HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models

Anonymous ACL submission

Abstract

Warning: *This paper contains examples of very offensive content.* The widespread use of social media necessitates reliable and efficient detection of offensive content to mitigate harmful effects. Although sophisticated models perform well on individual datasets, they often fail to generalize due to varying definitions and labeling of "offensive content." In this paper, we introduce *HateCOT*, an English dataset with over 52,000 samples from diverse sources, featuring explanations generated by GPT-3.5-Turbo and curated by humans. We demonstrate that pretraining on *HateCOT* significantly enhances the performance of open-source Large Language Models on three benchmark datasets for offensive content detection in both zero-shot and few-shot settings, despite differences in domain and task. Additionally, *HateCOT* facilitates effective K-shot fine-tuning of LLMs with limited data and improves the quality of their explanations, as confirmed by our human evaluation. Our repository is available at [REDACTED].

1 Introduction

As social media has become indispensable to modern discourse, this channel of communication has amplified the propagation of offensive content. Speech that promotes hateful sentiments thrives on such platforms, leading to real and significant harm on their audience (Giachanou and Rosso, 2020; Saha et al., 2019). However, "offensive content" is still a contested construct, as what is and is not allowed varies by platform. In research, different approaches analyze semantically similar but still distinct concepts: *Cyber-bullying, Toxicity, Sexist, Racist, Hate etc.* (Poletto et al., 2021; Fortuna et al., 2021; Nghiem et al., 2024), further highlighting this contestedness.

Compounding the challenge, reliable detection of offensive content typically requires significant amounts of data. Sophisticated models tend to be data-hungry, and the process of curating a dataset tailored to a specific use case can be costly, time-consuming, and emotionally challenging for annotators (Founta et al., 2018; Toraman et al., 2022). The typical pipeline consists of collecting samples based on topic-relevant key words, then recruiting either crowdworkers or experts to annotate data before developing classification models (Paullada et al., 2021). Each step incurs investment and may inject subtle idiosyncrasies proportionate to the size of the downstream dataset, further limiting transferable usefulness to related tasks (Fortuna et al., 2021). The size of a dataset also does not necessarily guarantee cross-domain transferrability (Poletto et al., 2021; Fortuna et al., 2021).

In practical settings, users may desire transparency from social media platforms. Therefore, the ability to provide human-understandable justification based on platform-specific policy becomes an attractive feature for content moderation. Nevertheless, current techniques often still fail to offer intuitive explanatory signals (Yadav et al., 2023; Babaeianjelodar et al., 2022; Ibrahim et al., 2022).

In this work, we attempt to simultaneously reduce the cost of *data curation*, enhance *cross-dataset generalization*, and address the necessity of *explainable decisions* for offensive content detection. Specifically, our main contributions are:

1. We release **HateCOT** (*Hate-related Chains-of-Thought*), a dataset of over $52,000$ samples consisting of input text, a hate speech label, and an explanation for that label. This corpus is constructed by merging eight datasets with explanations created by using GPT-3.5-Turbo to augment human annotations.

2. We demonstrate the benefits of using *HateCOT* as a pretraining corpus before finetuning on a target domain. Empirical results across 3 datasets show that open-source Large Language Models (LLMs) can effectively lever-

age definitions to adapt to new tasks using zero-shot and few-shot settings via finetuning and in-context learning .

3. We assess the quality of explanations generated by our finetuned models with respect to the criteria described in their corresponding definitions. These insights showcase LLM-generated explanations as a means to enhance transparency in content moderation.

2 Related Works

2.1 Offensive Speech Detection

Offensive speech detection has attracted considerable interest from the research community. Earlier approach typically investigated coarse-grained labels (e.g. *Hate* vs *Not Hate*) while subsequent efforts explored more diverse facets of offensive speech at higher granularity (Founta and Specia, 2021; Poletto et al., 2021; Vidgen and Derczynski, 2020). Increasingly more advanced models emerged over time with the diversity of datasets.

Cross-domain generalization, however, still remains a relevant challenge in the area. Fortuna et al. (2021) found empirically that cross-dataset transference is highly dependent on semantic similarity between their label spaces. Recent works have pretrained Transformer-based models, such as HateBERT and fBERT, on specialized corpora to enhance generalization to various levels of success (Caselli et al., 2021; Sarkar et al., 2021).

2.2 LLMs in Offensive Speech Classification

Zampieri et al. (2023) assessed a range of open-sourced LLMs on zero-shot prompting on the OffensEval task and found their performance trailing by a wide margin behind trained existing finetuned BERT-based systems. Chiu et al. (2021) and Han and Tang (2022) used the proprietary GPT-3 on a set of different datasets and noted that informative contexts and examples could boost the model's performance to competitive levels on a different set of data. Similarly, Roy et al. (2023) found that adding explanation to pipeline could result in 10 to 20% boost in performance of LLMs over baselines. (Yang et al., 2023)'s study found that training LLMs with step-by-step reasoning grounded by annotations could improve predictive power.

Pretrained language models have exhibited remarkable ability in text generation (Celikyilmaz et al., 2020). Recent large-size LLMs such as GPT-3 and later models are capable of astoundingly fluent, convincing and knowledge-infused outputs (Zhang et al., 2023). LLMs with hundreds of billions of parameters even exhibit reasoning capabilities (Wei et al., 2022, 2021), leading to a flurry of research on prompting techniques to harness their prowess, such as Chain-of-thought (COT), Tree-of-thought etc. (Yao et al., 2023; Diao et al., 2023). An interesting line of research leverages LLMs to efficiently generate high volumes of synthetic data for tasks with training resource is scarce (Puri et al., 2020; Bao et al., 2023; Whitehouse et al., 2023). We build upon these works to construct a dataset that can induce smaller LLMs to efficiently adapt to new categories of offensive content by leveraging their provided definitions.

3 Building *HateCOT*

We first describe the process to identify the candidate datasets from literature, and the procedure to obtain annotation-guided explanations from these samples (Section 3.1). We then perform a set of validation experiments to optimize the data's parameters for downstream tasks before augmenting our corpus to its eventual size (Section 3.2).

3.1 Data Selection

Datasets for Training. We use the following criteria to filter existing corpora related to offensive speech detection:

- ▷ *Size*: datasets should contain more than 5,000 samples to ensure adequate size for subsequent sampling.
- ▷ *Label*: datasets should contain diverse label space that address different facets of offensive language. Both neutral and non-neutral categories should be included for parity.
- ▷ *Definition*: each dataset should have the associated definitions with each label available (Figure 6). This criteria is important to generate informative explanations.
- ▷ *Target / Rationale*: the dataset should provide the *a)* targets and/or *b)* rationales, which are fragments of free texts by human annotators to convey some understanding of the corresponding post. Figure 4 shows several examples of rationales, demonstrating their unsuitability to be used as explanations in their native format.

These criteria significantly reduce the number of eligible candidates since many do not provide the required annotation on target and definition. Table

1a lists the 8 selected datasets.

Datasets for Evaluation. Using similar criteria, we select 3 additional datasets with different label spaces and definitions for downstream testing (details in Table 1b).

HateCheck was created by Röttger et al. (2021) with the explicit goal of evaluating hate speech detection models. 10 trained annotators labelled the dataset using a binary schema: *Hateful* and *Non-hateful*, with the reported inter-annotator agreement coefficient to be 0.93.

HateXplain was primarily collected from Twitter and the Gab platform (Mathew et al., 2021). In addition to the labels *Hate, Offensive, Normal*, annotators also provide justification for their selection by highlighting the span of tokens, called rationales, that contribute to their decision.

Latent_Hate was created on the premise that offensive speech classifiers tend to bias towards covert negative sentiment (ElSherief et al., 2021a). After discarding the augmented portion from the Social_Bias dataset to avoid contamination, 22,584 samples collected from Twitter remained. This dataset contains 3 coarse-grained labels *Not Hate, Explicit Hate, Implicit Hate*, while a subset contains 6 fine-grained categories, which we refer to as **Implicit_Hate** in subsequent test regimens.

Obtaining Annotation-Guided Explanation. Inspired by Yang et al. (2023)'s work that shows that GPT-3.5 could augment human-written rationales to create coherent texts that are still faithful to the original content, we use the prompt template in Figure 1 to generate the explanation, which is guided by the available annotations on label, target, and rationale from the chosen corpora. For datasets that contain multiple annotations per sample, we select the ultimate label via majority voting and concatenate annotations on the targets and/or rationales into a single string delimited by "|".

We use GPT-3.5-Turbo, accessed via the OpenAI's API, to generate explanations due to this model's affordability and its capability to follow instructions and generate coherent outputs (Ye et al., 2023a; Koubaa, 2023). For each of the 8 training datasets, we first randomly select and qualitatively analyze 20 samples to ensure the generated explanations are *a) stylistically coherent*, *b) consistent with the provided labels*, and *c) congruent with the criteria denoted by the definitions.* If deemed unsatisfactory, we iteratively adjust the input prompt until the quality threshold is achieved. Appendix A.4 describes this quality assurance process and the final prompts for different scenarios.

By elaborating on the provided Annotation, provide a brief paragraph to explain step-by-step why the post should be classified with the specified Label based on the given Definitions. Be as succinct as possible.

Definitions:
Hate : Contains direct attack on people based on protected characteristics such as race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.
Not Hate : Does not contain any attack on people based on protected characteristics.

Post:
It was for targeted harassment: either me reposting this pic or that a**hole is going down

Label:
Not Hate

Annotation:
Negative sentiment but not attacking on protected traits

Explanation:
This post should be labelled as "Not Hate" because ...

Figure 1: Template used to obtain explanations from GPT-3.5-Turbo guided by human-annotated rationales.

3.2 Optimization of Synthesized Corpus

Extending previous works (Magister et al., 2022; Ho et al., 2022), we are interested in optimizing 2 parameters central to the construction of our corpus: the distribution of neutral vs. non-neutral classes in the data and the number of explanations *per sample*. The former has been noted to influence predictive powers (Rathpisey and Adji, 2019; Casula and Tonelli, 2020). The latter, also referred to as *degree of reasoning diversity*, could improve knowledge distillation (Ho et al., 2022). We use the open-source model **Llama 2 Chat-HF** of 7 billion parameters (hereby referred to as *Llama 7B*) (Touvron et al., 2023) to perform tuning experiments in this stage due to its manageable size and strong classification performance. These empirical findings then guide the final augmentation process.

3.2.1 Optimization Procedure

Below are the experiments we perform on a sample of the collected data to optimize these parameters.

Description of Procedure. We choose 1,000 samples from each of the eight training datasets based on the following distribution: 20% are selected from neutral samples (categories that do not indicate any offensive content, e.g., *Not Hate, Normal*), while the remaining samples are evenly distributed among the non-neutral categories. Inspired by Ho et al. (2022) that diverse reasoning paths could improve knowledge distillation, we collect 4 alternative explanations, or *degree of reasoning diversity*, generated by GPT-3.5-Turbo for these samples using temperature 0.7, resulting in 32,000 samples.

Figure 5 illustrates the *Alpaca*-styled template

Dataset	Total Size	Sample Size	Platform	Target	Ration.	Selected Labels
Salminen et al. (2018)	137,098	5,418	Y, F		✓	Hateful, Neutral
Qian et al. (2019)	34,000	5,034	G		✓	Not Hate, Hate
Sap et al. (2019)	44,671	6,033	G, R, T	✓	✓	Not Offensive, Offensive
Vidgen et al. (2021a)	27,494	6,717	R	✓	✓	Neutral, Person Directed Abuse, Affiliation Directed Abuse, Identity Directed Abuse
Vidgen et al. (2021b)	10,152	7,209	S	✓		None, Derogation, Dehumanization, Animosity, Support, Threatening
Sachdeva et al. (2022)	135,556	7,272	Y, T, R	✓		Not Hate speech, Hate Speech
Hartvigsen et al. (2022)	274,186	7,239	S	✓		Benign, Toxic
Toraman et al. (2022)	100,000	7,215	T		✓	Normal, Offensive, Hate
Total	-	**52,137**				

(a) Datasets used to create training corpus. *Sample Size* denotes the number of chosen samples from corresponding dataset included in the training corpus.

Dataset	Total Size	Split Ratio	K val	K test	Platform	Target	Ration.	Selected Labels
HateCheck	3,728	50:50	300	500	S	✓		Non-hateful, Hateful
HateXplain	20,148	60:40	200	400	G, T	✓	✓	Normal, Offensive, Hate
Latent_Hate	19,112	60:40	200	400	T	✓	✓	Not Hate, Explicit Hate, Implicit Hate
Implicit_Hate	4,153	60:40	-	150	T	✓	✓	White Grievance, Incitement to Violence, Inferiority Language, Irony, Stereotypes and Misinformation, Threatening and Intimidation

(b) Datasets for testing

Table 1: *Sample Size* denotes the number of entries in the final corpus. *Target* and *Ration.* indicates the availability of annotation on Target or Rationale in the dataset. For *Platform*, F: Facebook, Y: Youtube, G: Gab, R: Reddit, S: Synthetic, T: Twitter. *K val* and *K test* represent the number of sampler per class drawn during development of the training corpus and final testing, respectively. Full definitions in Table 6 and 7.

to format each post with its corresponding label, generated explanation, definitions along with the instruction into blocks of an input prompt (Taori et al., 2023). Using the described corpus, we supervised finetune *Llama 7B* via LoRA techniques (technical details specified in Appendix A.2) (Hu et al., 2021). Then, we perform zero-shot classification using the same template to prompt the finetuned model to generate the explanation and label for posts drawn from the test datasets *HateCheck, HateXplain, Latent_Hate*. We omit *Implicit_Hate* at this stage due to this set's markedly different 6-label space. For this part, posts are drawn using K-shot sampling (equal number of samples for each class) on the Validation portion of the test data, based on the values of *K val* shown in Table 2b.

Experiment Configurations. For the first experiment, the training data is split into subsets whose distribution between the neutral (*NE*) class and non-neutral class(es) (*NN*) described by the following formula: $NN = R * NE$, where $R \in \{1, 2, 3, 4\}$ is the ratio coefficient. We set the number of explanations per sample to 2 , the smallest value that still enables the benefit of reasoning diversity. For the second experiment, we construct the subsets by varying the *degree of reasoning diversity* $D \in \{1, 2, 3, 4\}$ of each post.

Answer Extraction. We extract the generated explanation and predicted labels after their respective tags. If the models generate multiple items from the dataset's label space, we select the first admissible label. If no acceptable output is obtained, we randomly select an item in the label space.

3.2.2 Insights and Augmentation

We report *Llama 7B*'s macro F1-scores on the validation set of each configuration in Table 2. A *balanced distribution* between neutral and non-neutral classes in the training corpus is beneficial, as reflected by the substantially high mean F1-score of 0.643 when R=1. On the other hand, having 3 explanations per sample (D=3) achieves the best overall performance across 3 test sets, consistent with Ho et al. (2022)'s findings on the benefit of multiple reasoning paths. However, performance markedly degrades when D=4. Our manual analysis reveals that the quality of generated outputs deteriorates as the degree of diversity increases, consequently affecting the performance of models trained on this data.

Guided by these empirical findings, we augment the training corpus by selecting approximately 1,800 extra samples from each of the 8 datasets while preserving the 1:1 balanced ratio of neutral to non-neutral classes. Then, we collect 3 explanations per sample using the described mechanism, resulting in a final corpus of 52,137 samples (Table 1a), hereby referred to as *HateCOT*.

	R=1	R=2	R=3	R=4
HateCheck	0.879	0.750	0.650	0.574
HateXplain	0.534	0.533	0.495	0.528
Latent_Hate	0.516	0.473	0.456	0.408
Average	**0.643**	0.585	0.534	0.503

(a) Results for Ratio configurations (R)

	D=1	D=2	D=3	D=4
HateCheck	0.851	0.879	0.864	0.783
HateXplain	0.549	0.534	0.597	0.607
Latent_Hate	0.480	0.516	0.477	0.465
Average	0.627	0.643	**0.646**	0.618

(b) Results for number of explanations per sample (D)

Table 2: Macro F1-scores for different configurations of distribution between neutral vs. non-neutral classes (top) and number of explanations per sample (bottom) on validation set. Best average performance in **bold**.

4 Experiments on Test Sets

We perform experiments to answer 3 questions. First, does *HateCOT* improve zero-shot classification of open-sourced LLMs on unseen datasets? Second, how much data is necessary to enable competitive performance via in-domain finetuning after pretraining on *HateCOT*? Finally, is in-context learning a viable alternative to finetuning?

Models In addition to **Llama 7B** in Section 3.2.1, the following open-sourced models are selected.

- ▷ **Llama 13B** A larger variant of the instruction-tuned Llama 7B with 13 billion parameters.
- ▷ **OPT-IML** Based on the original OPT (Open Pre-trained Transformer Language Models) (Zhang et al., 2022), this encoder-only model contains 1.3 billion parameters and was further trained on the Instruction MetaLearning (IML) dataset (Iyer et al., 2022).
- ▷ **Flan-T5-L** Chung et al. (2022) further instruction-finetuned the encoder-decoder T5 family of models (Raffel et al., 2020). We use the Large version of 780 million parameters.
- ▷ **COT-T5-XL** A variant of the Flan-T5-XL (3 billion parameters), this model is further finetuned on the CoT dataset, a collection of 1.8 million samples augmented with chain-of-thought-style explanations (Kim et al., 2023a).

4.1 Zero-shot Classification

We prompt the models to perform classification with no in-context examples via 2 modes: *No Explanation*, where the model directly predicts the label for the input, and *With Explanation*, where a justification is required before the predicted label. We finetune the base models using only *HateCOT* and evaluate their performance on the 4 test sets as in Section 3.2.2 (more details in Appendix A.2).

From results presented in Figure 2a, the smaller models *Flan-T5-L* and *OPT-IML* are unable to generate explanations when prompted. In contrast, their scaled-up counterparts could follow instructions at all settings. Asking base (off-the-shelve) models to generate an explanation before the label results in observable boost to *Llama* models on *HateCheck* and *HateXPlain*, but actually hampers performance on *Latent_Hate* and its derivative *Implicit_Hate*, which are notably challenging due to its covert nature (ElSherief et al., 2021b).

Model Choice Matters Pretraining on *HateCOT* unanimously enables all models to generate explanations. While smaller models receive no observable boost, larger models (*COT-T5-XL, Llama 7B, Llama 13B*) are considerably enhanced compared to their base counterparts. With the exception of *HateXplain*, the *HateCOT*-pretrained version of *COT-T5-XL* attains an increment in F1 scores of 7.6% on *HateXplain* and 9.5% on *Latent_Hate* over the base counterpart without explanations. Similarly, *Llama 7B* observes 23.9%, 25.6%, 10% increment on *HateCheck, HateXplain* and *Latent_Hate*, respectively. These statistics are 27.9%, 118.5%, and 10.2% for *Llama 13B*. Notably, all models yield non-competitive performance on *Implicit_Hate*.

The reduced performance of *HateCOT*-pretrained models compared to their base counterparts without explanations (e.g.: *Flan-T5-L* and *OPT-IML* on almost all test sets) is in line with literature as COT-style prompting tends to favor larger models (Wei et al., 2022; Wang et al., 2024; Suzgun et al., 2023). Even the reduced F1 scores of *COT-T5-XL* on *HateXPlain* and *Implicit_Hate* is consistent with this model's suboptimal performance relative to its larger variants, as showcased in Kim et al. (2023b). These results serve as an empirical reference for researchers to select the appropriate model size for their respective task.

4.2 In-Domain Finetuning

We further finetune *COT-T5-XL, Llama 7B, Llama 13B* using data from the training portions of the test datasets, including *Implicit_Hate*. To simulate low-resource settings, we choose 256 samples uniformly at random from each class, then augment them with explanations as described in Section 3.1. Both the *Base* and *Pretrained* versions of the models are then finetuned using various K-shot $\in \{32, 64, 128, 256\}$ training data from this pool.

In Figure 2b, the general superiority of finetuning models after *HateCOT* over their base counterpart indicates enhanced generalizability with limited in-domain data. However, too little training data (K$\leq$64) may impair models' performance compared to the zero-shot setting, likely a result of attempting to optimize a large number of parameters on limited signals. Stable gains are attained at K=128, and at K=256, significant boost over the non-finetuned zero-shot results are observed.

Interestingly, decoder-only models (*Llama*) considerably outperform encoder-decoder *COT-T5-XL* on the 2 and 3-way classification tasks, yet the reverse is observed for the nuanced 6-way *Implicit_Hate*. On this task, only *COT-T5-XL* consistently scales with the increment in training data to reach the max F1 score of 0.56, while *Llama* models plateau at sub-0.3 range even at K=256.

We further select the best performing model at K=256 for each dataset and finetune their *Base* versions with the entire training data and no explanations no definition. In Figure 2b and Table 3, in-domain finetuning after *HateCOT* achieves competitive results even with only a fraction of the full training data. Furthermore, prompting for explanations enables *Llama 13B* and *COT-T5-XL* to attain performance that surpasses using the full training data on *Latent_Hate* and *Implicit_Hate*.

4.3 In-context Learning

As an alternative to in-domain finetuning, we investigate the models' performance using in-context learning (ICL), when a number of complete examples are provided as part of the input prompt. We select 1 sample from *each* class in the training data of each dataset, then obtain its the associated explanation from GPT-3.5-Turbo. The sets of post, explanation and label are arranged in the same format in the same template shown in Figure 5. We run inference for classification results over 5 seeds, which also randomly permutes their order.

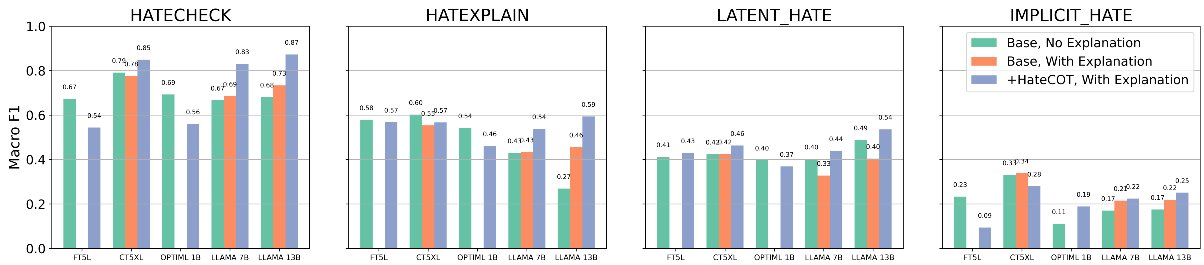
Figure 2c shows the mean, minimum and maximum values of macro F1 scores over the seeds for the base, *HateCOT*-pretrained only (*Pretrained)*, and in-domain finetuned (K=256) versions of *Llama 7B* and *Llama 13B*. *COT-T5-XL* regularly generates overly repetitive outputs, and thus omitted. The range of F1 scores is large regardless of settings, an observation in line with the variance of in-context learning in literature (Lu et al., 2022; Dong et al., 2022). Unsurprisingly, base models' performances tend to be inferior to their finetuned counterpart. Interestingly, the max F1 scores of finetuned models with ICL are not appreciably better than those in the zero-shot counterparts (Figure 2b). In contrast, except for *Llama 7B* on *HateXplain*, the best scores of pretrained models approach those of the finetuned models–particularly for *Llama 13B*.

This finding suggests another advantage of pretraining on *HateCOT*: boosting performance via ICL without additional in-domain finetuning, an area that has attracted growing attention (Min et al., 2021; Wang et al., 2023; Ye et al., 2023b). Nevertheless, there exists the trade-off: ICL examples with explanations extend significantly the context length, and ICL inferencing takes considerably more time compared to zero-shot, making the latter more resource-efficient overall.

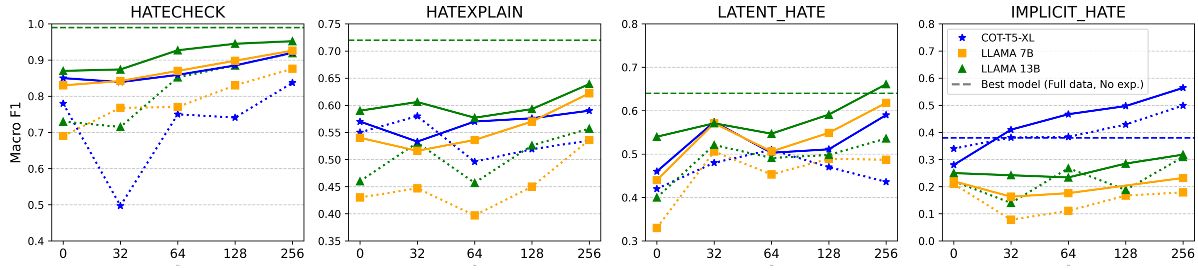
4.4 Assessment and Recommendations

From empirical observations, we make the following recommendations to construct a cost-efficient pipeline for classifier on novel domains:

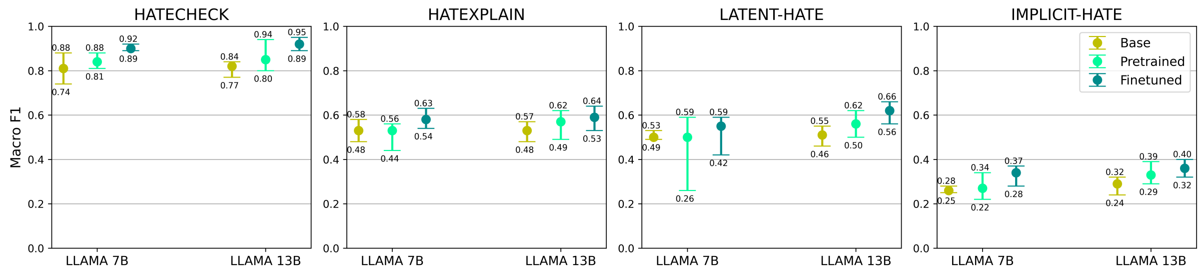
- ▷ The most consistent benefit of *HateCOT* is its capacity to enable data-efficient in-domain finetuning following pretraining.
- ▷ Practitioners should choose models of sufficient number of parameters for the task. Larger instruction-tuned LLMs appear to



(a) Macro F1 scores of LLMs in zero-shot setting using 3 configurations. *Base* refers to out-of-the-box models, *+HateCOT* denotes their pretrained counterpart on our dataset. *FT5L*: Flan-T5-L, *CT5XL*: COT-T5-XL. Results for Base Flan-T5-L and OPTIML models for *With Explanation* settings omitted to reflect their inability to generate explanation according to the prompt.



(b) Macro F1 scores for models in zero-shot setting with explanation after K-shot in-domain finetuning at various values of K. Dashed line represents finetuned base models, solid line represents models pre-trained on *HateCOT*, then in-domain finetuned. For each dataset, the horizontal dashed line represents the base version of the best performing model at K=256 which is finetuned using the entire training data *without* any rationale for comparison, denoted as *Best model (Full data, No. exp)*.



(c) Min, max and mean of macro F1 scores for Llama 7B and Llama 13B over 5 seed using in-context learning. *Pretrained* denotes models finetuned on *HateCOT* only. *Finetuned* denotes ICL performed on models that have been both *HateCOT*-pretrained then K=256 shot in-domain finetuned.

Figure 2: Performance resutls of LLMs on test sets in various settings.

Dataset	Best Model @K=256	F1 @K=256	F1 Base + Full	F1 %	Data Size @K=256	Data Size Full	Data %
HateCheck	LLAMA 13B	0.95	0.99	96%	512	1,864	27%
HateXplain	LLAMA 13B	0.64	0.72	89%	768	12,088	6%
Latent_Hate	LLAMA 13B	0.66	0.64	103%	768	11,460	7%
Implicit_Hate	COT-T5-XL	0.56	0.38	147%	1,536	2,707	57%

Table 3: Comparison of performance metrics for the best performing models finetuned using K=256 in-domain post *HateCOT* vs. finetuned on the full training set and no explanation nor definition. *F1%* denotes the percentage of macro F1 score of the K=256 finetuned model over that of the model trained on full data. Similarly, *Data %* denotes the percentage of data size used by the K=256 regimen over the full data.

more effectively capitalize on *HateCOT* pretraining regimen before in-domain finetuning.

▷ Instead of devoting resources to curate substantial training data, practitioners could focus on obtaining high quality annotations for representative rationales, and augment them into explanations using their LLM of choice. Alternatively, practitioners may choose to curate the explanations organically to achieve certain desired thematic qualities. This process may be iterated until targeted performance is reached according to some guiding metrics with acceptable quality of explanation.

5 Quality of Explanations

In addition to the enhanced classification prowess, we investigate whether pretraining on *HateCOT* also improves the quality of explanation LLMs. To this end, the following 2 human quality assurance (QA) experiments are conducted. In QA 1, we assess if human annotators prefer the explanations generated by the *base* or *HateCOT*-pretrained LLMs (*COT-T5-XL, Llama 7B, Llama 13B*). In QA 2, we perform in-domain K-shot finetuning on the aforementioned models and examine how annotators evaluate their generated explanations. An additional assessment of Target Identification is presented in Appendix A.6.

5.1 QA 1: Base vs. Pretrained

From the outputs of the test sets generated by the 3 LLMs, we select 50 samples uniformly at random whose explanations from the *Base* and *HateCOT*-pretrained versions agree on the predicted label, for a total of 150 samples and 300 explanations. We then recruit 13 annotators from the crowdsource platform Amazon Mechanical Turk and solicit their annotation on these explanations (Appendix A.5). Using the template in Figure 6, we ask the annotators to indicate their preferred explanation that better suits the purpose of content moderation based on fluency, soundness and the alignment with the definition of the chosen label. Each post is annotated by 5 humans, resulting in 750 annotations.

In Table 4, we observe that the raw frequency count for explanations generated by the *HateCOT*-pretrained models exceed their base version's. Similarly, even when tallying by majority vote–where the explanation is chosen by at least 3 out of 5 annotators–preference for those generated by the *Pretrained* models still prevails. We note that the preference margin is smaller for *Llama 13B Pretrained*, likely due to this model's already strong generative capabilities.

5.2 QA 2: Inter-model Comparison

Inspired by Wang et al. (2023); Lin et al. (2023); Yang et al. (2023), we assess the quality of explanations generated by finetuned models (K=256) on the following criteria:

- ▷ **Persuasiveness**: how convincingly the explanation justifies its chosen label for the post.
- ▷ **Soundness**: how valid and logical is the explanation with respect to the label's definition.

	Human (Frequency Count)		Human (Majority Vote)	
Model	**Base**	**Pretrained**	**Base**	**Pretrained**
COT-T5-XL	62 (24.8%)	188 (75.2%)	6 (12%)	44 (88%)
Llama 7B	109 (43.6%)	141 (56.4%)	19 (38%)	31 (62%)
Llama 13B	114 (45.6%)	136 (54.4%)	22 (44%)	28 (56%)

Table 4: Comparison of Base and Pretrained models in Human Evaluation. *Frequency Count* : count per annotation; *Majority Vote* indicates aggregate count by the version is preferred by at least 3 out of 5 annotators.

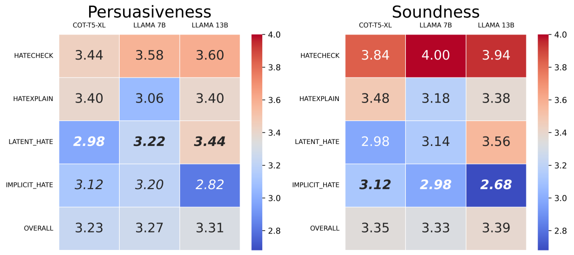


Figure 3: Heatmap for the average rating of explanations by finetuned *Model* (x-axis) and *Dataset* (y-axis) on 3 criteria from 1 (least) to 5 (very). *Overall* indicates average scores aggregated over all datasets. Triplets of scores *italicized* and in **bold** are those whose p-value < 0.05 by one-way ANOVA test that compare ratings of 3 models across the dataset on that row. *Italicized*-only scores indicate marginal significance (p-value $\approx$ 0.07).

For this QA task, we recruit 6 annotators also from Amazon Mechanical Turk (Appendix A.5). Using the template in Figure 7, we collect their numerical ratings on a scale from 1 (least) to 5 (very) on these criteria for 50 posts per model per dataset, for a total of 600 annotations. Figure 3 displays the mean ratings for each model-dataset pair, as well as *Overall* scores averaged across all datasets. The average *Ovearall* ratings for both *Persuasiveness* and *Soudness* are above 3.2 out of 5, indicating generally positive reception by human evaluators. Interestingly, there exists a degree of correlation between the models' better classification performance (Figure 2b) and higher mean ratings on each dataset, a useful artifact to calibrate models. These ratings may serve as benchmarks for future works.

6 Conclusion

We show empirically that our *HateCOT* dataset considerably enhances offensive speech detection even with limited training data while producing high-quality explanations. We invite future research to explore other benefits of LLM-augmented data and extend them to other related low-resource areas.

7 Limitations

We first acknowledge that our work is restricted to English corpora, a common limitation among literature on offensive speech (Yin and Zubiaga, 2021; Poletto et al., 2021). However, our approach sets a proof-of-concept for researchers to construct similar corpus in other languages by leveraging existing resources. Furthermore, our developmental pipeline is considerably more data-efficient than conventional approaches (Section 4.1, 4.2), potentially lowering the barrier of entry for practitioners without access to abundant resources. Therefore, this work invites further expansion on multilingual datasets, particularly to develop corpora with clearly defined definitions to facilitate synergy with other research.

Second, due to computational limitations, we could not perform experiments on larger open-source models. With the development of newer, more powerful models, it is reasonable to expect their performance to further improve though the use of this dataset of our corpus, as demonstrated by our empirical results.

Finally, we recognize the risk of propagating implicit biases that LLMs are known to carry (Cheng et al., 2023; Gupta et al., 2023). However, we note that the approach of using LLMs (GPT-3.5-Turbo in this paper) to bridge the logical gaps in original rationales has been shown to produce outputs less prone to logical failures (Yang et al., 2023). Biases in Pretrained Language Models have been attracting much attention in the research community. We invite further works to consider our approach to reduce hallucinations and biases in text generation.

8 Ethics Statement

We acknowledge the potential malicious usage of our corpus to generate content capable of evading detection, or jeopardizing classifiers' performance.

References

Marzieh Babaeianjelodar, Gurram Poorna Prudhvi, Stephen Lorenz, Keyu Chen, Sumona Mondal, Soumyabrata Dey, and Navin Kumar. 2022. Interpretable and high-performance hate and offensive speech detection. In *International Conference on Human-Computer Interaction*, pages 233–244. Springer.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Tommaso Caselli, Valerio Basile, Mitrovic Jelena, Granitzer Michael, et al. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25. Association for Computational Linguistics.

Camilla Casula and Sara Tonelli. 2020. Hate speech detection with machine-translated data: the role of annotation scheme, class imbalance and undersampling. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769. CEUR-WS. org.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

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

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021a. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021b. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

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.

Antigoni Founta and Lucia Specia. 2021. A survey of online hate speech through the causal lens. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 74–82, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anastasia Giachanou and Paolo Rosso. 2020. The battle against online harmful information: The cases of fake news and hate speech. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3503–3504.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Lawrence Han and Hao Tang. 2022. Designing of prompts for hate speech recognition with in-context learning. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–320. IEEE.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Muhammad Amien Ibrahim, Samsul Arifin, I Gusti Agung Anom Yudistira, Rinda Nariswari, Abdul Azis Abdillah, Nerru Pranuta Murnaka, and Puguh Wahyu Prasetyo. 2022. An explainable ai model for hate speech detection on indonesian twitter. *CommIT (Communication and Information Technology) Journal*, 16(2):175–182.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023a. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708, Singapore. Association for Computational Linguistics.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023b. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Anis Koubaa. 2023. Gpt-4 vs. gpt-3.5: A concise showdown.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

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, pages 14867–14875.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Huy Nghiem, Umang Gupta, and Fred Morstatter. 2024. "define your terms" : Enhancing efficient offensive speech classification with definition. In *Proceedings of the 18th Conference of the European Chapter of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1309, St. Julian's, Malta. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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 4755–4764.

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.

Heng Rathpisey and Teguh Bharata Adji. 2019. Handling imbalance issue in hate speech classification using sampling-based methods. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 193–198. IEEE.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

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.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, pages 255–264.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

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*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fbert: A neural transformer for identifying offensive content. *arXiv preprint arXiv:2109.05074*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

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.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, 15(12).

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. 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.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* *and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*.

Sargam Yadav, Abhishek Kaushik, and Kevin McDaid. 2023. Hate speech is not free speech: Explainable machine learning for hate speech detection in code-mixed languages. In *2023 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–8. IEEE.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023a. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2023b. In-context instruction learning. *arXiv preprint arXiv:2302.14691*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. Offenseval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*, 29(6):1416–1435.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendix

A.1 Data Pre-Processing

The datasets used in this work are released by their respective authors for research purpose. Aware of the risk of containing confidential in social media data, we anonymize posts during the curation process by replacing user handles with the string '<user>'. Our many layers of randomization provides further protection with respect to privacy.

A.2 Technical Specification

A.2.1 Inferencing

We used OpenAI's API to access the publicly available version of GPT-3.5 in November 2023, and GPT-4 in January 2024. To obtain explanations from the former (as described in 3.2.2), we choose temperature among candidates $\{0.3, 0.5, \mathbf{0.7}\}$ and settle on the last value during inference. This value is selected based on literature and multiple iterations of qualitative analysis of outputs (Yang et al., 2023; Kim et al., 2023a). For GPT-4 and other open-source models, we use greedy decoding.

A.2.2 Finetuning

To train models, we employ both full supervise finetuning (*FLAN-T5-Large, OPT-IML*) and LoRA parameter-efficient techniques (all other models). LoRA models set to 8-bit quantization using the BitsandByes library. Training *FLAN-T5-Large, OPT-IML, LLAMA 7B* models was done on 2 Nvidia RTX A6000 GPUs, whereas *COT-T5-XL, Llama 13B* used 4 GPUs. Hyperparameters for the following candidates are tuned on the validation set of sampling K=64 shots from the leftover training samples by optimizing macro F1-score metric. Options in bold indicate final chosen values among multiple across all models to finetune on *HateCOT*.

- Learning rate: {5e-5, 1e-4, **3e-4**}
- Training Epochs: {1, 2, **3**}
- LoRA Rank: {16, 32, **64**} (*alpha*=rank*2)
- LoRA Target Modules: {Q, V}
- Batch size: 2
- Gradient Accumulation Step: 2

For in-domain K-shot finetuning, the values above remain the same except for the following variations in Learning Rates, which is set to 1e-4 for *HateCheck, HateXplain, Latent_Hate*, and 3e-4 otherwise.

A.3 Variants of Prompt Template for Explanation

The 8 datasets introduced in Section 3.1 may not always have annotations on all required fields; thus, we modify the first sentence in the *Instruction* block in Figure 5 with the following variants when appropriate:

- Only *Target* is available: *'Provide a brief paragraph to explain step-by-step how the post targets the specified group or entity, and how that leads to the specified Label based on the given Definitions.'*
- Both *Target* and *Rationale* are available: *'By elaborating on the provided Annotation, provide a brief paragraph to explain step-by-step how the post targets the specified group or entity, and how that leads to the specified Label based on the given Definitions.'*

A.4 Quality Review of Sample Explanations

Elaborating on the criteria outlined in Section 3.1, we review the quality of the generated explanation by GPT-3.5-Turbo on the following items:

- Grammatically correct
- Succinct in their justification of the chosen label.
- Persuasive and logical in their reasoning for the chosen label

We discard explanations that are too verbose, and/or not choosing the label already provided by human annotation (fortunately, this scenario happens rarely, likely due to the presence of existing rationales guiding the extra generated outputs). We also remove explanations that conjure incorrect and/or irrelevant facts to the context to discourage hallucination and encourage a high degree of pertinent to the post at hand.

Dataset	COT-T5-XL	Llama 7B	Llama 13B
HateCheck	94.5 %	96.5 %	93.5 %
HateXplain	71.0 %	74.0 %	74.0 %
Latent_Hate	50.5 %	51.0 %	44.5 %

Table 5: Percentage out of 200 samples per dataset, where explanations correctly identify at least 1 of the targets listed by human annotators by each model.

A.5 Human Annotation for QA Experiment

A.5.1 QA 1

With approved IRB, we recruit 13 crowdsource workers using the Amazon Mechanical Turk platform to annotate 50 samples per model, for a total of 150 data points for the task described in Section 5.1. The annotators was paid a fair wage at $15 per hour, and forewarned about the nature of the task. Annotators must be fluent English speakers. We also limit each annotator to no more than 100 posts (60% of the total 150 samples per model) to maintain diversity of opinions. We observe that preference for explanations generated by *Pretrained* models remains consistent with GPT-4's.

The demographic breakdown of the 13 annotators are described below:

- **Gender**: Female (8), Male (5)
- **Age**: 18-29 (2), 30-39 (4), 40-49 (4), 50+ (3)
- **Education**: High School (2), 2-year college (5), 4-year college (4), Master's or Higher (2)

A.5.2 QA 2

For this experiment, each annotation task consists of the explanations generated by 3 models are grouped by the sample. This division resutls in 200 tasks. 6 Amazon Mechanical Turk workers are recruited, with similar qualification criteria as described above.

The demographic breakdown of the 6 annotators are described below:

- **Gender**: Female (3), Male (3)
- **Age**: 18-29 (1), 30-39 (2), 40-49 (3)
- **Education**: 2-year college (2), 4-year college (4)

A.6 QA 3: Target Identification

We investigate the in-domain finetuned models' capabilities to identify the target of the sentiments expressed by the post. We also randomly select 200 samples from each dataset that have the *Target* variable annotated by humans, then ask GPT-4 to judge whether the explanations from the models mention at least one of the listed targeted groups using the template in Figure 8. Note that we encourage GPT-4 to consider potential variance of expression and not restrict to exact matches.

Table 5 shows the percentage of accuracy on this task. Similar levels of descending performances are observed in the presented order of test datasets. However, this observation may be an artifact of the differences between the annotations targets among datasets: *HateCheck* has a limited number of discrete categories while *Latent_hate* contains multiple combinations of free-text labels. We urge practitioners to consider this factor while curating training data if target identification is desired.

Dataset	Definition
Salminen et al., 2018	**Neutral** : A post that is not offensive to any group of people. **Hateful**: An offensive post, motivated, in whole or part, by the writer's bias against an aspect of a group of people.
Qian et al., 2019	**Not Hate** : Does not contain any attack on people based on protected characteristics. **Hate**: Contains direct attack on people based on protected characteristics such as race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.
Sap et al., 2019	**Not Offensive** : not offensive to anyone. **Offensive** : denotes the overall rudeness, disrespect, or toxicity of a post. whether a post could be considered offensive to anyone.
Vidgen et al., 2021a	**Neutral**: Content that does not fall into other categories, usually entirely unrelated to abuse, hate, prejudice, or intolerance. **Identity Directed Abuse**: Content that directs abuse at an identity, which relates to fundamental aspects of individuals' social position, community and self-representation. An identity includes but is not limited to religion, race, ethnicity, gender, sexuality and sexual preferences, immigration status, nationality, ableness, physical appearance and class. **Affiliation Directed Abuse** : Content that directs abuse at people who have a voluntary affiliation with a profession, membership, association, ideology, or other well-defined group or collective. **Person Directed Abuse** : Content that directs abuse at an identifiable person.
Vidgen et al., 2021b	**Derogation** : content which explicitly attacks, demonizes, demeans or insults a group. **Animosity** : content which expresses abuse against a group in an implicit or subtle manner. **Threatening** : content which expresses intention to, support for, or encourages inflicting harm on a group, or identified members of the group. **Support For Hateful Entities** : content which explicitly glorifies, justifies or supports hateful actions, events, organizations, tropes and individuals collectively, entities. **Dehumanization** : content which perceives or treats people as less than human. **Not Hate** : content that falls into none of the other categories.
Basile et al., 2019	**Hate Speech** : language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate,or to insult the members of the group. may also be language that threatens or incites violence. **Offensive Language** : may contain offensive terms but targets disadvantaged social groups in a manner that is potentially harmful to them. **Neither** : language that does not all into either of the other categories.
Hartvigsen et al., 2022	**Benign**: Text that is not harmful nor offensive to anyone. **Toxic**: Text that could be seen as harmful to anyone, may contain offensive, rude humor, insults, personal attacks, profanity, aggression, may refer to targeted group with harmful intent that is expressed in stereotypes or lewd manners
Toraman et al., 2022	**Hate** : target, incite violence against, threaten, or call for physical damage for an individual or a group of people because of some identifying trait or characteristic. **Offensive** : humiliate, taunt, discriminate, or insult an individual or a group of people in any form, including textual. **Normal** : does not fall into any of the other categories.

Table 6: Labels and definitions of 8 datasets used to synthesize training corpus. Definitions are lifted directly from their original works, with exceptions of minor adjustment for stylistic consistency.

Dataset	Definition
HateCheck	**Non-hateful**: Language that does not contain any abusive sentiment targeted at a protected group. **Hateful**: Abuse that is targeted at a protected group or at its members being a part of that group. Protected groups include age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation.
HateXplain	**Normal** : neither hate speech nor offensive. **Hate Speech** : language which attacks, demeans, offends, threatens, or insults a group based on race, ethnic origin, religion, disability, gender, age, sexual orientation, or other traits. it is not the presence of certain words that makes the text hate speech, rather you should look the context the word is used in the text. **Offensive Language** : usage of rude, hurtful, derogatory, obscene or insulting language to upset or embarrass people.
Latent_Hate	**Not Hate** : speech or actions that do not involve any form of hatred, prejudice, or discrimination toward individuals or groups based on their characteristics. **Explicit Hate**: openly expressed, direct forms of hatred and prejudice toward individuals or groups based on their characteristics. **Implicit Hate**: coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity.
Implicit_hate	**White Grievance** : includes frustration over a minority groups perceived privilege and casting majority groups as the real victims of racism. This language is linked to extremist behavior and support for violence. **Incitement To Violence** : includes flaunting in group unity and power or elevating known hate groups and ideologies. **Inferiority Language** : implies one group or individual is inferior to another, and it can include dehumanization, denial of a person's humanity , and toxic language that compares the target with disease, insects, animals . Related to assaults on human dignity, dominance, and declarations of superiority of the in group. **Irony** : refers to the use of sarcasm , humor, and satire to attack or demean a protected class or individual. **Stereotypes And Misinformation** : associate a protected class with negative attributes such as crime, or terrorism. includes misinformation that feeds stereotypes and vice versa, like holocaust denial and other forms of historical negationism. **Threatening And Intimidation** : conveys a speaker's commitment to a target's pain, injury, damage, loss or violation of rights, threats related to implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation.

Table 7: Labels and definitions for 4 test datasets.

Post: "I am so fucking annoyed with these dumb ass Africans, I really can just beat all they asses"

Label: Hate

Target: N/A

Rationale: "Targeting people based on Race"

GPT-enhanced explanation: This post should be classified as Hate based on the given definitions because it contains language that specifically targets and insults a group of people based on their race. The use of derogatory language and the call for physical harm demonstrates the intention to incite violence and discrimination against individuals of African descent.

(a) Example from Toraman et al. (2022)

Post: "Wake up you white American TRASH and admit you all are a bunch of RACIST MOTHERFUCKERS."

Label: Hateful

Target: "Specific nation(s)/ white Americans"

Rationale: N/A

GPT-enhanced explanation: The post should be classified as "Hateful" based on the given definitions because it contains offensive language and is motivated by the writer's bias against white Americans, using derogatory terms to insult and generalize them as racist individuals.

(b) Example from Salminen et al. (2018)

Post: "So many b***** think they're so much hotter than they are"

Label: Offensive

Target: "Women"

Rationale: "women are b****** | women are stuck up"

GPT-enhanced explanation: The post targets women by making generalizations and perpetuating negative stereotypes about them. It suggests that women only date attractive men and implies that they are superficial and mean. By making these offensive statements, the post disrespects and devalues women, leading to the offensive label.

(c) Example from Sap et al. (2019)

Figure 4: Examples drawn from our training corpora showing their native *Post, Target and Rationale*, along with the corresponding GPT-3.5-Turbo-enhanced explanations. Due to their nature as fragmented annotations, verbatim *Rationales* are not serviceable explanation, but can serve as guiding signals that leverage GPT's generative capabilities to construct legible passages with detailed justifications.

```
### Instruction:
Perform this task by considering the following Definitions.
Based on the message, label the input as only one of the
following categories : Non-hateful, or Hateful.
Provide a brief paragraph to explain step-by-step why the
post should be classified with the provided Label based
on the given Definitions. If this post targets a group or
entity relevant to the definition of the specified Label,
explain who this target is and how that leads to that
Label.
Append the string '<END>' to the end of your response.
Provide your response in the following format:
EXPLANATION: [text]
LABEL: [text] <END>
### Definitions:
Hateful : Abuse that is targeted at a protected group or at
its members being a part of that group. Protected groups
include...
Non-hateful : Language that does not contain any abusive
sentiment targeted at a protected group.
### Input:
Do you not realize how hateful you are being when calling
*** people trash?
### Response:
```

Figure 5: Template used to prompt LLM for classification inference.

```
### Instruction:
Given a post, a label and its corresponding definition, you are presented 2
explanations denoted EXP_A and EXP_B. Each explanation is a short passage that
justifies the decision to select the label based on the content of the post.

Please act as an impartial judge to determine which explanation is better suited to
be used for content moderation on a social media platform.

Please consider the fluency, soundness, and the appropriateness of each
explanation's justification to select the label based on the label's definition in your
evaluation.
You must select either EXP_A or EXP_B, but never both. Answer only with your
choice. Do not provide any extra justification. Do not let the order in which the
passages are presented, or the length of the explanations, influence your decision.
### Post:
All trans people like them are vile.
### Definitions:
Hateful : Hateful : Abuse that is targeted at a protected group or at its members
being a part of that group. Protected groups include age, disability, gender
identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or
sexual orientation.
### EXP_A:
The post should be classified as Hateful based on the given definitions because it
contains abusive sentiment targeted at a protected group, specifically transgender
individuals. The statement "All trans people like them are vile" is derogatory and
offensive, expressing a negative and discriminatory view towards transgender
individuals. This language is offensive and disrespectful, making it a clear example
of hate speech. LABEL: Hateful
### EXP_B:
The given post contains a derogatory term for trans people and is therefore
hateful. LABEL: Hateful
### Evaluation:
```

Figure 6: Template for QA Experiment 1. In this example, EXP_A is preferred.

Instruction:
You are presented a post, a set of labels and their corresponding definitions, and an explanation that justifies the decision to select a label based on the relevance between the post and that label's definition. This explanation may be used to facilitate content moderation on a social media platform.

Please act as an impartial judge to evaluate the quality of the explanation on 3 aspects: Informativeness, Persuasiveness, and Soundness. Criteria for these aspects are described below:

Persuasiveness: Rate the explanation on whether it appears convincing in its justification of the chosen label for the post. Use a rating scale that ranges from 1 (not persuasive) to 5 (very persuasive).

Soundness: Rate the explanation on whether the explanation seems valid and logical with respect to its justification to select that label. Use a rating scale that ranges from 1 (not sound) to 5 (very sound).

Post:
this is just another good reason for white southerners to quit giving their money to the nfl

Labels and Definitions:
Incitement : includes flaunting in-group unity and power or elevating known hate groups and ideologies.
Inferiority : implies one group…

Explanation:
The post targets white southerners by suggesting that they should stop giving their money to the NFL. This implies that white southerners are being influenced or influenced by the NFL, which is a known hate group and ideology. By suggesting that white southerners should stop supporting the NFL, the post incites division and promotes a sense of in-group unity and power. This aligns with the definition of incitement, as it flaunts in-group unity and power while also elevating a known hate group and ideology. LABEL: Incitement

○ 1-Not Persuasive ○ 2 ○ 3-Persuasive ○ 4 ○ 5-Very Persuasive

○ 1-Not Sound ○ 2 ○ 3-Sound ○ 4 ○ 5-Very Sound

Figure 7: Template for QA Experiment 2.

Instruction:
Given a post, a label and its definition, you are presented an explanation that justifies the decision to select that label based on the relevance between the post and the label's definition. In addition, you are also presented a target, which may refer to one or multiple entities.

Please act as an impartial judge to perform the following.

Answer YES if the explanation mentions the target, or at least one of the targets if there are multiple, and generally describes this target or targets as the receiving entity of some sentiment expressed by the associated post.
If the explanation refers to a sub-category of at least one of the targets, it is also considered correct.
Answer NO otherwise.

Please note that it is not necessary for the target to be mentioned verbatim in the explanation. As long as the entity is referenced in a way that conveys its meaning, consider it acceptable. Feel free to recognize variations in wording that capture the core idea of the target rather than requiring an exact match.

Post:
at least 60 white south africans were …

Label and Definition:
Implicit Hate:

Target:
Black people

Explanation:
The post explicitly targets black people by making a generalization that they are racist. This statement indirectly disparages black people based on their race, which falls under the definition of implicit hate. LABEL: Implicit Hate

Evaluation:

Figure 8: Template for QA Experiment 3.