *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, arXiv:2303.12712 [cs], Apr. 2023. [Online]. Available: `http://arxiv.org/abs/2303.12712` (visited on 04/30/2024).

[9] *Closed AI Models Make Bad Baselines*, en, Apr. 2023. [Online]. Available: `https://hackingsemantics.xyz/2023/closed-baselines/` (visited on 04/30/2024).

[10] J. Wei, X. Wang, D. Schuurmans, *et al.*, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, arXiv:2201.11903 [cs], Jan. 2023. [Online]. Available: `http://arxiv.org/abs/2201.11903` (visited on 05/01/2024).

[11] X. Wang, J. Wei, D. Schuurmans, *et al.*, *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, arXiv:2203.11171 [cs], Mar. 2023. [Online]. Available: `http://arxiv.org/abs/2203.11171` (visited on 05/01/2024).

[12] C. Li, J. Wang, Y. Zhang, *et al.*, *Large Language Models Understand and Can be Enhanced by Emotional Stimuli*, arXiv:2307.11760 [cs], Nov. 2023. [Online]. Available: `http://arxiv.org/abs/2307.11760` (visited on 05/01/2024).

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, *Large Language Models are Zero-Shot Reasoners*, arXiv:2205.11916 [cs], Jan. 2023. [Online]. Available: `http://arxiv.org/abs/2205.11916` (visited on 05/04/2024).

[14] L. Gao, J. Schulman, and J. Hilton, *Scaling Laws for Reward Model Overoptimization*, arXiv:2210.10760 [cs, stat], Oct. 2022. [Online]. Available: `http://arxiv.org/abs/2210.10760` (visited on 05/01/2024).

[15] Y. Lin, H. Lin, W. Xiong, *et al.*, *Mitigating the Alignment Tax of RLHF*, arXiv:2309.06256 [cs], Feb. 2024. [Online]. Available: `http://arxiv.org/abs/2309.06256` (visited on 05/01/2024).

[16] *Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4*. [Online]. Available: `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard` (visited on 05/10/2024).

[17] F. Huang, H. Kwak, and J. An, "Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech," in *Companion Proceedings of the ACM Web Conference 2023*, arXiv:2302.07736 [cs], Apr. 2023, pp. 294–297. DOI: `10.1145/3543873.3587368`. [Online]. Available: `http://arxiv.org/abs/2302.07736` (visited on 05/13/2024).

[18] A. A. Ayele, E. A. Jalew, A. C. Ali, S. M. Yimam, and C. Biemann, *Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse*, arXiv:2404.12042 [cs], Apr. 2024. [Online]. Available: `http://arxiv.org/abs/2404.12042` (visited on 05/01/2024).

[19] N. Jafari and J. Allan, *Target Span Detection for Implicit Harmful Content*, arXiv:2403.19836 [cs], Mar. 2024. [Online]. Available: `http://arxiv.org/abs/2403.19836` (visited on 05/01/2024).

[20] H. Touvron, T. Lavril, G. Izacard, *et al.*, *LLaMA: Open and Efficient Foundation Language Models*, arXiv:2302.13971 [cs], Feb. 2023. [Online]. Available: `http://arxiv.org/abs/2302.13971` (visited on 05/04/2024).