HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning

Yongjin Yang∗ Joonkee Kim∗ Yujin Kim∗ Namgyu Ho
James Thorne† Se-young Yun†
KAIST AI
{dyyjkd, joonkeekim, yujin399, itsnamgyu, thorne, yunseyoung}@kaist.ac.kr

Abstract

With the proliferation of social media, accurate detection of hate speech has become critical to ensure safety online. To combat nuanced forms of hate speech, it is important to identify and thoroughly explain hate speech to help users understand its harmful effects. Recent benchmarks have attempted to tackle this issue by training generative models on free-text annotations of implications in hateful text. However, we find significant reasoning gaps in the existing annotations schemes, which may hinder the supervision of detection models. In this paper, we introduce a hate speech detection framework, HARE, which harnesses the reasoning capabilities of large language models (LLMs) to fill these gaps in explanations of hate speech, thus enabling effective supervision of detection models. Experiments on SBIC and Implicit Hate benchmarks show that our method, using model-generated data, consistently outperforms baselines, using existing free-text human annotations. Analysis demonstrates that our method enhances the explanation quality of trained models and improves generalization to unseen datasets. Our code is available at https://github.com/joonkeekim/hare-hate-speech.git.

1 Introduction

The increase in the use of online media has intensified the exposure to hate speech, prompting the need for effective detection systems (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). While early works have been limited to the classification of explicit hate speech (Caselli et al., 2020; Mathew et al., 2021), recent works have drawn our attention to implicit forms of hate speech which are more prevalent, yet subtle. (Jurgens et al., 2019).

To tackle these nuanced forms of hate speech, it is important for systems to not only identify hate speech but also provide interpretable explanations (Liu et al., 2019). This can help mitigate distributional biases inherent in simple classification, allowing people to understand and reason about the potential harms of hateful text (Sap et al., 2019b). Explanations can also improve the transparency of content moderation on social media (Gillespie, 2018).

Recent works on hate speech understanding (Sap et al., 2019b; ElSherief et al., 2021; Huang et al., 2022) have considered training autoregressive language models to generate underlying explanations on hate speech. The models are trained on human-written free-text rationales such as implied statements and targeted groups. However, despite the use of novel benchmark datasets, i.e., SBIC (Sap et al., 2019b) and Implicit Hate (ElSherief et al., 2021), the trained models struggle to generate detailed and comprehensive explanations. Moreover, we observe that the provided rationales give marginal improvement to detection performance under joint training.
A potential cause of the limited supervision provided by existing annotations on understanding and explaining hate speech may be the existence of critical gaps in reasoning. For example, as shown in Figure 1, the implied statement of the post “*How dark is my humour? It picks cotton*” is annotated as “*black folks are slaves*”, in SBIC. To understand this implication, one must understand that “dark” implies “black folks”, and the phrase “picks cotton” relates to the historical background of African Americans. While this may be obvious to human annotators, language models are known to lack societal knowledge and commonsense reasoning skills to understand these nuances (Talmor et al., 2019; Li et al., 2022; Choi et al., 2023). This leaves a significant gap between the training objectives of classification and generating annotated implications, which may harm supervision (Wiegrefe et al., 2021b; Wang et al., 2023a).

Drawing inspiration from the reasoning capabilities of large language models (LLMs) improved with chain-of-thought (CoT) reasoning (Wei et al., 2022), we present our novel approach “Explainable Hate Speech Detection with Step-by-Step REasoning (HARE)”. We leverage LLM-generated free-text rationales using CoT prompts to fill in the gaps of reasoning in existing hate speech annotations and enhance supervision of generative detection models. To create these rationales, we propose two approaches: (1) adopt CoT prompts to create comprehensive rationales that align with the given texts and (2) incorporate existing human annotations from benchmarks in the CoT prompts to bridge the logical gap between the input text and human annotations. When tested on the challenging SBIC and Implicit Hate datasets, our approach outperforms standard fine-tuning with given human annotations. When tested on the challenging SBIC and Implicit Hate datasets, the task can also be formulated to sequentially output $C$, $T$, and then $I$, as $p((C; T); I)|P)$. Hence, the model trained with human-written annotations is designed to generate $C$ and then provide explanations using annotations $T$ and $I$. However, due to a logical gap between the speech $P$ and the annotations $T$ and $I$, training a model with these annotated rationales does not significantly enhance the model’s ability to comprehend hate speech.

Chain-of-thought prompting (Wei et al., 2022) refers to a prompting method that generates a chain of reasoning to derive answers. Kojima et al. (2022) introduce a method of including the phrase “Let’s think step by step” at the end of the prompt to generate reasoning in a zero-shot setting.

### 2.2 Fr-HARE & Co-HARE

In this paper, we propose HARE, a framework that fine-tunes using rationales generated by LLMs through CoT prompts. By leveraging these rationales, the model is trained to discern whether a text is classified as hate speech or not. This approach of utilizing the LLM-generated rationales not only improve the model performance as in previous works (Ho et al., 2022; Wang et al., 2023a) but also enables the model to generate more plausible explanations. Subsequently, we introduce two variants that employ different methods for extracting rationales from LLMs.

**Fr-HARE** Our first strategy, Free-HARE (Fr-HARE), extracts rationales using LLM from the LLM without referencing the annotated rationales. In line with Kojima et al. (2022), we employ zero-shot prompts to prompt the LLM to extract rationales, as follows:

Determine whether the following post is offensive, and explain why.
Post: $P$
Answer: Let’s explain step by step.

From the given prompt above, both the rationale $R$ without any logical leap and the classification result $C$ are generated. Then, we utilize the generated rationale $R$ by training the model to sequentially generate $C$ and $R$, represented as $p((C; R)|P)$. Utilizing these rationales enhances the model’s ability to comprehend the context of hate speech, thus leading to explanations that are more intelligible to humans.

---

1We refer to the “offensive” category in SBIC as “hate” class. In the experiment, we use the wording “offensive” as is.
Co-HARE We also propose a second strategy, Conditioned-HARE (Co-HARE), which utilizes human-written rationales to formulate CoT prompts. Fr-HARE generates complete rationales entirely from scratch, disregarding human annotations included in the datasets. Although Fr-HARE offers the advantage of being universally applicable to any hate speech dataset, the LLM-generated rationales might lack consistency due to wide variation in reasoning sequences. To address this inconsistency, Co-HARE integrates human-written rationales about target groups $T$ and implied statements $I$ into the CoT prompt as follows:

Then, we train the model using extracted rationales, as in Fr-HARE. For samples labeled as “not hate” that do not include human-written rationales, we apply the prompt used in Fr-HARE. While Co-HARE requires human-written rationales, it generates rationales that are more tailored to the specific requirements and features of the dataset, due to its guided nature. Therefore, the model trained with Co-HARE can provide explanations that align more closely with the forms of rationales that humans construct.

Details of HARE Once we have extracted the rationales from the LLMs, we follow the approach of Kojima et al. (2022) to have the LLMs predict the class. Specifically, we employ a two-stage extraction process. In the first stage, we extract both the class $C$ and the rationale $R$ from the LLMs using our HARE method, represented as $p(C; R|P)$, as previously outlined. In the second stage, we prompt the LLMs again, this time to predict the class $C$ given the extracted rationales $R$ and the post $P$, denoted as $p(C|R, P)$. During fine-tuning on hate speech datasets, if the predicted class $C$ coincides with the true answer $C$, we concatenate $C$ with the extracted rationale $R$. If the predicted labels are incorrect, the models are solely trained to predict the class $C$. Furthermore, following the findings of Ho et al. (2022), we generate multiple distinct rationales to facilitate the learning process.

3 Experiments

3.1 Experimental Setup

We utilize SBIC and Implicit Hate datasets for our fine-tuning experiments. Our models are trained to classify the offensiveness and hatefulness of posts, using SBIC and Implicit Hate, respectively. It is noteworthy that in our Implicit Hate experiments, we combine both the explicit and implicit hate classes into a single “hate” category. We set up baselines with two families of models: $C$, a model trained exclusively for classification, and $C+T+I$, a model trained using human-written rationales. For Fr-HARE and Co-HARE, by using gpt-3.5-turbo-0613 that is known for its reasoning capabilities (Ouyang et al., 2022), we extract four and eight different rationales per each sample in SBIC and Implicit Hate, respectively, following the hyperparameter setting of Ho et al. (2022). Subsequently, we fine-tune the model, setting LLM-generated rationales $R$ and class $C$ as target sequence. For performance evaluation, we measure detection accuracy and compute the F1 score of classification, regarding “hate” as the positive class. We make use of Flan-T5 (Wei et al., 2021) with different model configurations: small, base, and large. We also conduct experiments using the large models of T5 (Raffel et al., 2020) and GPT-2 (Radford et al., 2019). A more detailed explanation of our experimental setup can be found in Appendix B.

3.2 Results and Discussions

Do LLM-generated rationales improve detection performance? Table 1 presents the performance of hate speech detection according to different methods on the SBIC and Implicit Hate datasets. Our strategies Fr-HARE and Co-HARE consistently exhibit superior performance over other baseline methods, regardless of the model size. This suggests that even though the baseline method is trained using human-written rationales, the more detailed and logically-sequenced LLM-generated rationales of HARE can further aid the model in understanding the input text and accurately classifying it as hate speech. Therefore, the results demonstrate that the quality of rationales has a strong impact on classification. Furthermore, the performance of our method consistently improves as the model size increases, in contrast to baselines. This suggests that diverse reasoning becomes increasingly beneficial as scale grows. This notable im-

---

**Table 1: Performance of hate speech detection**

<table>
<thead>
<tr>
<th>Method</th>
<th>SBIC</th>
<th>Implicit Hate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr-HARE</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>Co-HARE</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.80</td>
<td>0.75</td>
</tr>
</tbody>
</table>

---

**Details of Prompt**: Determine whether the following post is offensive, and explain why. Provide a brief, step-by-step explanation of how the post targets the specified group and how it leads to the implied statement provided.

**Prompt**:

Post: $P$
Target: $T$
Implied statement: $I$

Answer: Let’s explain step by step.
Table 1: The performance of fine-tuning on SBIC and Implicit Hate dataset with various models and size.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>SBIC Acc</th>
<th>SBIC F1</th>
<th>Implicit Hate Acc</th>
<th>Implicit Hate F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3.5-turbo-0613</td>
<td>ZS</td>
<td>80.06</td>
<td>81.75</td>
<td>73.58</td>
<td>65.66</td>
</tr>
<tr>
<td></td>
<td>CoT</td>
<td>73.48</td>
<td>79.07</td>
<td>73.98</td>
<td>67.19</td>
</tr>
<tr>
<td>Flan-T5 small</td>
<td>C</td>
<td>82.56</td>
<td>84.05</td>
<td>77.58</td>
<td>71.98</td>
</tr>
<tr>
<td></td>
<td>C+T+I</td>
<td>82.99</td>
<td>84.05</td>
<td>77.63</td>
<td>72.39</td>
</tr>
<tr>
<td></td>
<td>Fr-HARE</td>
<td>84.18</td>
<td>85.18</td>
<td>79.33</td>
<td>73.29</td>
</tr>
<tr>
<td></td>
<td>Co-HARE</td>
<td>84.44</td>
<td>85.35</td>
<td>78.54</td>
<td>73.49</td>
</tr>
<tr>
<td>Flan-T5 base</td>
<td>C</td>
<td>82.35</td>
<td>83.71</td>
<td>78.03</td>
<td>72.17</td>
</tr>
<tr>
<td></td>
<td>C+T+I</td>
<td>82.54</td>
<td>84.41</td>
<td>79.77</td>
<td>73.15</td>
</tr>
<tr>
<td></td>
<td>Fr-HARE</td>
<td>84.20</td>
<td>85.46</td>
<td>79.84</td>
<td>74.84</td>
</tr>
<tr>
<td></td>
<td>Co-HARE</td>
<td>84.65</td>
<td>85.76</td>
<td>80.38</td>
<td>75.69</td>
</tr>
<tr>
<td>Flan-T5 large</td>
<td>C</td>
<td>81.70</td>
<td>82.84</td>
<td>78.42</td>
<td>72.92</td>
</tr>
<tr>
<td></td>
<td>C+T+I</td>
<td>83.48</td>
<td>83.70</td>
<td>80.14</td>
<td>73.10</td>
</tr>
<tr>
<td></td>
<td>Fr-HARE</td>
<td>85.21</td>
<td>86.16</td>
<td>80.49</td>
<td>74.62</td>
</tr>
<tr>
<td></td>
<td>Co-HARE</td>
<td>84.93</td>
<td>85.57</td>
<td>81.49</td>
<td>76.71</td>
</tr>
<tr>
<td>T5 large</td>
<td>C</td>
<td>83.03</td>
<td>83.53</td>
<td>78.79</td>
<td>72.50</td>
</tr>
<tr>
<td></td>
<td>C+T+I</td>
<td>84.23</td>
<td>85.21</td>
<td>79.61</td>
<td>73.80</td>
</tr>
<tr>
<td></td>
<td>Fr-HARE</td>
<td>85.27</td>
<td>86.32</td>
<td>81.61</td>
<td>75.59</td>
</tr>
<tr>
<td></td>
<td>Co-HARE</td>
<td>85.35</td>
<td>85.93</td>
<td>80.98</td>
<td>75.88</td>
</tr>
<tr>
<td>GPT-2 large</td>
<td>C</td>
<td>81.39</td>
<td>82.68</td>
<td>73.32</td>
<td>66.68</td>
</tr>
<tr>
<td></td>
<td>C+T+I</td>
<td>82.80</td>
<td>83.43</td>
<td>75.95</td>
<td>65.25</td>
</tr>
<tr>
<td></td>
<td>Fr-HARE</td>
<td>83.92</td>
<td>85.48</td>
<td>78.47</td>
<td>71.35</td>
</tr>
<tr>
<td></td>
<td>Co-HARE</td>
<td>84.64</td>
<td>85.67</td>
<td>80.07</td>
<td>71.58</td>
</tr>
</tbody>
</table>

Table 2: Cross Evaluation results on HateXplain (Mathew et al., 2021) and DynaHate (Vidgen et al., 2020). We utilize Flan-T5-large fine-tuned on SBIC using each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>HateXplain Acc</th>
<th>HateXplain F1</th>
<th>DynaHate Acc</th>
<th>DynaHate F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>64.40</td>
<td>74.18</td>
<td>64.35</td>
<td>67.41</td>
</tr>
<tr>
<td>C+T+I</td>
<td>68.84</td>
<td>74.52</td>
<td>64.72</td>
<td>67.41</td>
</tr>
<tr>
<td>Fr-HARE</td>
<td>70.69</td>
<td>78.91</td>
<td>68.06</td>
<td>75.15</td>
</tr>
<tr>
<td>Co-HARE</td>
<td>71.62</td>
<td>78.52</td>
<td>69.98</td>
<td>75.01</td>
</tr>
</tbody>
</table>

Figure 2: The result of GPT-4 evaluation following Zheng et al. (2023). The bar and line represent the average scores that range from 1 to 10 and 95% confidence interval, respectively. We utilize Flan-T5-large fine-tuned on SBIC using each method.

(a) GPT-4 Single answer grading

(b) GPT-4 Pairwise comparison

**Does HARE improve the quality of generated explanations?** To evaluate the quality of the generated explanations, we conduct automated evaluation using GPT-4 (OpenAI, 2023) which has demonstrated its high alignment with human evaluation (Liu et al., 2023; Zheng et al., 2023). We randomly sample 50 “hate” instances from SBIC test set where all methods are correct and apply single-answer grading (Zheng et al., 2023) on explanations generated by models trained with C+T+I, Fr-HARE, and Co-HARE. Detailed explanation about GPT-4 evaluation can be found on Appendix C.

The results depicted in Figure 2 (a) show that both Fr-HARE and Co-HARE are capable of generating higher quality explanations, with Fr-HARE exhibiting slightly superior performance, suggesting that the flexibility of Fr-HARE leads to a more qualified explanation.
Post: I like my women like I like my cigars 7 years old and coming from cuba in a burlap sack.
(A) Offensive
Label: offensive
Annotation: girls are not worthy of equal life / girls are property / girls are used for sex / trivializes harm to victims
C/T+I
The answer is: (B) Not offensive
Fe-HARE
The answer is: (A) Offensive. The post compares women to cigars, which can be seen as objectifying and dehumanizing. 2. It also makes a reference to a specific age group, which can be seen as inappropriate and offensive. 3. (…)
The answer is: (A) Offensive 1. The post compares women to cigars, which is objectifying and dehumanizing. (…)
Fr-HARE
The answer is: (A) Offensive 1. The post starts with a quote from Anne Frank. But lately she’s been proven to have not said anything for almost a century.
(A) Offensive
Label: offensive
Annotation: jewish people were rounded up and killed in the holocaust / jewish people were victims of a genocide
C/T+I
The answer is: (B) Not offensive
The answer is: (A) Offensive. The post starts with a quote from Anne Frank, a Jewish girl who died during the Holocaust. 2. The quote itself is a statement about her lack of speaking much. (…)
Fr-HARE
The answer is: (A) Offensive. The post starts with a quote from Anne Frank, a Jewish girl who wrote a diary during the Holocaust. 2. (…) 3. (…)
The answer is: (A) Offensive 1. The post starts with a quote from Anne Frank, a Jewish girl who wrote a diary during the Holocaust. (…)
Co-HARE
The answer is: (A) Offensive 1. The post starts with a quote from Anne Frank, a Jewish girl who died during the Holocaust. This quote targets Holocaust victims and Jewish folks by making a joke about Anne Frank’s lack of speaking for almost a century. It implies that she has been silent or silent about her experiences during the Holocaust. This trivializes the harm and suffering that Holocaust victims endured. (…)

Figure 3: Model outputs using baseline methods and our framework HARE. The samples are from SBIC test set. Note that the answers are abbreviated by (…), and the full context is reported in Appendix D.

As Co-HARE is designed to create rationales that align better with human-written rationales, we also conduct pairwise comparison between Co-HARE and C+T+I to determine which method aligns better with human-written labels. Figure 2 (b) clearly shows that the rationales generated by the model trained utilizing Co-HARE are aligned more to actual human-written rationales than the ones generated by the model trained directly with human-written rationales. This also highlights the value of Co-HARE, as it aligns better with natural intuition, indicating that both Fr-HARE and Co-HARE can be utilized for different purposes.

How does HARE detect and contextualize hate speech? Figure 3 presents model outputs using HARE and baseline method. A more detailed qualitative study of randomly selected samples is available in Appendix D. In the first sample, the post makes light of harming a young girl using phrases like “I like like my cigar” and “burlap sack”. While our model, Fr-HARE, does not explicitly connect the “burlap sack” to the idea of a kidnapped girl, it does recognize the harmful implications towards the girl. Furthermore, Co-HARE understands the historical context behind the term “burlap sack”. The presence of the positive term “like” and the absence of overtly derogatory words might lead some models to classify the statement as non-offensive. However, our approach appropriately identifies the underlying hateful context.

In the second sample, the post contains hate speech targeting Jewish victims of the Holocaust by referencing Anne Frank. Our model accurately recognizes the historical background of Anne Frank as a Holocaust victim. While the baseline seems to overlook the historical significance associated with Anne Frank, our method correctly identifies her and assumes that the reference constitutes harassment against a Jewish victim, even though there is a slight misunderstanding about the context of “lack of speaking” in Fr-HARE.

Is GPT-3.5 a qualified teacher? Since our framework is based on distillation of generated rationales from GPT-3.5 to smaller models, it is crucial to verify whether the teacher is qualified. Figure 4 displays rationales produced by GPT-3.5-turbo, which is employed to train the student model. This example illustrates that the LLM not only discerns the hateful nuances towards both white and black individuals, but also offers more detailed explanations compared to rationales written by humans. Notably, it accurately correlates the historical context, associating the word “slaves” with “pets”. More analysis of rationales from GPT-3.5-turbo can be found in Appendix D.2.

4 Conclusion
In this paper, we present HARE framework to improve the ability of the language model to understand hate speech and provide clearer explanations for its decisions. We propose utilizing CoT reasons extracted from LLMs in two variants to overcome the logical gaps in human-annotated rationales. When fine-tuned on the SBIC and Implicit Hate datasets, our methods achieve superior detection performance and better qualified explanations.
Limitations

While we assess the quality of explanations generated by HARE using GPT-4, we do not conduct human evaluations, which are crucial for tasks requiring human-readable explanations. The primary reason for this omission is that the hate speech content and its respective explanations could be excessively offensive for annotators and GPT-4 already aligns with the level of inter-human agreement. In addition, the "verbosity bias," characterized by a preference for the longer text of GPT-4 as indicated by (Liu et al., 2023), may also serve as a limitation in our evaluation process.

Ethics Statement

Predicting whether an online post contains hate-speech is both technically and socially challenging. While methods for automating hate speech detection have utility in an online platform, it is critical that these are tuned and used appropriately. False-positive errors have potential to censor online speech, further marginalizing specific user groups, for example: use of n***** in AAVE English may be flagged. It is critical to understand specific reasoning behind a classification including deeply social reasons. While language models act as a mechanism to generate reasonable explanations, it is critical that they are used appropriately to prevent them from inadvertently educating users on how to craft more subtle and toxic language. We used automated evaluation metrics in this paper to prevent exposure of toxic language to human annotators. However, real-world usage would require validation that deeply rooted social issues are expressed correctly by these models.

It is also important to note that there might be concerns about the inherent bias in the GPT-3.5 model. While not flawless, GPT-3.5 has demonstrated its impartiality regarding gender, race, ethnicity, and religion by achieving the highest grade on the Harmfulness metric within the FLASK evaluation framework (Ye et al., 2023). Crucially, we only select rationales that align with the ground truth label for training, thereby mitigating biases not in sync with human annotators. Analysis of GPT-3.5-turbo can be found in Section 3 and Appendix D.2.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by Korea government (MSIT) [No. 2021-0-00907, Development of Adaptive and Lightweight Edge-Collaborative Analysis Technology for Enabling Proactively Immediate Response and Rapid Learning, 90%] and [No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST), 10%].

References


A Related Work

Hate Speech Detection Hate speech (Waseem et al., 2017) is a form of language designed to offend a particular individual or groups. In this study, we expand this definition by incorporating the broader concept of offensive language as in (Burnap and Williams, 2016; Ribeiro et al., 2018). Numerous recent works on hate speech detection have delved into providing underlying explanations of prediction on hate speech (Sap et al., 2019a,b; Mathew et al., 2021; ElSherief et al., 2021; Lin, 2022). One line on research focuses on keyword-based explanations (Sap et al., 2019a; Davidson et al., 2019; Mathew et al., 2021; Kim et al., 2022), but this approach often fails to capture implicit hatefulness that is not explicitly present in the text. Another approach involves explanations utilizing external knowledge sources (Sridhar and Yang, 2022; Lin, 2022), but these methods aim to solely improve classification performance. Yet another studies involve training generative models with human-written free-text rationales (Sap et al., 2019b; ElSherief et al., 2021; Huang et al., 2022) present in multiple benchmarks (Sap et al., 2019b; ElSherief et al., 2021). Nevertheless, due to the existence of logical gaps in these human-annotated rationales (Aggarwal et al., 2021; Sun et al., 2022), relying solely on these rationales results in sub-optimal detection and explanation quality. Our proposed HARE shows its effectiveness by incorporating LLM-generated rationales, which include logical completeness and abundant explanatory power extracted with our CoT prompting.

Self-Rationalization Self-rationalization, a technique where models provide explanations for their predictions, has been extensively studied to make models more understandable and transparent (Marasović et al., 2021; Wiegreffe et al., 2021a,b). Recent studies leverage rationale-augmented exemplars to few-shot prompt LLMs (Wei et al., 2022; Wang et al., 2022; Lampinen et al., 2022), while others fine-tune smaller models using the rationales generated by LLMs (Ho et al., 2022; Wang et al., 2023a; Chan et al., 2023; Wang et al., 2023b). Several works apply their self-rationalization methods to detect hate speech and generate underlying explanations (Marasović et al., 2021; Ludan et al., 2023; Chen et al., 2023). However, they directly use human-written rationales in their framework, which may lack logical completeness. Instead, our method adopts the approach of utilizing the rationales generated by LLMs.

B Implementation Details

Training Details All the experiments utilize an Adafactor (Shazeer and Stern, 2018) optimizer with batch size of 32 and learning rate \( \in \{5e^{-3}, 5e^{-4}, 5e^{-5}\} \) without learning rate scheduling. We train all models for 10 epochs and utilize the checkpoint with the highest validation F1 score to evaluate on the test set. Top-k sampling with \( k=20 \) is adopted as a decoding method. For dataset split, we use official dataset split for SBIC and random split with 6:2:2 ratio for train, validation, and test sets for Implicit Hate. All our experiments are conducted using a single NVIDIA A6000 GPU.

C Details on GPT-4 Evaluation

To conduct GPT-4 evaluation, we utilize the prompts proposed by Zheng et al. (2023). For single answer grading, the prompt format for GPT-4 is written as follows:

```
[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: 
"

[Question]
Determine whether the following post is offensive or not. You must choose one of the options.
Post: P
Options: (A) Offensive (B) Not offensive
[The Start of Assistant’s Answer] C, R
[The End of Assistant’s Answer]
```

To compare the alignment of generated rationales with human-written rationales, we conduct a pairwise comparison of explanations against ground-truth rationales. These ground-truth rationales encompass the annotated target groups and implied statements from the instruction, and we compare the results. Additionally, to minimize bias from the order of candidate answers, we adopt the approach of Zheng et al. (2023), considering both original and swapped orders of predicted explanations. If \( C + T + I \) and \( C = H A R E \) are chosen alternately, it is deemed a tie. Should one method...
be selected following a tie, that method is considered as the chosen one. The prompt format for the pairwise comparison is provided below:

D.1 Qualitative Study of HARE

Figures 5, 6, 7, and 8 showcase results generated by the fine-tuned Flan-T5-1arge model using HARE and $C+T+F$, based on test samples from SBIC. Although a brief explanation is provided in Section 3.2, we delve deeper with an extended analysis of the 20 examples from our qualitative study. These 20 samples were randomly chosen in proportion to their correct and incorrect predictions across the different methods.

When comparing human-written annotations with HARE, it becomes evident that the annotated rationales in SBIC often take the form of implied statements, following a simple Hearst-like pattern (Sap et al., 2019b). Learning from such rationales, which are closely tied to the conclusion, creates a logical gap for the model and makes interpretation challenging for humans. For instance, understanding hate speech without background knowledge references, such as 'burlap sack', can make it difficult to see the connection between the statement "girls are not worthy of equal life" and the provided sentence. Figures 5 and 6 showcase successful cases where models have attempted to bridge this reasoning gap through HARE, offering more detailed rationales that encompass the context. Furthermore, these models exhibit capabilities not seen in previous research, such as detecting terms with historical significance (e.g., 'burlap sack' or 'Anne Frank') or common words that may carry hateful connotations (e.g., 'reds'), thus enhancing the intermediate reasoning process.

However, when examining the failure cases in Figures 7 and 8, the results show that HARE sometimes fails due to increased sensitivity to potentially harmful terms, thereby classifying them as offensive. While this increased sensitivity can be viewed as a drawback, there are instances, such as with the Alzheimer example, where an expression might be interpreted as hateful depending on the individual. This suggests that HARE aims to classify a post as hateful if it could be considered offensive to certain groups. Moreover, considering the David Bread Katz example, it is also challenging for HARE to decide if the post is offensive if it post with background that it hasn’t encountered, possibly due to a lack of background knowledge regarding the implied shooting incident, illustrating the limitation of LLM distillation.

D.2 Qualitative Study on GPT-3.5 rationale

When comparing annotations with rationales generated by GPT-3.5, we observe that human-written rationales from SBIC use implied statements that follow simple Hearst-like patterns (Sap et al., 2019b). In contrast, models like LLM (e.g., GPT-3.5) tend to provide detailed, step-by-step explanations, often complemented by relevant social background information, which is immensely beneficial. For example, while earlier rationales might omit mentioning Bill Cosby’s conviction of sexual assault, GPT-3.5 explicitly informs us of this fact, greatly enhancing comprehension. A particularly striking example is the "pet" case. While a human annotator perceived it as hate speech targeted at black individuals, GPT-3.5 points out that it could also be used derogatorily against white individuals, thereby emphasizing the potential biases in hate speech detection.
When rationales are categorized and structured, as seen in SBIC, instead of being tailored to individual posts, they may not be sufficient for learning implications. This could explain why, as suggested by Table 2, there’s a decrease in generalization for $C+T+I$. Our approach offers aligned rationales for each post at minimal cost, enabling the learning of diverse reasons for potential hate, which in turn leads to enhanced generalization.
I like my women like I like my cigars 7 years old and coming from Cuba in a burlap sack, (A) Offensive
 girls are not worthy of equal life / girls are property / girls are used for sex / trivializes harm to victims
 The answer is: (B) Not offensive

The answer is: (A) Offensive 1. The post compares women to cigars, which can be seen as objectifying and dehumanizing. 2. It also makes a reference to a specific age group, which can be seen as inappropriate and offensive. 3. The post also mentions Cuban cigars, which can be seen as promoting or glorifying illegal activities. Based on these steps, the post can be considered offensive. It objectifies and dehumanizes women, promotes illegal activities, and includes references to illegal activities.

The answer is: (A) Offensive 1. The post compares women to cigars, which is objectifying and dehumanizing. This targets women by reducing them to objects and implying that they are only valuable for their physical appearance. 2. The post specifically mentions that the cigars should be "7 years old," which is highly inappropriate and offensive as it promotes child exploitation and sexual abuse. This targets women by implying that they are sexual objects and should be subjected to the same level of abuse as cigars. 3. The post also mentions that the cigars should be "coming from Cuba in a burlap sack," which is a clear reference to the sex trafficking scandal that occurred in the United States in the past. This targets women by implying that they are only valuable for their sexual appeal and that their worth is solely based on their ability to fulfill sexual desires. Overall, the post is offensive as it objectifies and devalues women, promotes child exploitation, and makes light of a serious and sensitive topic.

"I think a lot, but I don't say much." ~ Anne Frank

Let's hope these reds aren't mistaken for the usual white targets of these jungle savages! (A) Offensive

The answer is: (A) Offensive 1. The post starts with a quote from Anne Frank, a Jewish girl who died during the Holocaust. 2. The quote itself is a statement about her lack of speaking much. 3. The post then makes a comparison by stating that she has been proven to have not spoken anything for almost a century. Based on this analysis, the post can be considered offensive. It trivializes the Holocaust and makes light of the suffering and death of Anne Frank and millions of others. The comparison between her lack of speaking and her lack of speaking is insensitive and disrespectful.

The answer is: (A) Offensive 1. The post starts with a quote from Anne Frank, a Jewish girl who wrote a diary during the Holocaust. 2. The quote states, "I think a lot, but I don’t say much." 3. The post then adds, "But lately she’s been proven to have not said anything for almost a century." This post targets Holocaust victims and Jewish folks by making a joke about Anne Frank's lack of speaking for almost a century. It implies that she has been silent or silent about her experiences during the Holocaust. This trivializes the harm and suffering that Holocaust victims endured. Therefore, the post can be considered offensive as it makes light of a tragic event and trivializes the harm suffered by Holocaust victims.

Actually I never had too many women until I became a white nationalist. (A) Offensive

The answer is: (A) Offensive 1. The post states that the person has never had many women until they became a white nationalist. This implies that the person believes that white nationalists are superior to other racial or ethnic groups. 2. The post specifically mentions that the person has never had too many women before becoming a white nationalist. This implies that the person believes that white nationalists are superior or more desirable than other racial or ethnic groups. 3. The post targets "black folks, latino/latina folks, native american/first nation folks" by implying that they are inferior or less desirable than other racial or ethnic groups. 4. The post implies that the person has never had many women until they became a white nationalist. This implies that the person's lack of women is due to their inferiority or lack of desirability. Therefore, the post can be considered offensive.

I can't stand three "grow" as females &/or/120131c12797912000a2640edf893129318c12797912000a2640edf how you ** grow ** but childish as hell! (B) Not offensive

The answer is: (B) Not offensive 1. The post uses derogatory language by referring to females as "grow" and "childish." Language is offensive as it dehumanizes and insults women. 2. The post also uses emojis that can be interpreted as mocking or belittling the mentioned females. This can be seen as offensive as it belittles and mocks them. Based on these steps, the post can be considered offensive.

The answer is: (A) Offensive 1. The post uses derogatory language by referring to females as "grow" and "childish." Language is offensive as it dehumanizes and insults women. 2. The post also uses emojis that can be interpreted as mocking or belittling the mentioned females. This can be seen as offensive as it belittles and mocks them. Based on these steps, the post can be considered offensive. It uses derogatory language, degrades women, and includes derogatory terms.

Dunnes strikers to attend funeral - BelfastTelegraph.co.uk

Let's hope these reds are n't mistaken for the usual white targets of these jungle savages! (A) Offensive

The answer is: (A) Offensive 1. The post mentions Dunnes strikers attending a funeral. 2. It then makes a statement about the reds being mistaken for the usual white targets of these "jungle savages." Based on this explanation, the post can be considered offensive. The statement about the reds being mistaken for the usual white targets of these "jungle savages" implies a negative stereotype about a specific racial group, suggesting that they are inherently dangerous or primitive. This kind of language promotes racism and discrimination.

The answer is: (A) Offensive 1. The post mentions Dunnes strikers attending a funeral. 2. It then states, "Let's hope these reds aren't mistaken for the usual white targets of these jungle savages." The term "reds" is a derogatory term for black people, implying that they are inferior or less than human. 3. The post also uses the term "jungle savages" to refer to black people, further dehumanizing them. 5. The post implies that black people are savages, which is a derogatory and offensive stereotype. Overall, the post targets black people by using derogatory language and perpetuating harmful stereotypes. It implies that black people are inferior, savages, and savages. Therefore, the post is offensive.

Figure 5: Success cases of HARE
Four-year-old boy: Why did she get to walk in first? Dad: Because she's a girl. And girls go first. Four-year-old boy: Why? That's DUMB! Dad: Because girls need protecting from strong boys like you.

The post does not appear to be offensive. It is simply providing information about a game and stating the average answers.

The answer is: (B) Not offensive

1. The post mentions a specific game, BG, and a specific error in the game. 2. It states that the average answers are 5 / 5, with one less or more errors not being changed.

The post does not appear to be offensive. It is simply discussing the number of errors and suggesting that there should not be any change in the number of answers.

The answer is: (B) Not offensive

1. The post expresses the person's enjoyment of horror, thriller, and crime movies and their desire to watch them together.

The post does not appear to be offensive. It simply expresses the person's enjoyment of horror, thriller, and crime movies.

The answer is: (B) Not offensive

1. The post states a historical fact about English entrepreneurs setting sail to found Jamestown in what became Virginia the first of the 13 colonies that became the USA. 2. It mentions that Jamestown was the first of the 13 colonies that became the USA.

The post does not appear to be offensive. It simply provides historical information about the founding of Jamestown and the subsequent colonies.

The answer is: (B) Not offensive

1. The post mentions that the person loves horror, thriller, and crime movies. 2. They also mention that their friend and they always watch horror together.

The post does not appear to be offensive. It simply expresses the person's interest in horror, thriller, and crime movies and their friend's interest in watching them together.

The answer is: (B) Not offensive

1. The post mentions Frederick Nietzsche, who is a historical figure known for his role in the Holocaust and the persecution of millions of people. 2. The post suggests that Nietzsche's beliefs about the need to kill weaker elements to prevent contamination of the gene pool are not based on scientific evidence.

The post does not appear to be offensive. It is simply discussing the number of errors and suggesting that there should not be any change in the number of answers.

The answer is: (B) Not offensive
what's the worst thing about being told you have Alzheimer's? It doesn't happen at once.

The answer is: (B) Not offensive.

The post mentions David "Bread" Katz being eliminated from the Madden 19 Tournament. It states that he was eliminated from a tournament and his decision to eliminate the competition from IRL. Based on this information, the post does not appear to be offensive. It simply states a fact about David "Bread" Katz being eliminated from a tournament and his decision to eliminate the competition from IRL.
Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Don't get it twisted, carnal living doesn't please God; Fix your mind & heart on the WORD to live a life pleasing to Father God! 😇

The answer is: (B) Not offensive.

Figure 8: Failure cases of HARE

- Post: Nigga you made it RT @HelloCupkake: Too many good single girls, too many hoes taken.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: A Madden player was fatally shot during a touchdown... Trueboy really was in the End Zone
  - Label: (A) Offensive
  - Annotation: The answer is: (A) Offensive.

- Post: one time for the real niggas , two times for the bad bitches.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: RT @HelloCupkake: Too many good single girls, too many hoes taken.
  - Label: (A) Offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: A Madden player was fatally shot during a touchdown... Trueboy really was in the End Zone
  - Label: (A) Offensive
  - Annotation: The answer is: (A) Offensive.

- Post: one time for the real niggas , two times for the bad bitches.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: Nigga you made it RT @HelloCupkake: Too many good single girls, too many hoes taken.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: A Madden player was fatally shot during a touchdown... Trueboy really was in the End Zone
  - Label: (A) Offensive
  - Annotation: The answer is: (A) Offensive.

- Post: one time for the real niggas , two times for the bad bitches.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: Nigga you made it RT @HelloCupkake: Too many good single girls, too many hoes taken.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.

- Post: A Madden player was fatally shot during a touchdown... Trueboy really was in the End Zone
  - Label: (A) Offensive
  - Annotation: The answer is: (A) Offensive.

- Post: one time for the real niggas , two times for the bad bitches.
  - Label: (B) Not offensive
  - Annotation: The answer is: (B) Not offensive.
Why do so many white people own pets? Because they can’t own people anymore.

Label: (A) Offensive
Annotation: black people are inferior / black people should be slaves
GPT-3.5 rationales:
1. The post makes a generalization about white people and their ownership of pets.
2. It suggests that white people historically owned people, implying slavery.
3. It implies that white people now own pets as a replacement for owning people. This post can be considered offensive because it perpetuates a harmful stereotype about white people and makes light of a dark and painful history of slavery. It reduces the complex issue of slavery to a simplistic and offensive joke. While humor can be subjective, this post crosses the line by making inappropriate and insensitive comparisons.

Had a threesome with Bill Cosby Unlike the woman we were with, it was a night I’ll never forget.

Label: (A) Offensive
Annotation: makes a joke at the expense of women who were raped / makes light of their struggle
GPT-3.5 rationales:
1. The post mentions a threesome involving Bill Cosby, who has been accused and convicted of sexual assault. This brings up sensitive and potentially offensive subject matter.
2. The post suggests that the person had a memorable night, contrasting the experience with the woman involved, potentially implying that the woman had a negative experience or that the post is making light of the seriousness of the situation.
Based on these steps, the post can be considered offensive because it trivializes allegations of sexual assault and could be perceived as insensitive towards the victims.

Freedom of the press was only so the Press could report the truth not to be a Democrat propaganda machine that’s exactly what Hitler did

Label: (A) Offensive
Annotation: compared to nazi party, democrats control the press / democrats spread propaganda, suppresses truth
GPT-3.5 rationales:
1. The post starts by mentioning "freedom of the press", which is a constitutional right that allows journalists to report without censorship or interference from the government.
2. The post states that this freedom was intended for the press to report the truth.
3. However, it then takes a political stance by claiming that the press is behaving like a "Democrat propaganda machine".
4. The post compares this behavior to Hitler’s propaganda tactics. Based on this analysis, the post can be considered offensive for multiple reasons. Firstly, it accuses the press of being a biased propaganda machine, which can be seen as an attack on the integrity and credibility of journalists. Secondly, it makes an unfair and unsupportable comparison between the press and Hitler’s propaganda tactics, which is highly offensive and diminishes the seriousness of the atrocities committed by Hitler and the Nazi regime.

What do you call that useless flesh around the vagina that doesn’t do anything but lie around? The baby

Label: (B) Not offensive
Annotation: -
GPT-3.5 rationales:
1. The post contains offensive language and derogatory statements about the female anatomy.
2. The post uses derogatory language to refer to a specific part of the female anatomy, equating it with being useless and inactive.
3. The post implies a negative view of pregnancy and childbirth, suggesting that the purpose of the “useless flesh” is solely for the baby’s presence.
4. The post objectifies and devalues women by reducing their bodies to something insignificant and purposeless.
Based on these explanations, the post is offensive because it uses derogatory language, devalues women, and promotes negative stereotypes about the female anatomy.

(a) Success cases of GPT-3.5-turbo

(b) Failure cases of GPT-3.5-turbo

Figure 9: Success cases and failure cases of GPT-3.5-turbo when prompted with our CoT prompt.