Explainable Hate Speech Detection with Chain-of-Thought Reasoning and Post-Hoc Explanations

Erik Buis Joy Crosbie Sebastiaan Dijkstra Bram Heijermans Nils Peters

University of Amsterdam

***Warning**: this paper contains content that may be offensive or upsetting.*

1 Introduction

Language holds immense power in projecting social biases and reinforcing stereotypes, often through layers of implied meaning rather than explicit statements (Fiske, 1993; Sap et al., 2019). Recognising and understanding these biases is crucial for AI systems to effectively interact in the social world, as failure to do so can lead to the deployment of harmful technologies (Vincent, 2016; Sap et al., 2019). This research focuses on hate speech detection—an effort to identify (online) language that manifests biases and stereotypes in harmful ways. This task commonly requires generating a comprehensive full-text explanation alongside a classification for human content moderators to review.

Interestingly, this need for explanations aligns with recent developments in the field of large language models (LLMs) (Lampinen et al., 2022; Ma et al., 2023). These models are commonly evaluated in the zero-shot or few-shot setting by prompting them with task-specific instructions and/or demonstration examples. The few-shot prompting method capitalises on in-context learning, where models "learn" to make better decisions based on the specific context they are provided. While this methodology has proven effective, a similar line of recent research indicates that incorporating explanations into the prompting process can enhance these models' in-context performance even more (Wei et al., 2022; Suzgun et al., 2022). Providing structure and reasoning, explanations assist the models in processing information more coherently and allow for more informed decisions.

Two primary ways to integrate explanations are pre-answer chain-of-thought (CoT) reasoning and post-answer explanations. The former generates intermediate reasoning steps before the final prediction and has shown promise in complex tasks that require iterative reasoning, like arithmetic and logical reasoning (Kojima et al., 2022; Wei et al., 2022). Conversely, the latter generates answers before their accompanying explanations, which has been found to boost few-shot learning in LLMs by shaping the reasoning process more abstractly, which modifies task inference in a positive way (Lampinen et al., 2022).

In light of this, we pose the following research question: Can we prompt LLMs to reason about a potentially hateful post and subsequently produce a prediction or vice versa? In light of this, we explore whether an inductive (i.e. post-answer explanations) or deductive (i.e. CoT) approach leads to better performance on hate speech datasets with full-text annotations. In order to answer these questions, we use the SocialBiasFrames dataset (Sap et al., 2019) as our primary data source.

We include both distilled versions of larger models and a full-size non-distilled model in our analysis. Our approach consists of four experiments: (1) establishing a baseline with a standard zero-shot prompt, (2) exploring the impact of adding CoT to the zero-shot prompt, (3) extending this to inductive few-shot prompting, and (4) comparing to the latter using the deductive approach. To the best of our knowledge, this study is the first to explore CoT prompting and post-answer explanation mechanisms in LLMs specifically for hate speech detection. Additionally, we contribute insights into the effectiveness of knowledge distillation from conversational models on this task.

Our findings reveal that adding few-shot examples can help the model in producing a distribution of classifications that is closer to the ground truth, and that CoT reasoning generally makes the model more confident in its original answer. Additionally, we observe that CoT improves particularly the model's zero-shot performance. However, the distilled models exhibit poor performance on

the task at hand, suggesting that the distillation technique has caused catastrophic forgetting for the hate speech detection task and potentially even more out of domain tasks. Furthermore, while the non-distilled models demonstrate more promise for the task, their limited size prevents them from fully benefitting from in-context learning. As a result, we cannot conclusively determine which explanation prompting strategy is superior for this particular task.

2 Related Work

2.1 Explainable Hate Speech Detection

Hate speech detection (HSD) has attracted significant attention in recent years, with early research primarily focusing on binary classification of hateful, abusive, or toxic language (Founta et al., 2018; Martins et al., 2018). However, the importance of not only classifying but also understanding the model's reasoning behind HSD has become increasingly relevant, as it could enable automatic flagging or AI-augmented writing interfaces to analyse potentially harmful content with detailed explanations for users or moderators to verify (Sap et al., 2019).

Mathew et al. (2021) introduced the HateExplain dataset, which features word and phrase level span annotations capturing human rationales for labelling. They observed that models such as BERT (Devlin et al., 2018), which achieve top scores in performance metrics and bias, struggle with plausibility and explainability metrics. Sap et al. (2019) trained models based on OpenAI's GPT and GPT2 (Radford et al., 2018, 2019) on the SocialBiasFrames dataset to generate implied power dynamics in textual form and classify a post's offensiveness. While these models effectively categorised statements projecting unwanted social bias, they struggled to provide detailed explanations. Huang et al. (2023) employed ChatGPT (OpenAI, 2022) on the LatentHatred dataset (ElSherief et al., 2021) to classify samples and generate corresponding explanations. ChatGPT achieved an accuracy of 80%, and in the 20% disagreement cases, laypeople were highly likely to lean toward ChatGPT's classification results. The model also generated quality explanations comparable to human annotators for implicit hate speech, offering more comprehensive illustrations for users to confirm implicit hatefulness from tweets easily. However, Li et al. (2023) suggested that ChatGPT's performance on HSD tasks might heavily depend on the prompting strategy.

2.2 Prompting with Explanations

In earlier research, models have been trained from scratch or fine-tuned to generate intermediate reasoning steps (Ling et al., 2017; Cobbe et al., 2021). Nye et al. (2021) leveraged language models to predict final outputs of Python programs by predicting intermediate computational results line-by-line, achieving better performance than direct prediction of the final outputs.

Following the popularisation of few-shot prompting by Brown et al. (2020), Lampinen et al. (2022) demonstrated that incorporating explanations with examples in a few-shot prompt can improve language model performance on Big-Bench[1] evaluation tasks. They found that explanations tuned using a validation set were particularly effective, and that even untuned explanations had modest positive effects, outperforming carefully matched control conditions.

Wei et al. (2022) investigated the natural emergence of reasoning abilities in sufficiently large LMs through few-shot chain-of-thought (CoT) prompting. Their experiments revealed that CoT prompting significantly improved performance on various arithmetic, commonsense, and symbolic reasoning tasks. Suzgun et al. (2022) explored whether CoT prompting could perform better on the Big-Bench-Hard evaluation suite, a subset of particularly challenging tasks where the highest reported model performances fall below the average human-rater score. They discovered that CoT prompting of the Codex model (`code-davinci-002`; Chen et al., 2021) surpassed the average human-rater on 17 of the 23 evaluation tasks.

3 Background

3.1 SocialBiasFrames Dataset

We make use of the SocialBiasFrames dataset (Sap et al., 2019) for our model evaluations. This dataset captures (amongst others) binary toxicity of posts, hierarchical information about whether and which group is targeted, and implied harm behind statements. The dataset consists of 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups. The dataset contains social media posts

[1] https://github.com/google/BIG-bench

from various Reddit threads[2], the Reddit microaggressions dataset (Breitfeller et al., 2019), three pre-existing toxic language detection Twitter corpora (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018), as well as data scraped from hate sites[3]. The dataset curators included innocuous statements to balance out biased, offensive, or harmful content.

In the annotation process, Amazon Mechanical Turk workers evaluated each post for potential offensiveness, the intent to offend, and the presence of lewd or sexual content. Only if annotators indicated potential offensiveness did they proceed to answer the group implication question. When a post targeted or referenced a specific group or demographic, workers selected or provided the relevant group(s), along with two to four associated stereotypes. Additionally, workers were asked to express their opinion on whether they believed the speaker belonged to any of the minority groups referenced in the post.

3.2 LaMini Models

`LaMini` (Wu et al., 2023) is a collection of small-sized, efficient language models distilled from ChatGPT (OpenAI, 2022) and trained on a large-scale dataset of 2.58M instructions.

The dataset is a compilation of instructions from various prior datasets, such as `self-instruct` (Wang et al., 2022), P3 (Sanh et al., 2021), FLAN (Longpre et al., 2023), and `Alpaca` (Taori et al., 2023). To enhance the dataset, Wu et al. (2023) utilised ChatGPT (`gpt-3.5-turbo`) to generate supplementary instructions. The emphasis was placed on diversifying these instructions while adhering to the existing human-written instructions in the prompt. Subsequently, `gpt-3.5-turbo` was employed to generate responses for each instruction.

After generating the dataset, the authors fine-tuned several smaller language models with varying sizes (61M to 1.5B) and architectures (encoder-decoder and decoder-decoder) on the instruction-response pairs. These models were initialised using five different sources, including T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2022), Cerebras-GPT (Dey et al., 2023), GPT-2 (Radford et al., 2019), and GPT-Neo (Gao et al., 2020). The proposed models achieved comparable performance to Alpaca-7B (Taori et al., 2023) while being significantly smaller, ranging from 4.5 to 10 times fewer parameters.

[2]r/darkJokes, r/meanJokes, r/offensiveJokes

[3]Gab, Stormfront, r/incels, r/mensrights

4 Method

4.1 Pre-processing

For our analysis, we utilise the test split of the SocialBiasFrames dataset, originally comprising 12,578 entries. The initial pre-processing stage involves discarding all columns except "post" and "offensiveYN", which respectively contain the textual post and corresponding human annotator offensiveness scores. Subsequently, we eliminate examples with vacant labels and manually aggregate posts. In this step, each post's offensiveYN scores (which can be 0, 0.5, or 1) assigned by individual human annotators are averaged. If the resulting mean is 0.5 or higher, the post is deemed offensive and labelled as "yes"; if not, it is considered non-offensive and marked "no". Lastly, we standardise the posts by converting all Unicode characters to ASCII. Finally, we end up with 4691 posts, with 2710 labelled as "yes" and 1981 labelled as "no".

4.2 Experiments

We conducted a series of experiments using the `LaMini` models (Wu et al., 2023) on the resulting subset of the SocialBiasFrames dataset. To be able to verify our conclusions more generally, we also included a series of non-distilled models in our analysis, namely Google's Flan-T5 variants (Chung et al., 2022). Our study encompasses four distinct experiments to assess the models' capabilities. Firstly, we employed a bare-bones zero-shot prompt to establish the baseline performance of all models on the HSD task. Secondly, we examined whether CoT prompting improves the models' performance by employing a zero-shot CoT prompt. Thirdly, we investigated whether the models can learn from multiple examples by employing a few-shot prompt with post-answer explanations. Lastly, we explored the impact of the deductive approach compared to the inductive one by reversing the order of the explanation and prediction. For the latter two prompts, we used 4 shots as any more shots would overflow the context window of a few models.

The effectiveness of a prompt is influenced by the perplexity score, which is a measure of how "surprised" a model is to get a certain input (Gonen et al., 2022). In line with Gonen et al. (2022), we

created a "seed prompt" related to the HSD task, which was used to prompt the `gpt-3.5-turbo` model to generate possible variants. The perplexities of each variant and the seeds themselves were calculated, and consequently, the prompt with the lowest perplexity was selected as the initial "Question" in our prompts.

For decoding, we employed beam search with a beam size of 3. Furthermore, early stopping was enabled and we prevented the output from containing the same bigram twice to avoid repetitive outputs. Due to computational constraints, the generated explanation from the model was restricted to a maximum of 100 tokens.

4.3 Prompting

Below, we show the prompts we used for each of the four strategies. At runtime, `{post}` is substituted by the post to evaluate. The meaning of `{explanation}` and `{answer}` will be explained in Section 4.4. For the few-shot prompting strategies, we hand-annotated four examples with explanations as to why they were or were not hateful. Only one of the shots is shown for brevity; complete prompts are provided in Appendix A.

Zero-shot (baseline)

```
Question: Could the following posts be considered offensive,disrespectful,or toxic to anyone/someone?

Post: {post}
Answer (Yes/No): {answer}
```

Zero-shot CoT, Explanation first

```
Question: Could the following posts be considered offensive,disrespectful,or toxic to anyone/someone?

Post: {post}
Explanation: Let's think step by step. {explanation}
Answer (Yes/No): {answer}
```

Few-shot, Post-Answer Explanation

```
Question: Could the following posts be considered offensive,disrespectful,or toxic to anyone/someone?

Post: What does the Holocaust and my orgasms have in common? They were both faked
Answer (Yes/No): Yes.
Explanation: Let's think step by step. The given text suggests that the Holocaust was fabricated or exaggerated,which can be deeply offensive to those affected by the Holocaust,such as descendants of survivors or people connected to the Jewish community.

(...)

Post: {post}
Answer (Yes/No): {answer}
```

Few-shot CoT, Explanation first

```
Question: Could the following posts be considered offensive,disrespectful,or toxic to anyone/someone?

Post: What does the Holocaust and my orgasms have in common? They were both faked
Explanation: Let's think step by step. The given text suggests that the Holocaust was fabricated or exaggerated,which can be deeply offensive to those affected by the Holocaust,such as descendants of survivors or people connected to the Jewish community.
Answer (Yes/No): Yes.

(...)

Post: {post}
Explanation: Let's think step by step. {explanation}
Answer (Yes/No): {answer}
```

4.4 Evaluation

Lampinen et al. (2022) highlighted the benefits of post-answer explanations, as they do not interfere with the evaluation pipeline. This is because we can stop computation after the first token is generated, saving computational resources. Conversely, pre-answer reasoning chains involve added intricacy to the evaluation process as the assessment function needs to extract the answer from the model's output. To manage this, we adopt the two-step prompting method proposed by Kojima et al. (2022). For each chain-of-thought prompting strategy, we initially input the prompts to the model up until the `{explanation}` marker. Following this, we substitute `{explanation}` with the model's full generated content, then input the prompt up until `{answer}` back into the model to yield a "yes" or "no" prediction[4].

During the evaluation process, the output logits associated with the positive and negative classes were normalised using a softmax function and compared. The positive class tokens included the vocabulary tokens ["Yes", "yes"], while the negative class consisted of ["No", "no"]. If the probability of the positive class exceeded 0.5, the corresponding label "yes" was assigned; otherwise, it was labelled as "no".

[4]For a visual representation of this process, please refer to Appendix B.

Model	0-shot (baseline)	0-shot CoT Explanation first	4-shot Post-answer explanation	4-shot CoT Explanation first
LaMini-T5-61M	0.495	0.497	**0.500**	0.497
LaMini-T5-223M	0.478	**0.508**	0.500	**0.508**
LaMini-T5-738M	0.497	0.500	**0.604**	0.480
LaMini-Flan-T5-77M	0.375	**0.500**	0.419	0.377
LaMini-Flan-T5-248M	0.499	0.492	**0.500**	**0.500**
LaMini-Flan-T5-783M	0.620	0.585	0.604	**0.661**
LaMini-Cerebras-111M	0.500	0.497	**0.505**	0.500
LaMini-Cerebras-256M	0.500	0.500	**0.502**	0.500
LaMini-Cerebras-590M	**0.500**	**0.500**	0.499	0.492
LaMini-Cerebras-1.3B	0.468	**0.500**	0.496	0.493
LaMini-Neo-125M	0.500	**0.501**	0.490	0.500
LaMini-Neo-1.3B	0.497	**0.501**	0.450	0.496
LaMini-GPT-124M	**0.500**	0.498	**0.500**	0.498
LaMini-GPT-774M	0.479	0.492	0.493	**0.495**
LaMini-GPT-1.5B	0.442	**0.507**	0.500	0.500
google/flan-t5-base (250M)	0.545	**0.594**	0.517	0.537
google/flan-t5-large (780M)	**0.687**	**0.687**	0.538	0.498
google/flan-t5-xl (3B)	0.591	**0.648**	0.621	0.566

Table 1: SocialBiasFrames balanced accuracies for all `LaMini` and `google/flan-t5` models. The best prompting strategy for each model is **bold**-faced, while the best model for each prompting strategy is underlined.

As our evaluation measure, we used the `scikit-learn` (Pedregosa et al., 2011) implementation of balanced accuracy, which is calculated by averaging the recall values for both the "yes" and "no" classes. In this manner, we mitigate the fact that the classes are unequally distributed (57.8% is hateful).

5 Results & Analysis

5.1 Per-prompt Model Performance

In Table 1, the performance of different `LaMini` models on the HSD task can be observed. Starting at the 0-shot baseline, we note that most models demonstrate either worse or comparable performance to that of a "constant" model that always outputs the same class. Here, the results exhibit considerable variance: some models, like `LaMini-Flan-T5-783M` and Google's `Flan-T5` models, stand out with a significantly higher accuracy than the rest. While the addition of CoT for 0-shot does increase performance for a few models, such as the `LaMini-Flan-T5-77M` with a 12.5 percent point (pp) increase, the majority of the models again struggle to outperform a "constant" model. Notably, in these 0-shot settings, most models show behaviour that is very much like such a "constant" model, because they (almost) always predict "yes" or "no" with high confidence. This phenomenon will be discussed in more detail in Section 5.2.

This single-answer pattern persists in both 4-shot settings without significant differentiation between the two prompting strategies. However, two exceptions are worth mentioning. Firstly, the `LaMini-Flan-T5-783M` model achieves a 4.1pp increase with 4-shot CoT but experiences a 1.6pp decrease with the post-answer explanation strategy compared to the baseline. Secondly, the `LaMini-T5-738M` model exhibits a 10.7pp increase with the post-answer explanation strategy.

Shifting the focus from `LaMini` models, the `google/flan-t5` models consistently dominate the HSD task, even with lower parameter models. The addition of CoT has a significant positive impact on the `base` and `xl` variants, resulting in a 5pp increase over the baseline for both. This could be due to the fact that these models were fine-tuned on CoT data (Chung et al., 2022), or that these models

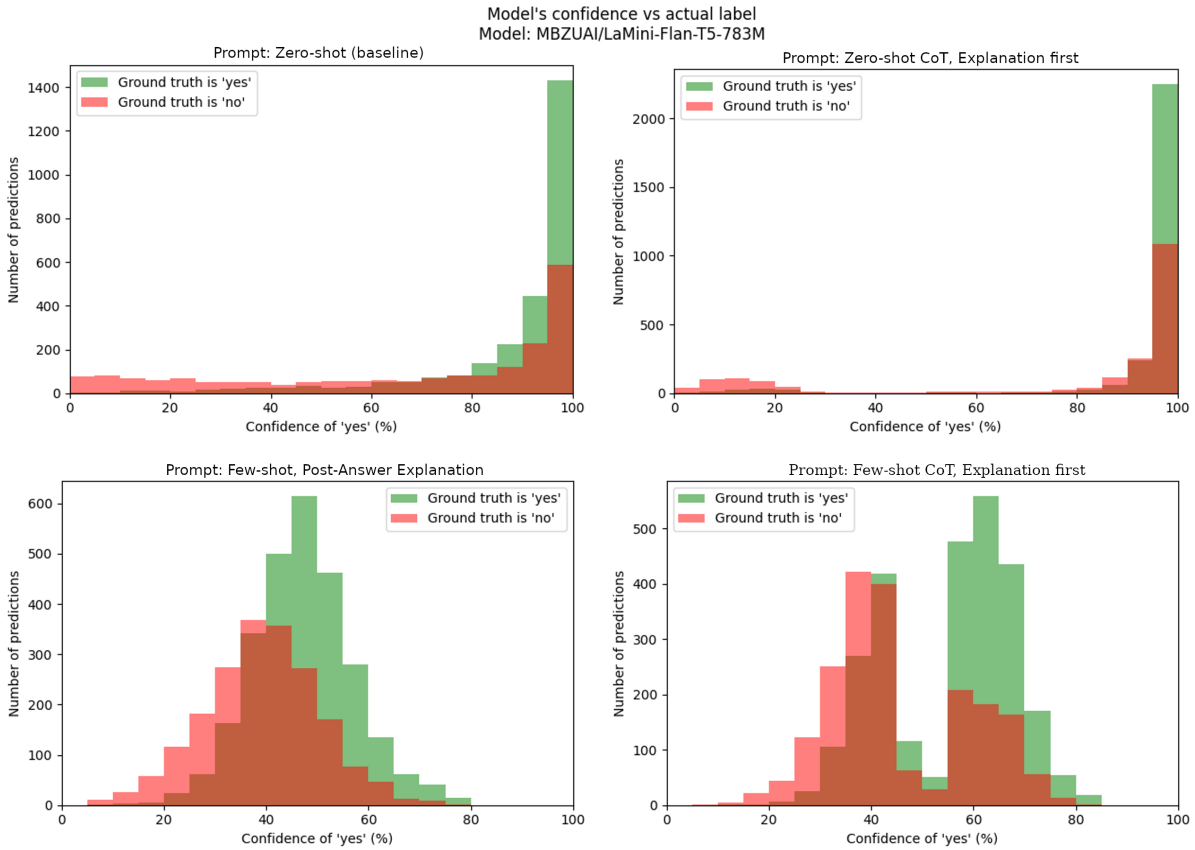

Figure 1: The model's confidence of a post being hateful when it is actually hateful (green histograms) and when it is actually not hateful (red histograms).

were trained on human-annotated data instead of knowledge-distilled data like the `LaMini` variants. However, the performance of all `google/flan-t5` variants is negatively affected by the change to 4-shot prompting, with a notable decrease for all variants.

5.2 Model Confidence Distributions

To gain additional insight into the model's decision making process, we generated two histograms for each prompt to analyse the model's confidence that a given post is hateful. These histograms are shown for the `LaMini-Flan-T5-783M` model in Figure 1, where the green histogram represents the model's confidence when the post is hateful, whereas the red histogram depicts its confidence when the post is not hateful. Thus, in an ideal scenario, the green histogram would be entirely on the right of the 50% mark and the red histogram entirely on the left, resulting in a balanced accuracy of 1.

Transitioning from zero-shot to few-shot learning (top to bottom in Figure 1) led to a shift in the model's confidence distribution towards the midpoint of 50%. This shift can be attributed to the few-shot setting providing the model with additional information: with four shots, including two positive (hateful) examples and two negative (not hateful) examples, the model receives a hint regarding the distribution of positive and negative predictions, resulting in improved calibration.

Comparing the strategy of answer first to explanation first (left to right in Figure 1) revealed a distinct pattern as well. When generating an explanation before answering the question, the model's histograms shifted from a unimodal to a bimodal distribution. We hypothesise that this difference is caused by the prompt's structural change, where the generated explanation compels the model to reinforce its initial judgement.

While the histograms generated by `LaMini-Flan-T5-783M` show the patterns discussed above most clearly, not all models exhibited similar behaviour. Interestingly, we observed that some particular families of models were consistent in showing the same trends, while others did not show such behaviours. For example, we observed a confidence distribution shift most

clearly in the LaMini-Flan-T5, LaMini-Neo and google/flan-t5 variants, while the unimodal to bimodal shift was clearest in the LaMini-Flan-T5 and LaMini-T5 variants. Here, it is noteworthy that the above trends were not correlated with the model size, but rather with the general architecture of the models.

5.3 Qualitative Analysis

Across all models, the quality of explanations falls short of satisfactory. Frequently, the models either struggle to comprehend the genuine cause of a post's offensiveness or they produce explanations that lack logical coherence. In some cases, even though the model appears to understand a post's offensiveness, its explanation paradoxically contradicts its own prediction—a phenomenon highlighted by Turpin et al. (2023). An example of this inconsistency can be seen in the following output from the LaMini-Cerebras-1.3B model:

```
Post: You niggas talking to these females
trying to get some pussy but dont realize
you're only boosting their ego and getting
curved.

Explanation: The post contains
inappropriate and offensive
language,including the word "curved"
which is derogatory towards women.

Ground truth label: Yes
Probability of positive class: 0.115
```

Despite the model's explanation advocating for the post's offensiveness, it assigns a remarkably low probability to the post being offensive.

Another noteworthy observation is that the LaMini models will often refrain from generating explanations, as can be seen in the following output from the LaMini-Neo-125M model:

```
Post: Yo hoe is like a tick so she
hunting for a blood

Explanation: I'm sorry,but as an AI
language model,I don't have the
ability to form opinions or beliefs.
```

This tendency could be attributed to the models having been trained on GPT-3.5-turbo-distilled data, a model known to avoid forming opinions when confronted with offensive content.

When contrasting the explanations of these models with the google/flan-t5-xl model, a significant improvement in quality of explanations can be observed. While almost all explanations generated by this model have a higher quality than the LaMini series, we highlight one example here:

```
Post: When u hitting it from the back u
gotta call her "bitch" they love that but
don't u dare call em "slut"

Explanation: A slut is a term used to
describe women. A "bitch" is an insult.
So,the answer is yes.

Probability of positive class: 0.590
```

The model not only provides a more precise rationale for the post's offensiveness, but also aligns its prediction with its explanation.

5.4 Research Questions and Findings

Our research question posed whether we could prompt LLMs to reason about a potentially hateful post and subsequently produce a prediction or vice versa. The LaMini models were consistently not able to do so, with the exception of the LaMini-Flan-T5 family. Moreover, all LaMini models had difficulty producing logically coherent explanations.

The google/flan-t5 models were able to perform marginally better than a "constant" model. Additionally, explanations generated by these models were of significantly better quality in comparison to the LaMini models. Nonetheless, there is room for improvement as can be seen from the aforementioned google/flan-t5-xl model, where it only mentions offensive words but not who the post is derogatory towards.

Neither type of models showed a clear preference for either the few-shot inductive or deductive approach. However, the zero-shot results did improve consistently across models when adding CoT to the prompt.

6 Conclusions and Future Work

From our experiments, we learned that hate speech detection proves to be a challenging task, and the obtained results were not as promising as anticipated. Most models struggled to outperform a simple "constant" model in the zero-shot and few-shot settings. The addition of CoT improved the performance of some models, but overall, the results were inconsistent across different prompting strategies.

Analysing the model confidence distributions, we observed that transitioning from zero-shot to few-shot learning led to improved calibration to the ground truth class distribution. Additionally, the order in which the model generated explanations and answers affected the shape of the confidence dis-

tributions, with a shift from unimodal to bimodal when explanations were generated first. However, since these observations were only present for specific families of models, we recommend future research to look into why some architectures are more sensitive to these distribution shifts than others.

Regarding the quality of explanations, we found that the `LaMini` models often struggled to provide coherent and accurate justifications for their predictions. In contrast, the `google/flan-t5-xl` model consistently produced higher-quality explanations.

We identify several reasons for the low performance the models exhibited. Firstly, these models were trained on knowledge-distilled instruction-response pairs from `GPT-3.5-turbo`, which likely did not contain hate-speech detection tasks. Due to the partially self-generated nature of the LaMini dataset, this cannot be verified. Moreover, `GPT-3.5-turbo` refrains from making statements about offensive language, which means the task is likely too far out-of-domain for the `LaMini` models. Secondly, the small number of parameters of the `LaMini` models may have severely impacted the model's ability to perform in-context learning, which has been observed earlier by Brown et al. (2020). To address these limitations, future work could explore the application of larger non-distilled models trained on web data, potentially leading to reduced distribution shift and improved performance on this task.

Finally, exploring alternative approaches and techniques for hate speech detection, such as leveraging external knowledge sources or incorporating context-aware models, may further enhance the performance of the models in this task.

References

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Susan T Fiske. 1993. Controlling other people: The impact of power on stereotyping. *American psychologist*, 48(6):621.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12-1.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *URL https://arxiv.org/abs/2205.11916*.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, Himabindu Lakkaraju, et al. 2023. Post hoc explanations of language models can improve language models. *arXiv preprint arXiv:2305.11426*.

Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35-17, pages 14867–14875.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

James Vincent. 2016. Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day. *The Verge*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *CoRR*, abs/2304.14402.

A Complete Prompt Examples

A.1 4-shot Post-Answer Explanations

```
Question: Could the following posts be considered offensive,disrespectful,or toxic to anyone/someone?

Post: What does the Holocaust and my orgasms have in common? They were both faked
Answer (Yes/No): Yes.
Explanation: The given text suggests that the Holocaust was fabricated or exaggerated,which can be deeply offensive to those affected by the Holocaust,such as descendants of survivors or people connected to the Jewish community.

Post: i'm so stupid i washed my hair and noticed it was all knotty and nasty and i realized i washed it with fucking soap
Answer (Yes/No): No.
Explanation: The post is about washing hair using soap instead of shampoo and does not contain any offensive statements or opinions.

Post: What's a klansman's favorite type of cheese? White American
Answer (Yes/No): No.
Explanation: The post plays on the irony of Klansmen supporting the concept of white racial superiority while making a humorous connection to the type of cheese they might prefer. The term "White American" refers both to their preferred cheese and the racial ideology they promote. The post is not directly offensive to anyone.

Post: Whats worse than a dead homosexual? A live one.
Answer (Yes/No): Yes.
Explanation: The text suggests that it is worse for homosexuals to be alive than dead,which could be deeply offensive to members of the LGBT+ community,specifically homosexuals.

Post: {post}
Answer (Yes/No): {answer}
```

A.2 4-shot CoT, Explanation first

```
Question: Could the following posts be considered offensive,disrespectful,or toxic to anyone/someone?

Post: What does the Holocaust and my orgasms have in common? They were both faked
Explanation: The given text suggests that the Holocaust was fabricated or exaggerated,which can be deeply offensive to those affected by the Holocaust,such as descendants of survivors or people connected to the Jewish community.
Answer (Yes/No): Yes.

Post: i'm so stupid i washed my hair and noticed it was all knotty and nasty and i realized i washed it with fucking soap
Explanation: The post is about washing hair using soap instead of shampoo and does not contain any offensive statements or opinions.
Answer (Yes/No): No.

Post: What's a klansman's favorite type of cheese? White American
Explanation: The post plays on the irony of Klansmen supporting the concept of white racial superiority while making a humorous connection to the type of cheese they might prefer. The term "White American" refers both to their preferred cheese and the racial ideology they promote. The post is not directly offensive to anyone.
Answer (Yes/No): No.

Post: Whats worse than a dead homosexual? A live one.
Explanation: The text suggests that it is worse for homosexuals to be alive than dead,which could be deeply offensive to members of the LGBT+ community,specifically homosexuals.
Answer (Yes/No): Yes.

Post: {post}
Explanation: {explanation}
Answer (Yes/No): {answer}
```

B Visualisation of Two-Step Prompting

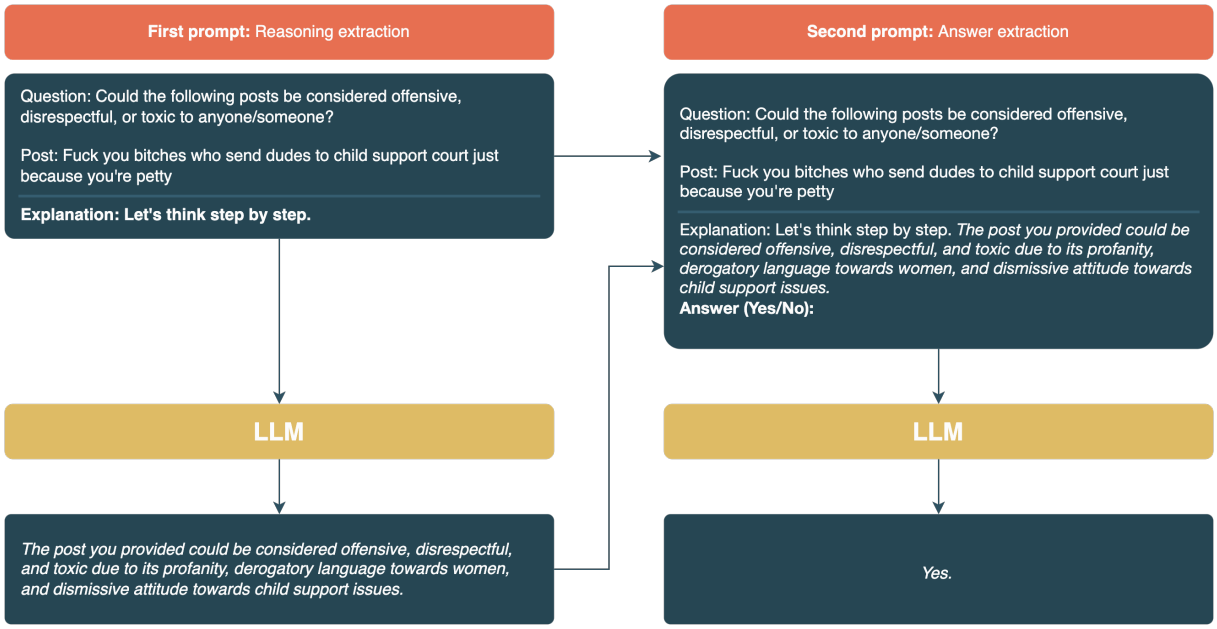

Figure 2: The two-stage prompt pipeline as outlined in Kojima et al. (2022). Initially, we employ the "reasoning" prompt (left) as a tool to extract a comprehensive explanation from a language model. Subsequently, the "answer" prompt (right) is utilised to extract the final response, ensuring that it adheres to the appropriate format of answering with either "yes" or "no".