

# SMARTER: A Data-efficient Framework to Improve Toxicity Detection with Explanation via Self-augmenting Large Language Models

Huy Nghiem, Advik Sachdeva, Hal Daumé III

University of Maryland  
{nghiemh, asachde1, hal3}@umd.edu

## Abstract

**WARNING:** *This paper contains examples of offensive materials.* To address toxic content on social media, we introduce *SMARTER*, a data-efficient 2-stage framework for explainable content moderation using Large Language Models (LLMs). In Stage 1, we leverage LLMs' own outputs to generate synthetic explanations for correct and incorrect labels, enabling preference optimization with minimal supervision. In Stage 2, we refine explanation quality through cross-model training, allowing weaker models to align with stronger ones. Experiments on 3 classification tasks (*HateXplain*, *Latent Hate*, *Implicit Hate*) show *SMARTER* achieves up to 13% macro-F1 improvement over few-shot baselines using only 6-57% of training data. Our framework offers a scalable strategy for low-data settings by harnessing LLMs' self-improvement for explainable moderation.

## 1 Introduction

In recent years, social media platforms have enabled millions of users to virtually interact at a global scale. While a vital channel to disseminate information and knowledge, these platforms also facilitate the spread of harmful materials (Hwang and Kim, 2015; Cinelli et al., 2021; Nguyen et al., 2024). Arguably the most concerning among them is toxic content, which could induce serious negative psychological impact on the audience (Castaño-Pulgarín et al., 2021; Windisch et al., 2022; Tontodimamma et al., 2021). As a result, content moderation is crucial to detect and mitigate its harm.

Toxic content encompasses a wide spectrum of terminologies whose definitions vary by platform: hate speech, cyberbully, sexist, racist etc. (Gelber, 2021; Anjum and Katarya, 2024; Wang et al., 2025). Typically, human moderators must manually adjudicate the target content, an inefficient process taxing to the moderators themselves (Baker et al., 2020; Spence et al., 2023). Recently, social

media platforms can significantly reduce this overhead by integrating machine learning models to automate the detection pipeline (d'Sa et al., 2020; Sumanth et al., 2022).

While capable of achieving impressive classification performance, these increasingly sophisticated models require ample resources to train, especially when the target label space covers nuanced concepts (Fortuna et al., 2021; Sap et al., 2021; Zhao et al., 2021). Developing a robust dataset for training typically demands significant effort in curation and human annotation, often involving iterative updates to adapt the model to evolving linguistic trends (Toraman et al., 2022; Bespalov et al., 2023; Rad et al., 2025). In addition, this pipeline is limited to decision-only pipeline. Failure to provide users with human-understandable signals renders the adjudication process opaque, inviting further issues on transparency and trustworthiness.

In this paper, we address the pressing need for intuitive explanations as a cornerstone of content moderation by leveraging Large Language Models (LLMs), which have shown remarkable capabilities in reasoning-related tasks (Wei et al., 2021; Yao et al., 2024). Specifically, our contributions are:

- ◊ We introduce **SMARTER**<sup>1</sup> (Self-augMentAtion Regimen Towards Efficient Content ModeRation), a 2-stage framework where (1) each LLM self-augments on few-shot data using preference optimization with synthetic explanations, (2) cross-model refinement to further improvement by training one LLM with responses generated by another on held-out sets.
- ◊ Experiments conducted on 3 classification tasks (*HateXplain*, *Latent Hate*, *Implicit Hate*) with Llama-3.1-8B-Instruct and COT-T5-XL demonstrate up to 13.5% increase in macro-F1

<sup>1</sup>Our repository can be found at [https://github.com/nghiem-nlp/hate\\_dpo\\_public](https://github.com/nghiem-nlp/hate_dpo_public)

scores in few-shot settings with high-quality explanations, outperforming a range of commercial models.

While commercial APIs may incur high costs and limited transparency, SMARTER delivers competitive performance and explainability with 6-57% of training data, ensuring deployment control.

## 2 Related Work

**Robust Toxicity Detection** Existing research (Caselli et al., 2021; Sarkar et al., 2021; Toraman et al., 2022) has proposed strategies to enhance cross-domain robustness in toxicity detection, such as combining multiple datasets to improve generalization (Antypas and Camacho-Collados, 2023) and incorporating label definitions to support few-shot adaptation (Nghiem et al., 2024). Building on these advances, recent studies leverage LLMs for toxicity detection, demonstrating strong zero- and few-shot performance (Masud et al., 2024; Jahan et al., 2024; Roy et al., 2023; Kumarage et al., 2024). LLMs can also generate textual explanations, enhancing interpretability for moderators and users (Di Bonaventura et al., 2024; Yang et al., 2023; Calabrese et al., 2024; Almohaimeed et al., 2025). We extend this line of work by training open-source LLMs to jointly perform classification and explanation to promote transparency.

**Aligning LLMs to Human Preferences** In addition to extensive pretraining, modern LLMs also undergo post-training that both unlocks additional capabilities (Longpre et al., 2023) and aligns the outputs towards human preferences (Ouyang et al., 2022). Recent Reinforcement Learning (RL)-based techniques (Shao et al., 2024; Zweiger et al., 2025; Nghiem et al., 2025) enable promising LLM self-alignment at expensive computational cost. Another body of work showcases LLMs to self-improve via their own synthetic data (Tang et al., 2023; Li et al., 2023), such as Self-Refine (Madaan et al., 2023) and Self-Instruct (Wang et al., 2022). Building on these ideas, our framework presents a simple yet principled approach that allows LLMs to iteratively improve with minimal training cost while amenable to human supervision.

## 3 Overview of Framework

As shown in Algorithm 1, our framework consists of two stages. First, we focus on optimizing the LLMs’ classification performance through alignment tuning using self-augmenting synthetic data

---

### Algorithm 1 High-Level Overview of SMARTER

---

```

1: Input: LLMs  $\mathcal{L}$ , performance metric  $\mathcal{M}$ 
   Datasets  $\mathcal{D}_{\text{train}}$  (with explanation),  $\mathcal{D}_{\text{val}}$ ,  $\mathcal{D}_{\text{test}}$ ,
2: Optional: Initial SFT pretraining on relevant data
3: Stage 1: Individual LLM Self-Augmentation
4: for  $K \in \{K_{\text{start}}, \dots, K_{\text{end}}\}$  do
5:   Sample  $K$ -shot subset  $\mathcal{D}^{(K)}$ 
6:   for all  $\text{LLM}_i \in \mathcal{L}$  do
7:     SFT on  $\mathcal{D}^{(K)}$ 
8:     Collect chosen responses for gold labels
9:     Collect rejected responses for incorrect labels
10:    Align via preference optimization using synthetic
       preference data; evaluate on  $\mathcal{D}_{\text{val}}$ 
11:   end for
12: end for
13: Stage 2: Cross-Model Refinement
14: for  $\text{LLM}_a \neq \text{LLM}_b \in \mathcal{L}$  do
15:   Select  $K_{\text{check}} \in \{K_{\text{start}}, \dots, K_{\text{end}}\} \rightarrow \mathcal{D}^{(K_{\text{check}})}$ 
16:   SFT  $\text{LLM}_a$  on  $\mathcal{D}^{(K_{\text{check}})}$ 
17:   Self-augment  $\text{LLM}_a$  with responses from  $\text{LLM}_b$  on
        $\mathcal{D}^{(K'_{\text{check}})} \neq \mathcal{D}^{(K_{\text{check}})}$ 
18: end for
19: return Model with best performance  $\mathcal{M}$  on  $\mathcal{D}_{\text{val/test}}$ 

```

---

generation. In the second stage, we refine the quality of the generated explanations by cross-model refinement, leveraging the synergy between different LLMs. Section 4 and 5 present the experimental setup and findings for Stage 1. Section 6 focuses on cross-model improvement in Stage 2. Section 7 analyzes the consistency of explanations with respect to the definitions and labels.

## 4 Preliminary Setup

In this section, we provide details about the empirical setup on data, model, and a description of classification technique via prompting.

### 4.1 Data

We select 2 datasets that cover distinct yet still relevant concepts to toxic content moderation and derive from them 3 classification tasks.

- ◊ **HateXplain**, introduced by Mathew et al. (2021). Consisting of posts collected from 2 social media sites Gab and Twitter (now X), this dataset contains 20,148 samples annotated by crowdworkers of the Amazon Mechanical Turks (AMT) platform on 3 labels *Hate*, *Offensive* and *Normal*.
- ◊ **Latent Hate**, curated by ElSherief et al. (2021), focuses on covert hate speech collected from Twitter. With 19,112 samples also annotated by AMT workers, this dataset’s primary label space contains 3 high-level categories: *Not Hate*, *Implicit Hate*, *Explicit Hate*.

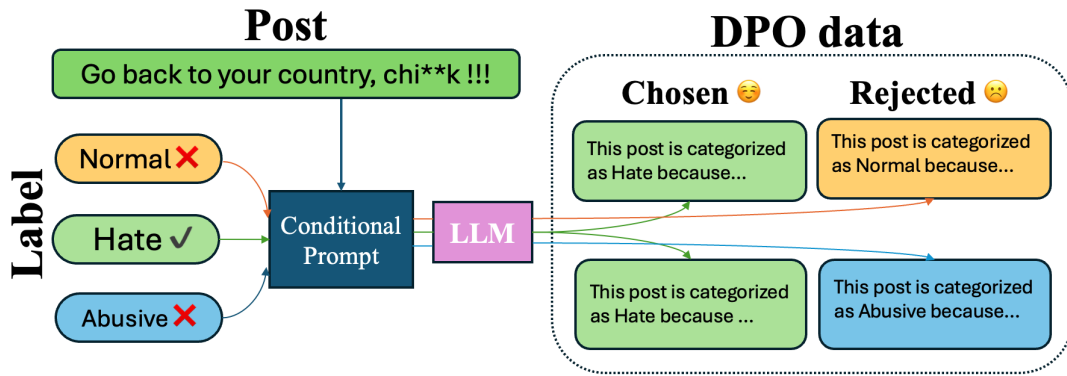


Figure 1: Self-augmenting pipeline: for each post, explanations are conditionally generated for the gold label and all incorrect labels using prompt template in Figure 7. For DPO, data consists of matching the explanation of the correct label with another incorrect label’s.

**Input:** It means people aren’t buying the lie! diversity k\*\*s now prove me wrong!  
**Explanation:** The post uses coded or indirect language to disparage a person or group based on their race. Overall, the use of the phrase "diversity k\*\*s" and the challenge to "prove me wrong" indicate a negative view of diversity, suggesting a belief that diversity is harmful. This coded language indirectly disparages people of color, making it an example of implicit hate.  
**Label:** Implicit Hate

Figure 2: Example of explanation and classification label produced by *Llama* at  $K = 128$  shot on the dataset *Latent Hate*. For definition of categories, see Table 10.

The authors also assigned the subset of 4,153 tweets of the *Implicit Hate* category a label among 6 secondary fine-grained categories: *White Grievance, Incitement to Violence, Inferiority Language, Irony, Stereotypes and Misinformation, Threatening and Intimidation.*

Label definitions are included, enabling context integration into explanations following Yang et al. (2023) and Nghiem and Daumé (2024). We denote these three tasks as *HateXplain, Latent Hate,* and *Implicit Hate.* Table 4 summarizes data splits; preprocessing details are in Appendix B. Figure 2 provides an example of the explanation produced by our model on *Latent Hate.*

## 4.2 Models

Extending our focus on accessibility, we employ 2 open-source LLMs that have exhibited strong performance in similar classification tasks:

- ◊ **COT-T5-XL:** A variant of the base encoder-

decoder Flan-T5-XL (Longpre et al., 2023) architecture, this model consists of 3 billion parameters and is further trained on the COT corpus, a collection of 1.8 million samples with chain-of-thought style explanations, which enhances its capabilities on a variety of reasoning tasks (Kim et al., 2023).

- ◊ **Llama-3.1-8B-Instruct:** A decoder-only LLM released by Meta (Dubey et al., 2024). Consisting of 8 billion parameters, this model is further instruction-tuned to aligned with human preferences in addition to being trained on extensive linguistic corpora.

For brevity, we refer to the models as *T5* and *Llama* respectively for the remainder of this paper.

## 4.3 Classification via Prompting

We apply the Chain-of-Thought (COT) paradigm (Wei et al., 2022) in our classification tasks to obtain the explanation from the LLMs necessary for content moderation. Inspired by Nghiem and Daumé (2024) and Yang et al. (2023), we use the prompt in Figure 7, which incorporates the target post, the set of labels corresponding to the task and their associated definitions as context for the LLM to provide its explanation and predicted category.

## 4.4 Auxiliary Pretraining

Nghiem and Daumé (2024) found that supervised finetuning (SFT) off-the-shelf LLMs on *HateCOT* – a general compilation of toxic posts, labels, definitions and synthetic explanations – can boost their performance in downstream hate speech detection. To maximize expected gains with minimal in-domain data, we adopt this strategy and perform SFT on the 2 chosen LLMs using *HateCOT* to

prime them for our classification tasks. Finetuning is performed by training LoRA adapters (Hu et al., 2021), a parameter-efficient approach implemented by the HuggingFace library (see Appendix F) as the basis for downstream experiments.

## 5 Stage 1: Individual Model Training

We demonstrate SMARTER’s data efficiency by training models at extreme low-shot settings and evaluating against off-the-shelf, commercial, and full-training baselines.

### 5.1 K-shot In-domain Finetuning

For each task, we select uniformly at random  $K$ -shot without replacement  $\in \{16, 32, 64, 128, 256\}$  samples from the training set. We utilize the seed explanations by Nghiem and Daumé (2024) to further SFT the existing *HateCOT*-pretrained *T5* and *Llama* models similar to the auxiliary pretraining phase above.

**Self-augmenting via Alignment Tuning** We optimize the LLMs’ performance while not having access to more training data by exploiting the models’ sycophantic tendency. More concretely, we augment the training sample size by prompting the LLMs to generate explanations *conditioned on the incorrect labels* (as illustrated in Figure 1). This process supplies us with synthetic post-explanation pairs that are by design *dispreferred* compared to the true *preferred* sample pairs, enabling the use of the following alignment techniques to further finetune the models.

- ◇ **Direct Preference Optimization (DPO)**: An offline RL-based technique that optimizes the policy via an implicit reward model using preference data (Rafailov et al., 2024). DPO requires human preference data where a pair of chosen vs. rejected responses is presented for each prompt.
- ◇ **Kahneman-Tversky Optimization (KTO)**: An alternative to DPO, KTO leverages prospect theory to construct human-aware losses (HALOs) to optimize the policy (Ethayarajh et al., 2024). KTO only requires list-wise human preference data: a data point consists of a prompt, a response and a binary flag that indicates the acceptability of the response.

Appendix F.3 offers more technical details on these techniques. Our self-augmenting methodology naturally procures training datasets in the required format for KTO by simply designating

the explanation conditioned on the gold label as positive, and otherwise negative. For DPO, we match the explanations conditioned on the correct-incorrect pair sequentially to satisfy the pairwise chosen-rejected format. As an example, 100 posts of *HateXplain* (whose label space is of size 3), should produce  $100 * 3 = 300$  training samples for KTO, and  $100 * (3 - 1) = 200$  for DPO.

#### 5.1.1 Experimental Pipeline

The experiments follow a two-stage process. We first finetune *T5* and *Llama* at  $K \in \{16, 32, 64, 128\}$  followed by DPO self-augmentation. At  $K = 256$ , we compare KTO and DPO alignment.<sup>2</sup> For comparison, we finetune both LLMs on full training data for classification without explanations (*Full* models), and similarly train ModernBERT (large) (Warner et al., 2024) as an additional baseline.

In addition, we examine 2 partial sampling strategies to circumvent the proportionate growth of self-augmenting synthetic data to label space when  $K$  is large. In the first, we construct augmented data by selecting 128 and 192 shots of samples uniformly at random from the Baseline’s 256-shot pool. In the second, instead of  $K$ -shot sampling, we select  $K' * |S| * (|S| - 1)$  samples from all possible post-label pairs of the 256-shot Baseline pool (incorrect labels included) to create augmented data, where  $K' \in \{128, 192\}$  and  $|S|$  is the label space’s size. While both approaches should yield identical size of training data for the same  $K'$ , the first method, denoted *DPO-K*, preserves post-label parity whereas the second method, denoted *DPO-N*, yields higher diversity of post content.

Experiments are implemented with the HuggingFace library. We sample  $K = 50$  shot on the held out Validation extraction of each dataset to tune hyperparameters, and evaluate on samples from the Test set as shown in Table 4. Technical details are included in Appendix F.

## 5.2 Classification Results

### 5.2.1 Results for $K \leq 128$

We report macro-F1—unweighted mean of per-class F1 scores—as the evaluation metric (Christen et al., 2023). Figure 3 displays the bar plots of the LLM’s Baseline and DPO-enhanced performance on the test sets.

<sup>2</sup>KTO trainer for *T5* is unavailable, so we evaluate KTO only for *Llama*.

Self-augmentation with DPO generally boosts classification performance over Baseline counterparts, with the sole reversal for *T5* on *Latent Hate* at  $K = 16$  (2.1% reduction), along with static scores for *T5* on *HateXplain* at  $K = 32$  and *Llama* on *Latent Hate* at  $K = 64$ . These infrequent and negligible exceptions highlight our method’s consistent enhancement. For instance, F1 gains reach 14.5% for *Llama* (*HateXplain*,  $K = 128$ ) and 25% for *T5* (*Implicit Hate*,  $K = 64$ ).

Notably, DPO self-augmentation improves performance even at  $K \leq 64$ , showcasing our method’s advantage in low-resource settings. *Llama* Baseline models appear to consistently outperform *T5* across all tasks. However, with DPO self-augmentation, *T5* is able to narrow the margin on *Latent Hate* and *Implicit Hate*. In contrast, *Llama* still dominates on *HateXplain*. These discrepancies suggest an interaction between dataset and model choice, and practitioners may benefit from contrasting different options with our self-augmenting techniques for best results.

### 5.2.2 Results for $K = 256$

In [Figure 3](#), we observe the macro-F1 scores for models  $K = 256$  on the subfigures on the right. DPO-augmentation continues to enhance Baseline models which already benefits from more training data relative to lower  $K$  values, with improvement in scores reaching up to 13.2% for *Llama* and 14% for *T5*. In contrast, KTO-augmentation appears ineffectual on *HateXplain* and even hinders performance on *Latent Hate* and *Implicit Hate*, showing notable reduction from Baseline scores, respectively. With this, we hypothesize that DPO’s contrastive nature reinforces signals that allow the LLM to learn from its own mistakes, thereby sharpening its reasoning ability for this type of discriminatory task. On the other hand, KTO’s reliance on implicit signals from listwise data is not sufficient to improve discernment of classes when the semantic distinction between labels are nuanced for *Latent Hate* and *Implicit Hate*.

Results also show that DPO-augmented subsampling strategies still demonstrate generally better F1 scores than Baselines, further reinforcing the benefit of DPO-augmentation. There appears no clear winner between *DPO-K* and *DPO-N* among all 3 datasets. Augmenting with larger  $K'$  values (approaching 256) consistently yields scores closer to the maximal F1, indicating its role as a hyperparameter that practitioners need to decide for the

trade-off between computational cost and benefit.

**Ablation and Baseline Comparisons.** To isolate SMARTER’s contribution, [Table 5](#) shows off-the-shelf models without HateCOT pretraining achieve 0.52 F1 on *HateXplain*. Adding HateCOT (Baseline  $K = 256$ ) improves to 0.58 (+0.06), while SMARTER’s DPO self-augmentation reaches 0.64 (+0.06 additional), contributing performance gain.

We evaluate commercial models (GPT-4o-mini, GPT-4.1, GPT-5-chat, Qwen-32B) under zero-shot and 16-shot ICL (3 seeds; [Table 6](#)). In [Table 1](#), our DPO-augmented models at  $K = 256$  consistently outperform both paradigms. Remarkably, 16-shot ICL often degrades performance: GPT-4o-mini drops from 0.50→0.29 on *HateXplain* with high variance ( $\pm 0.10$ ), while even GPT-5-chat (0.62 mean) trails our *Llama\_DPO-256* (0.64). Following *ModernBERT*’s setup, we also train base models without explanations. At  $K = 256$ , our DPO variants retain 86%+ of *Full* performance on *HateXplain* and surpass it on *Latent Hate*.

## 6 Stage 2: Cross-model Refinement

Having shown individual models self-improve efficiently, we explore whether cross-model training can further enhance performance by transferring explanation quality between LLMs. We first conduct human evaluation of both models’ explanations, then train each model on the other’s outputs.

### 6.1 Human Evaluation

Due to resource constraints, we only perform this evaluation pipeline for *HateXplain*. We opt to collect outputs from the SFT+DPO models at  $K = 128$  to preserve the rest of the training data for further refinement experiments. To make a fair comparison on the outputs to agree on the predicted label, we use the prompt in [Figure 8](#) to obtain explanations conditioned on the correct gold label. This technique allows us to obtain 342 samples with consistent labels, or 89% of the possible  $128 * 3 = 384$  possible samples. The other 42 samples contain discrepant predicted labels, where the LLMs revert to behaviors ingrained by previous training.

To simulate real-life deployment, we solicit 14 crowdworkers from diverse demographic backgrounds on the platform Amazon Mechanical Turks (AMT) to annotate their preference on the explanations offered by the LLMs in this study. Annotators are encouraged to evaluate the pairs of explanation for each post based on the criteria of *Clarity*,

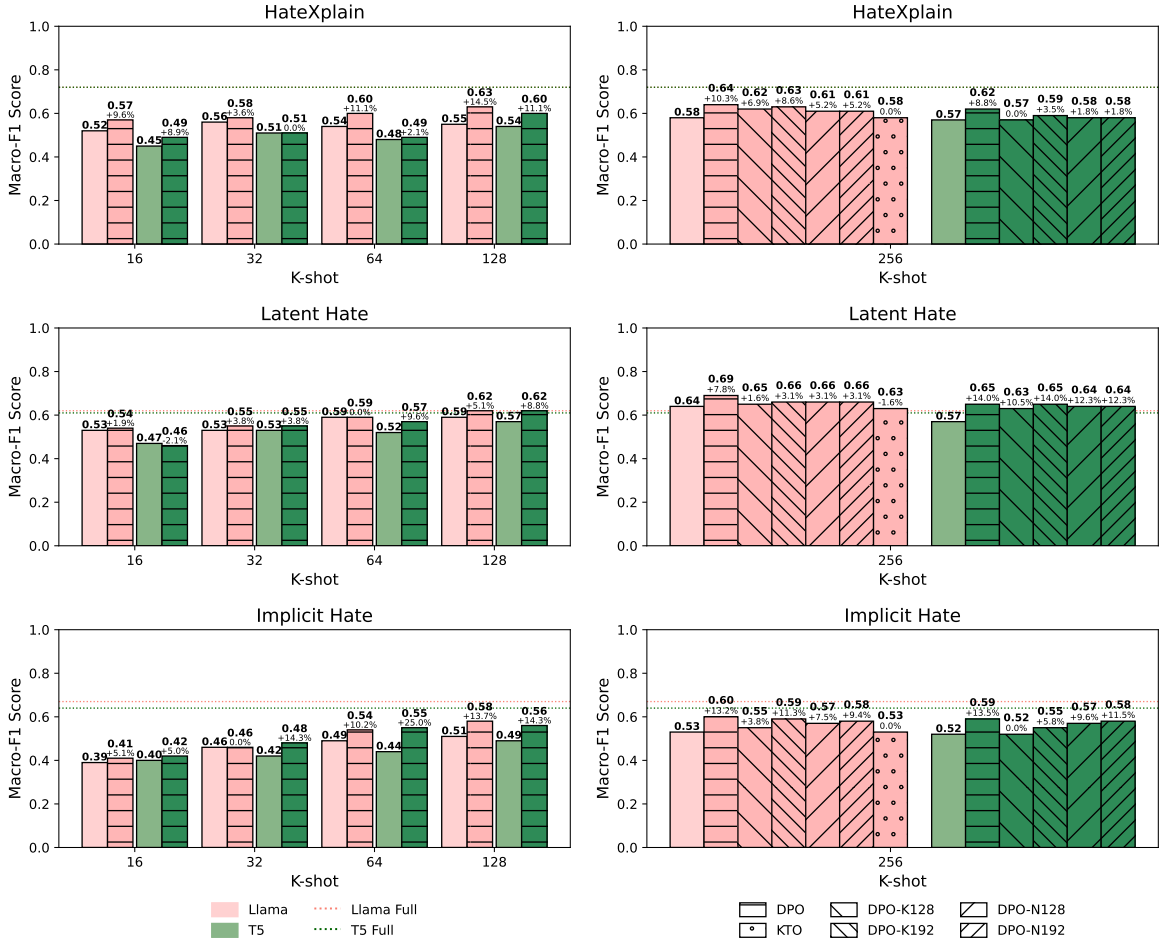


Figure 3: Bar plots for  $K$ -shot classification experiments on 3 datasets using *Llama* and *T5* models. Macro-F1 scores and percentage change over Baseline are displayed on top. Results for Baselines and DPO-augmented variants for  $K \in \{16, 32, 64, 128\}$  are displayed on the left subfigures. Results for  $K = 256$  of Baseline, KTO, DPO-augmented and its other variants on various sub-sampling strategies (section 5.1.1) are shown on the right. Horizontal lines show the F1 scores for *Full* models that use all training data.

*Reasoning* and *Alignment* as illustrated with the template in Figure 5. To avoid position bias, we present the pairs of explanations in randomized orders. All annotators are compensated fairly.

**Annotation Results** Table 2 shows the annotation results, aggregated per sample by majority voting. Across the 3 labels in *HateXplain*, annotators show no significant preference for explanations by either model on *Offensive* and *Hate*. On the other hand, *Llama*'s explanations for *Normal* posts are overwhelmingly preferred.

## 6.2 Cross-model Refinement

Motivated by *Llama*'s superior human-rated explanations, we investigate whether models can benefit from each other's outputs through cross-training. Below we describe our cross-model methodology, evaluate its impact on classification accuracy, and analyze how explanation styles evolve.

### 6.2.1 Cross-model Training Methodology

While *Llama* produces human-preferred explanations and is the natural deployment choice, we investigate whether models can adopt each other's explanation styles without degrading classification performance. Using the prompting technique described in Section 6.1, we use the SFT+DPO  $K = 128$  variants of both *Llama* and *T5* LLMs to generate explanations conditioned on the gold labels of the held-out 128-shot portion of the training data (unseen by either model). We then perform cross-model training: *T5* is finetuned on *Llama*'s outputs and vice versa, applying both SFT and DPO self-augmentation as in Section 5.1.

### 6.2.2 Cross-model Classification Results

Figure 4 shows the macro-F1 scores of both the original single-model SFT+DPO and cross-model finetuned variants on the test datasets.

Model	HateXplain			Latent Hate			Implicit Hate		
	F1	F1%	Data%	F1	F1%	Data%	F1	F1%	Data%
<i>SMARTER (Ours)</i>									
Llama_DPO-256	0.64	89%	6%	<b>0.69</b>	100%	7%	0.60	90%	57%
T5_DPO-256	0.62	86%	6%	0.65	94%	7%	0.59	88%	57%
<i>Full Training Baselines</i>									
Llama_Full	<b>0.72</b>	100%	100%	0.62	90%	100%	<b>0.67</b>	100%	100%
T5_Full	<b>0.72</b>	100%	100%	0.61	88%	100%	0.64	96%	100%
ModernBERT	0.70	97%	100%	0.61	88%	100%	0.64	96%	100%
<i>Commercial: Zero-shot</i>									
GPT-4o-mini	0.50	69%	–	0.54	78%	–	0.42	63%	–
GPT-4.1	0.55	76%	–	0.60	87%	–	0.57	85%	–
GPT-5-chat	0.56	78%	–	0.51	74%	–	0.58	87%	–
Qwen-32B	0.54	75%	–	0.47	68%	–	0.49	73%	–
<i>Commercial: 16-shot ICL</i>									
GPT-4o-mini	0.29±0.10	40%	–	0.25±0.09	36%	–	0.15±0.01	22%	–
GPT-4.1	0.52±0.07	72%	–	0.63±0.05	91%	–	0.38±0.11	57%	–
GPT-5-chat	0.62±0.01	86%	–	0.60±0.06	87%	–	0.40±0.11	60%	–
Qwen-32B	0.55±0.04	76%	–	0.57±0.03	83%	–	0.48±0.04	72%	–

Table 1: Comparison of performance and data usage across training paradigms. SMARTER models use DPO with  $K=256$  shots; commercial models evaluated in zero-shot and 16-shot in-context learning (ICL, mean±std over 3 seeds). We report macro-F1 (**F1**), relative score (**F1%**), and training data fraction (**Data%**). Models with the DPO suffix are DPO-augmented, while *Full* uses the entire training set. Within each dataset, best **F1** is bolded and **F1%** is normalized to that best model (100%). **Data%** is measured relative to the corresponding *Full* baseline (100%). ModernBERT is shown for reference.

Label	T5	Llama	Row Total
Normal	25 (25.5%)	73 (74.5%)	98
Offensive	63 (49.6%)	64 (50.4%)	127
Hate	54 (46.2%)	63 (53.8%)	117

Table 2: Annotator preferences for explanations on 342 *HateXplain* samples (majority vote).

**Cross-model training enhances T5 while impairing Llama.** Applying SFT training using explanations from *Llama* improves performance over the baseline *T5* SFT+DPO on *HateXplain* (F1 scores 0.62 from 0.60) and *Latent Hate* (0.64 from 0.62), with a slight reduction on *Implicit Hate* (0.54 from 0.56). On the other hand, applying SFT+DPO augmentation using *Llama*’s explanations from the complementary  $K=128$  shot data allows *T5* to **exceed** single-model performance on the full  $K=256$  set across the board. The new maximum macro-F1 scores for *T5* are 0.66 on *HateXplain* (up from 0.62 of single-model), 0.66 on *Latent Hate* (up from 0.65), and 0.61 on *Implicit Hate* (up from 0.59). Notably, these scores are superior to even *Llama*’s single-model performance at  $K = 256$  on *HateXplain* and *Implicit Hate*.

In contrast, SFT cross-training with *T5*’s output reduces *Llama*’s F1 scores to 0.58 from 0.63 on *HateXplain* while improving to 0.66 from 0.62 on *Latent Hate* and plateauing at 0.58 on *Implicit Hate*.

Additional SFT+DPO cross-model augmentation slightly boosts performance on *HateXplain* and *Implicit Hate* while slightly impairing on *Latent Hate*. None of the cross-model *Llama* variants matches the single-models’ F1 scores at  $K = 256$ .

### 6.2.3 Cross-model Style Analysis

We additionally train a style classifier based on the BERT architecture using explanations generated individually by *T5* and *Llama* of the SFT+DPO variant at  $K = 128$  (Devlin et al., 2019). A classification head is added to BERT, which leverages the *[CLS]* token from each explanation’s embedding to predict whether it was produced by *T5* or *LLaMA*. Technical details are reported in Appendix F.3.

Figure 12 shows individually trained models exhibit their own styles, while cross-model SFT aligns outputs with the alternative’s style. Except for *T5* on *HateXplain*, cross-model DPO slightly amplifies the base model’s style, suggesting this mechanism induces mild stylistic convergence.

**Cross-model explanations may improve weak categories in weaker models.** We compare *Precision*, *Recall*, and *F1-score* of base and cross-trained models at  $K = 256$  (Figure 11) to assess cross-model refinement. Improvements often occur in categories where the paired model excels, indicating transferable reasoning patterns, while declines appear when the counterpart is weak. This

Table 3: NLI-based consistency between explanations, predicted labels, and label definitions across datasets. Both DPO and cross-refined (XMOD) models show high entailment rates (>96%) with marginal percentages of contradictions (Contra.), neutral or undefined (Undef.) edge cases. Categories with all 0 values are omitted.

Dataset	Model	Train	Label Consistency (%)			Definition Consistency (%)		
			Entail	Contra.	Undef.	Entail	Contra.	Neutral
HateXplain	T5	DPO	99.2	0.8	0.0	98.2	1.5	0.3
		XMOD	96.7	1.5	1.8	99.0	0.9	0.2
	Llama	DPO	97.5	1.2	1.3	97.2	2.6	0.2
		XMOD	96.4	3.6	0.0	96.7	3.2	0.1
Latent Hate	T5	DPO	98.8	1.2	0.0	97.6	2.4	0.0
		XMOD	97.4	1.9	0.7	97.6	2.3	0.1
	Llama	DPO	97.3	2.7	0.0	97.3	2.3	0.3
		XMOD	96.8	3.2	0.0	97.5	2.5	0.0
Implicit Hate	T5	DPO	99.6	0.4	0.0	99.0	1.0	0.0
		XMOD	98.3	1.4	0.4	97.5	2.0	0.5
	Llama	DPO	98.1	0.4	1.5	97.5	1.8	0.8
		XMOD	96.9	2.8	0.1	96.0	3.6	0.4

asymmetry underscores our recommendation in line 19 of Algorithm 1 to validate and select the final model using a robust performance metric.

## 7 Analysis of Explanations

Using Natural Language Inference (NLI), we assess explanation consistency with labels and definitions, then analyze cross-model stylistic changes.

### 7.1 Analysis of Consistency

**Explanation-Label Consistency** We prompt GPT-4o-mini with the template shown in Figure 9 to judge whether the offered explanations are logically consistent with the predicted label for each sample across 3 test sets via the 4 categories: Entail, Contradict, Neutral and Undefined (noise cases where either the predicted label or the explanation is missing). An author’s independent annotation of 150 random samples from this pool under the same instruction achieves nearly unanimous agreement with GPT-4o-mini (Cohen’s  $\kappa = 0.92$ ), an observation inline with recent works that validate LLM-as-a-judge approach in NLI applications (Negru et al., 2025; Gallipoli and Cagliero, 2025).

In Table 3, Entail values exceeding 96% indicate that *explanations are largely consistent with their respective labels*. However, cross-model training induces a marginal 2-3% increase in Contradiction compared to the DPO augmentation.

**Explanation-Definition Consistency** Applying the same method to *only the Entail samples* from the previous analysis, using the prompt shown in

Figure 10 (Cohen’s  $\kappa = 0.96$ ), we again observe consistently high entailment rates and a slight increase in contradiction under cross-model training that mirrors the aforementioned trend.

Overall, *LLM-generated explanations remain predominantly consistent with both predicted labels and their definitions*. Cross-model refinement can enhance classification but may introduce a slight risk of inconsistency, warranting periodic monitoring. Appendix E offers qualitative analysis and insights for further quality checks for deployment.

## 8 Discussion and Conclusion

Our empirical analysis yields several key recommendations to guide practitioners in deployment.

**SMARTER delivers controllable, cost-efficient moderation.** Our staged pipeline achieves 86-100% of full-model performance using only 6-57% of training data (Table 1). Notably, we demonstrate this efficiency on challenging multi-class settings, whereas prior works tend to focus on simpler tasks (Guo et al., 2023; Kim et al., 2024). Unlike commercial APIs—where 16-shot ICL exhibits high variance ( $\pm 0.10$ ) and can degrade performance (Table 6)—SMARTER produces deterministic, explainable outputs suitable for production deployment at a fraction of the cost. When labeled data is scarce and explainability is required, SMARTER provides cost-efficient, transparent production deployment versus commercial APIs.

**Cross-model refinement enables targeted improvement.** T5’s superior cross-trained F1 scores

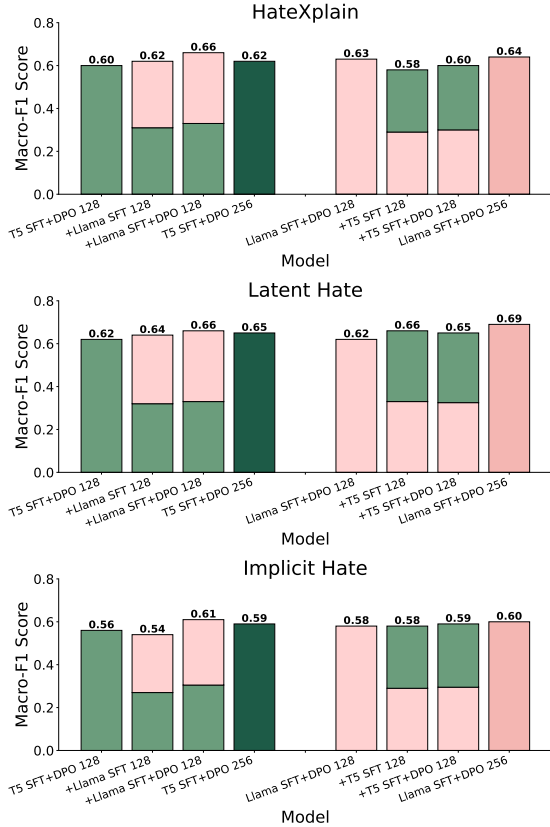


Figure 4: Macro-F1 scores on test portion of the 3 test sets for *T5* and *Llama* cross-model refinement regimen. In each figure: the first and last bars are scores reported with only *T5* model at  $K = 128$  and  $K = 256$  (directly from Figure 3); the middle bars are results after further finetuned using data from the complementary  $K = 128$  shots of counterpart model. Split-color bars: bottom color indicates the original model; top color indicates the counterpart model for additional cross-training.

(Figure 4) show that weaker models can adopt stronger models’ reasoning patterns. Practitioners facing category-specific weaknesses (Figure 11) can train multiple LLMs, then apply cross-model finetuning at high-quality checkpoints (e.g.,  $K = 128$ ) to transfer strengths. Architectural diversity (encoder-decoder vs. decoder-only) appears beneficial, as *Llama* enhances reasoning diversity that *T5* leverages (Hao et al., 2024).

**Human oversight remains essential.** Human evaluation (Table 2) shows category-specific preferences: *Llama* excels on *Normal* posts (3:1) while matching on *Hate*. Practitioners should periodically validate explanations, especially after cross-model training, to maintain consistency (Table 3).

We hope that SMARTER will enable practitioners to effectively and efficiently deploy explanation-based content moderation pipelines.

## 9 Limitations

We discuss several limitations in our research, along with their implications for future explorations.

**Generalization beyond English** Our work focuses exclusively on English corpora. In reality, social media exists for every language. As both English-based LLMs and research literature often attract much more attention, we hope this framework will be adapted for other languages.

**Augmentation to base framework** In our experiments, samples are randomly sampled from the label space. However, more strategic sampling choices (e.g., selecting based on how the level of classification difficulty calibrated on validation set) may boost performance. We encourage practitioners to explore augmentation to maximize the utility of our framework.

**Model selection** While we perform extensive experiments, our selection of 2 models could be expanded. At the cost of more training resource, training more models with different architectures or generation offer extra options for performance comparison for mix-and-match cross-training. We encourage other researchers to experiment with more LLMs using our framework.

**Limited scope of human validation** Due to budget constraints, we are only able to perform human validation for 1 dataset. In practice, it is recommended to solicit a larger pool of annotators to ensure appropriate amount of human oversight.

**Risk of bias in automated explanation** While our framework aims to improve performance and versatility in content moderation, it does not alleviate the fundamental sensitive nature of this task. Content moderation could be misappropriated to suppress free speech or harm marginalized groups (Dias Oliva, 2020; Kozyreva et al., 2023). Other works have noted that LLMs exhibit bias behaviors in certain applications (Nguyen et al., 2023; Gu et al., 2025). We urge researchers and practitioners to maintain vigilance when integrating our framework with human supervision to ensure ethical standards (Lai et al., 2022; Cao et al., 2024).

## 10 Ethical Considerations

The authors are not aware of any ethical problems in the development of this work. This research

uses publicly available dataset. We also further synonymize the content to the best of our ability to provide extra caution for privacy. We acknowledge that our framework and associated techniques could be abused for harmful purposes.

## 11 Acknowledgment

This work is funded by the NSF under Grant No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS). We thank the service of ACL ARR reviewers, area chairs and the editors of the ACL conference for our paper’s publication.

## References

- Saad Almohaimeed, Saleh Almohaimeed, Damla Turgut, and Ladislau Bölöni. 2025. Towards generalizable generic harmful speech datasets for implicit hate speech detection. *arXiv preprint arXiv:2506.16476*.
- Anjum and Rahul Katarya. 2024. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, 23(1):577–608.
- Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242.
- Stephanie Alice Baker, Matthew Wade, and Michael James Walsh. 2020. <? covid19?> the challenges of responding to misinformation during a pandemic: content moderation and the limitations of the concept of harm. *Media International Australia*, 177(1):103–107.
- Dmitriy Bespalov, Sourav Bhabesh, Yi Xiang, Liutong Zhou, and Yanjun Qi. 2023. Towards building a robust toxicity predictor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 581–598.
- Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408.
- Yang Cao, Lovely-Frances Domingo, Sarah Gilbert, Michelle Mazurek, Katie Shilton, and Hal Daumé Iii. 2024. Toxicity detection is not all you need: Measuring the gaps to supporting volunteer content moderators through a user-centric method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3567–3587.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 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.
- Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and violent behavior*, 58:101608.
- Peter Christen, David J Hand, and Nishadi Kirielle. 2023. A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):1–24.
- Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2024. Is explanation all you need? an expert survey on llm-generated explanations for abusive language detection. In *Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*.
- Thiago Dias Oliva. 2020. Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review*, 20(4):607–640.
- Ashwin Geet d’Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on: “Organization of Knowledge and Advanced Technologies”(OCTA)*, pages 1–5. IEEE.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

- 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.
- Giuseppe Gallipoli and Luca Cagliero. 2025. It is not a piece of cake for gpt: Explaining textual entailment recognition in the presence of figurative language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9656–9674.
- Katharine Gelber. 2021. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*.
- Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *arXiv preprint arXiv:2503.06778*.
- Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE.
- Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*.
- 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*.
- Hyesun Hwang and Kee-Ok Kim. 2015. Social media as a tool for social movements: The effect of social media use and social capital on intention to participate in social movements. *International Journal of Consumer Studies*, 39(5):478–488.
- Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddia, Nabil Arhab, et al. 2024. A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms. *arXiv preprint arXiv:2404.00303*.
- Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.
- Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16177–16188.
- Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. 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.
- Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120.
- Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *arXiv preprint arXiv:2403.08035*.
- Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of llms in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863.
- 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.
- Vlad Andrei Negru, Robert Vacareanu, Camelia Lemnar, Mihai Surdeanu, and Rodica Potolea. 2025.

- Morphnli: A stepwise approach to natural language inference using text morphing. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6938–6953.
- Huy Nghiem and Hal Daumé. 2024. [HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5938–5956, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Huy Nghiem, S. Panda, D. Khatwani, H. V. Nguyen, K. Kenthapadi, and H. D. Nguyen. 2025. [Balancing safety and helpfulness in healthcare ai assistants through iterative preference alignment](#). In *MLAH Symposium*. ArXiv preprint arXiv:2512.04210.
- Thu T Nguyen, Junaid S Merchant, Xiaohu Yue, Heran Mane, Hanxue Wei, Dina Huang, Krishik N Gowda, Katrina Makres, Crystal Najib, Huy T Nghiem, et al. 2024. A decade of tweets: Visualizing racial sentiments towards minoritized groups in the united states between 2011 and 2021. *Epidemiology*, 35(1):51–59.
- Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. Towards conceptualization of “fair explanation”: Disparate impacts of anti-asian hate speech explanations on content moderators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9696–9717.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Melissa Kazemi Rad, Huy Nghiem, Andy Luo, Sahil Wadhwa, Mohammad Sorower, and Stephen Rawls. 2025. Refining input guardrails: Enhancing llm-as-a-judge efficiency through chain-of-thought fine-tuning and alignment. *arXiv preprint arXiv:2501.13080*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- 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.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.
- Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fbert: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(4).
- Pabba Sumanth, Syed Samiuddin, K Jamal, Srikanth Domakonda, and Pathi Shivani. 2022. Toxic speech classification using machine learning algorithms. In *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, pages 257–263. IEEE.
- Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.
- Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.
- Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yılmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.
- Yulong Wang, Hong Li, and Ni Wei. 2025. Sahsd: Enhancing hate speech detection in llm-powered web applications via sentiment analysis and few-shot learning. In *Proceedings of the ACM on Web Conference 2025*, pages 3014–3025.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

- Jason Wei, Maarten Bosma, Vincent Y 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. *arXiv preprint arXiv:2109.01652*.
- 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.
- Steven Windisch, Susann Wiedlitzka, Ajima Olaghere, and Elizabeth Jenaway. 2022. Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell systematic reviews*, 18(2):e1243.
- 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, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, pages 500–507.
- Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. 2025. Self-adapting language models. *arXiv preprint arXiv:2506.10943*.

## Appendix

### A Amazon Mechanical Turk Annotation

We obtained IRB approval before soliciting annotation from the crowd workers. 14 workers were recruited to annotate the 342 explanations introduced in Section 6.1. Samples are assigned at random by the AMT platform to each annotator. Due to budget constraint, we opt to collect annotation for *HateXplain* to preserve statistical power instead of spreading across all 3 datasets. The breakdown of the annotators' demographic is below:

- ◇ **Gender:** Female (10), Male (3), Non-binary (1)
- ◇ **Age:** 18-29 (2), 30-39 (5), 40-49 (5), 50+ (2)
- ◇ **Race/Ethnicity:** Asian (3), Black (2), Hispanic/Latino (1), White (8)
- ◇ **Country of residence:** United States (14)
- ◇ **Education level:** 2-year college or equivalent (4), 4-year college or equivalent (6), High school or equivalent (3), Master degree or above (1)

### B Data Pre-processing

The datasets used in this work are open-source. We peruse them as research artifacts accordingly to their license. To further mitigate risks of privacy, we anonymize posts by replacing user handles with the string `<user>` and remove URLs if they appear in the context of the posts.

We obtain the seed explanations from the repository associated with [Nghiem and Daumé \(2024\)](#) and merge them with the correct identifier keys.

### C Obtaining Explanations for In-domain Data

The prompt template in [Figure 8](#) is used to collect explanation conditioned on a single label. Note that more labels (and their definitions) may also be added to this section as desired.

### D Stylistic Analysis

In [Figure 12](#), we show the distribution of *Llama* vs *T5* style pre- and post cross-model refined as recognized by the BERT style classifier.

### E Qualitative Analysis of Explanation

We examine the relatively small amount of non-Entail samples in Section 7 to observe general trends in the models' inconsistency with respect to both the predicted label and definition.

First, *T5* models typically exhibit higher consistency in explanation compared to their *Llama* counterparts as shown in [Table 3](#). Even after cross-model refinement (XMOD), the percentage of Contradiction increases by at most 1% (*Implicit Hate*). While a conclusive explanation is out of scope, we hypothesize that *T5*'s encoder-decoder architecture supports higher robustness compared to *Llama*'s decoder-only.

Second, qualitative analysis reveals different patterns of inconsistency in the explanations offered by these 2 models. For *HateXplain* and *Implicit Hate*, *Llama*'s Contradict-tagged explanations tend to consistently mentions some negative and/or offensive sentiments in the post, yet ultimately yielding the conflicting label *Normal* ([Figure 16](#)). In contrast, *T5* Contradict explanations – much fewer in overall number – generally asserts a label at the beginning yet explicitly rejects this label at the end. For the more challenging *Implicit Hate*, both models' Contradict explanations typically contain "lost in context" inconsistency, where the justification is grounded in some other labels' definition rather than the stated one, possibly due to the nuanced complexity of these finer-grained categories.

**Our NLI-based consistency check also emerges as a simple yet useful quality check for downstream deployment.** In addition to classification metrics, moderators may opt to implement a secondary lightweight NLI-checker to alert when a threshold of inconsistency is reached.

### F Technical Details

#### F.1 Hardware

All experiments are carried out on a maximum of 2 Nvidia RTX A6000 GPUs. We download the models from their Huggingface<sup>3</sup> checkpoints. Finetuning is implemented with LoRA techniques ([Hu et al., 2021](#)) via the TRL library<sup>4</sup>. The LoRA configurations for all settings are:

- ◇ Rank: 64
- ◇ *alpha*: 128
- ◇ Dropout: 0.05
- ◇ Target modules: *q* and *v* in projection layers

#### F.2 Inference Parameters

All models in our experiments use the following parameters during inference:

<sup>3</sup><https://huggingface.co/>

<sup>4</sup><https://huggingface.co/docs/trl/en/index>

Dataset	Total Size	Split Ratio	K val	K test	Platform	Label Space
HateXplain	20,148	60:20:20	50	400	G, T	Normal, Offensive, Hate
Latent Hate	19,112	60:20:20	50	400	T	Not Hate, Explicit Hate, Implicit Hate
Implicit Hate	4,153	50:20:30	50	150	T	White Grievance, Incitement to Violence, Inferiority Language, Irony, Stereotypes and Misinformation, Threatening and Intimidation

Table 4: Datasets for evaluating our framework. *Split Ratio* designates the proportion of the train:validation:test split. *K val* and *K test* are the number of data points per class to perform *K*-shot sampling from the Validation and Test set. For *Platform*, *G* denotes Gab, *T* for X (formerly Twitter).

Dataset	Model	Macro F1
HateXplain	Llama_OTS	0.52
	T5_OTS	0.56
Latent Hate	Llama_OTS	0.53
	T5_OTS	0.42
Implicit Hate	Llama_OTS	0.32
	T5_OTS	0.33

Table 5: Comparison of off-the-shelf (OTS) model performance across datasets. LLMs are prompted to produce answers without explanations in this setting as shown in Figure 6.

Dataset	Model	Min	Mean	Max	Range
HateXplain	GPT-4o-mini	0.22	0.29	0.42	0.20
	GPT-4.1	0.47	0.52	0.60	0.13
	GPT-5-chat	0.61	0.62	0.63	0.02
	Qwen-32B	0.52	0.55	0.59	0.07
Latent Hate	GPT-4o-mini	0.17	0.25	0.34	0.17
	GPT-4.1	0.58	0.63	0.67	0.09
	GPT-5-chat	0.54	0.60	0.65	0.11
	Qwen-32B	0.54	0.57	0.60	0.06
Implicit Hate	GPT-4o-mini	0.14	0.15	0.16	0.02
	GPT-4.1	0.30	0.38	0.51	0.21
	GPT-5-chat	0.29	0.40	0.51	0.22
	Qwen-32B	0.44	0.48	0.51	0.07

Table 6: Detailed 16-shot in-context learning (ICL) results for commercial models across 3 random seeds. For each dataset, we report the minimum, mean, and maximum Macro F1 scores, along with the range (Max – Min) to illustrate performance variability. Models were prompted with 16 examples per class. High variance, particularly for GPT-4o-mini and on complex tasks like Implicit Hate, demonstrates ICL’s brittleness for nuanced multi-class toxicity detection.

- ◇ temperature: 0.0
- ◇ max\_tokens/max\_new\_tokens:
  - ◇ 512 for classification with explanation
  - ◇ 20 for classification without explanation

The library `vLLM`<sup>5</sup> is used coupled with the `openAI v1/chatcompletion` endpoint to enable fast inference in our experiments. The Ope-

<sup>5</sup><https://docs.vllm.ai/en/latest/>

nAI model GPT-5-chat-latest endpoint was accessed on September 12, 2025. We also use the Qwen2.5-32B-Instruct-AWQ version on HuggingFace.

### F.3 Preference Optimization

**Direct Preference Optimization** DPO is an alignment technique used to fine-tune large language models (LLMs) to better align with human preferences. Unlike traditional methods such as Reinforcement Learning from Human Feedback (RLHF), DPO simplifies the process by directly optimizing the policy model using a binary cross-entropy objective, thereby eliminating the need for a separate reward model and the complexities of reinforcement learning. This approach reparameterizes the reward model such that its optimal policy can be expressed in a closed form, transforming preference optimization into a classification problem (Rafailov et al., 2024).

**Kahneman-Tversky Optimization** KTO is another model alignment technique that aims to directly maximize the utility of a model’s generations by applying principles from Kahneman-Tversky prospect theory (Kahneman and Tversky, 2013). KTO evaluates outputs based on their perceived gains or losses relative to a reference point, incorporating the concept of loss aversion. This method utilizes data that simply indicates whether an output is desirable or undesirable for a given input, which can be more readily available and less expensive to collect than the paired preference data required by DPO (Ethayarajh et al., 2024).

**Implementation** We primarily use the implementation of the `DPOTrainer` and `KTOTrainer` classes via the `TRL` library hosted on the HuggingFace platform. For the `DPOTrainer`, we set  $\beta = 0.1$ , and the loss to be default *sigmoid* loss. For the `KTOTrainer`, we similarly set the  $\beta = 0.1$  and use

the default KTO loss option.

#### **F.4 Hyperparameter Tuning**

For ModernBERT, we use the HuggingFace Trainer and train for 4 epochs with AdamW optimizer. We use the associated tokenizer and pad the inputs to maximum length (512) for each batch.

For Base SFT finetuning, we use 3 epochs and learning rate  $3e-4$ . We present the chosen parameters for DPO/KTO alignment tuning in [Table 7](#), [Table 8](#), [Table 9](#) below.

#### **G Examples of Explanations**

We provide some examples of the explanations generated during our cross-training phase in [Figure 13](#), [Figure 14](#), and [Figure 15](#).

K	Technique	T5		Llama	
		Epochs	LR	Epochs	LR
16	DPO	3	5e-05	3	1e-05
32	DPO	3	5e-05	4	1e-05
64	DPO	3	1e-05	3	1e-05
128	DPO	3	5e-05	3	1e-05
256	DPO	4	1e-04	3	1e-05
256	KTO	–	–	3	5e-07

(a) Hyperparameters for regular training.

K	Technique	T5		Llama	
		Epochs	LR	Epochs	LR
256	DPO-K128	3	1e-05	4	1e-04
256	DPO-K192	3	1e-04	1	7e-05
256	DPO-N128	3	5e-05	5	5e-06
256	DPO-N192	3	5e-05	5	5e-06

(b) Hyperparameters for subsampling training.

Table 7: *HateXplain*.

K	Technique	T5		Llama	
		Epochs	LR	Epochs	LR
16	DPO	3	5e-05	3	1e-06
32	DPO	3	1e-04	3	1e-06
64	DPO	3	5e-05	3	1e-06
128	DPO	4	5e-05	3	1e-04
256	DPO	3	1e-04	3	1e-04
256	KTO	–	–	3	5e-07

(a) Hyperparameters for regular training.

K	Technique	T5		Llama	
		Epochs	LR	Epochs	LR
256	DPO-K128	4	1e-04	4	1e-04
256	DPO-K192	3	1e-04	3	1e-04
256	DPO-N128	4	1e-04	3	1e-04
256	DPO-N192	4	1e-04	5	1e-04

(b) Hyperparameters for subsampling training.

Table 8: *Latent Hate*.

K	Technique	T5		Llama	
		Epochs	LR	Epochs	LR
16	DPO	3	5e-05	3	5e-06
32	DPO	3	5e-05	4	5e-05
64	DPO	3	1e-04	3	1e-06
128	DPO	1	7e-05	3	1e-04
256	DPO	1	5e-05	1	5e-05
256	KTO	–	–	3	5e-07

(a) Hyperparameters for regular training.

K	Technique	T5		Llama	
		Epochs	LR	Epochs	LR
256	DPO-K128	1	7e-05	1	1e-05
256	DPO-K192	1	7e-05	1	5e-05
256	DPO-N128	1	7e-05	1	1e-05
256	DPO-N192	1	7e-05	1	5e-05

(b) Hyperparameters for subsampling training.

Table 9: *Implicit Hate*

## Please read instruction carefully before attempting !

Instruction

Labels and Definitions

**Instructions:**

You are presented with a Post, a set of Labels and their Definitions, and two explanations EXP\_A and EXP\_B.

Your overall objective is to evaluate both explanations and choose the one that better provides a clear, logical, and convincing justification for the chosen label.

To achieve this, follow these steps:

- Review the entire set of labels and their definitions. Ensure that the label chosen in each explanation aligns with the content of the post.
- Evaluate both explanations based on the following criteria:
  - **Clarity:** How clearly is the reasoning communicated?
  - **Reasoning:** Is the reasoning behind the label choice logical and well-supported?
  - **Alignment:** Does the explanation adhere to the label's definition and apply it accurately to the post?
  - **Format:** The explanation should follow the format:  
EXPLANTION: [text] LABEL: [text]
- Finally, select the explanation that best fits the criteria above.

**Labels and Definitions:**

**Hate:** 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 for the context the word is used in the text.

**Offensive:** language that is rude, hurtful, derogatory, obscene or insulting, or to upset or embarrass people.

**Normal:** language that is neither Hate speech nor Offensive speech.

**Post:**

#{post}

**EXP\_A:**

#{exp\_a}

**EXP\_B:**

#{exp\_b}

**Select the better explanation:**

EXP\_A

EXP\_B

Submit

Figure 5: Template to collect annotation for preference on explanations for Amazon Mechanical Turks crowdworkers.

```
### Instruction:
By considering their corresponding Definitions, label the following post with
only one of these categories: {categories}.
Provide your response in the following format:
LABEL: [text]

### Definitions:
{definitions}

### Post:
{post}

### Response: [model output]
```

Figure 6: Prompt template for classification tasks **without explanation generation**. The model is instructed to directly output a label based on the provided definitions.

```
### Instruction:
By considering their corresponding Definitions, label the following post with
only one of these categories: categories. 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/are and how that justifies the chosen Label.
Provide your message in the following format:
EXPLANATION: [text]
LABEL: [text]

### Definitions:
Hate: Language that targets [...]
Abusive: Language that is [...]
Normal: Does not contain [...]

### Post:
{post}

### Response: [model output]
```

Figure 7: Prompt template for classification tasks. The definition block should contain the set's full label space

```
### Instruction:
By considering the following Post, the Label and its Definition, provide a brief
paragraph to explain step-by-step why the post should be classified with the
provided Label based on the given Definition.
If this post targets a group or entity relevant to the definition of the specified
Label, explain who this target is/are and how that justifies the chosen Label.
Provide your message in the following format:
EXPLANATION: [text]
LABEL: [text]

### Definitions:
Hate: Language that targets [...]

### Post:
{post}

### Response:
```

Figure 8: Prompt template to obtain explanation conditioned on a single label and its definition.

```
### Instruction:
You are given a model's prediction and its explanation for a social media post.
Your task is to judge whether the explanation is logically consistent with the
predicted label.

Possible outputs:
- ENTAIL: The explanation supports and is faithful to the predicted label.
- CONTRADICT: The explanation contradicts the predicted label.
- NEUTRAL: The explanation is unrelated, vague, or does not clearly support the
predicted label.
- UNDEFINED: Either the predicted label or the explanation is missing or empty.

Return only one word: ENTAIL, CONTRADICT, NEUTRAL, or UNDEFINED.

### Predicted Label:
{pred_label}

### Model Explanation:
{explanation}

### Response: [model output]
```

Figure 9: Prompt template for evaluating **explanation–label consistency**. The model is asked to judge whether a generated explanation is logically consistent with its predicted label.

```
### Instruction:
You are given a model's predicted label and its corresponding explanation.
Your task is to judge whether the explanation is logically consistent with the
definition of that label.

Possible outputs:
- ENTAIL: The explanation clearly aligns with and supports the definition.
- CONTRADICT: The explanation conflicts with the definition.
- NEUTRAL: The explanation is vague, unrelated, or does not clearly support the
definition.

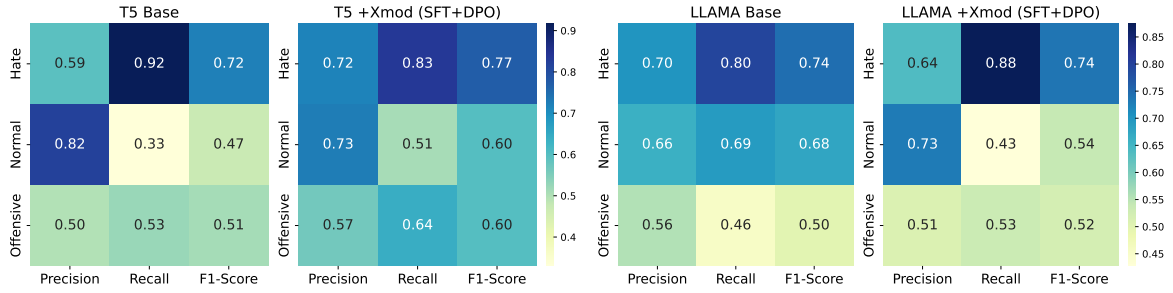
Return only one word: ENTAIL, CONTRADICT, or NEUTRAL.

### Official Definition:
{definition}

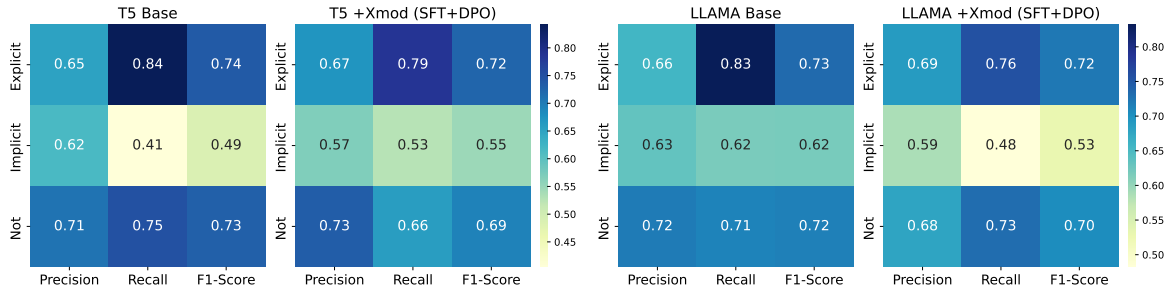
### Model Explanation:
{explanation}

### Response: [model output]
```

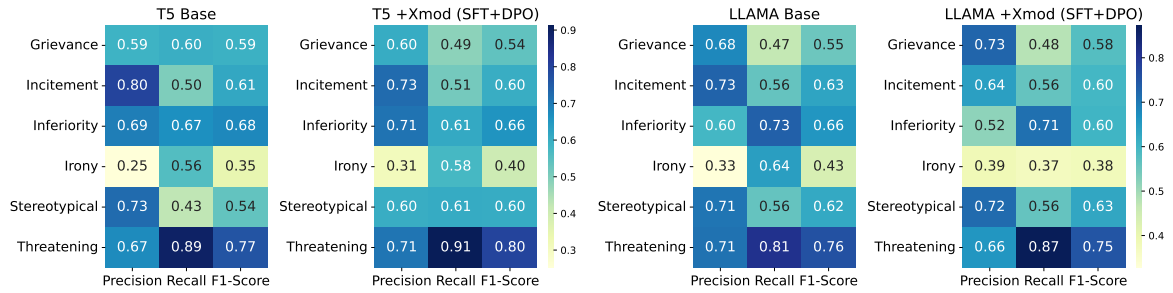
Figure 10: Prompt template for evaluating **explanation–definition consistency**. The model is asked to assess whether a generated explanation logically aligns with the formal definition of the predicted label.



(a) HateXlain



(b) Latent Hate



(c) Implicit Hate

Figure 11: Comparison of Precision, Recall, and Macro F1-scores for the *T5* and *Llama* model families under two training regimes: self-augmentation only (*Base*) and cross-model refinement (*+Xmod*) at  $K = 256$ . Models trained with cross-model refinement (right plot within each model family) generally show improvements in label categories where their counterparts already perform well, and conversely exhibit declines in areas corresponding to their weaknesses.

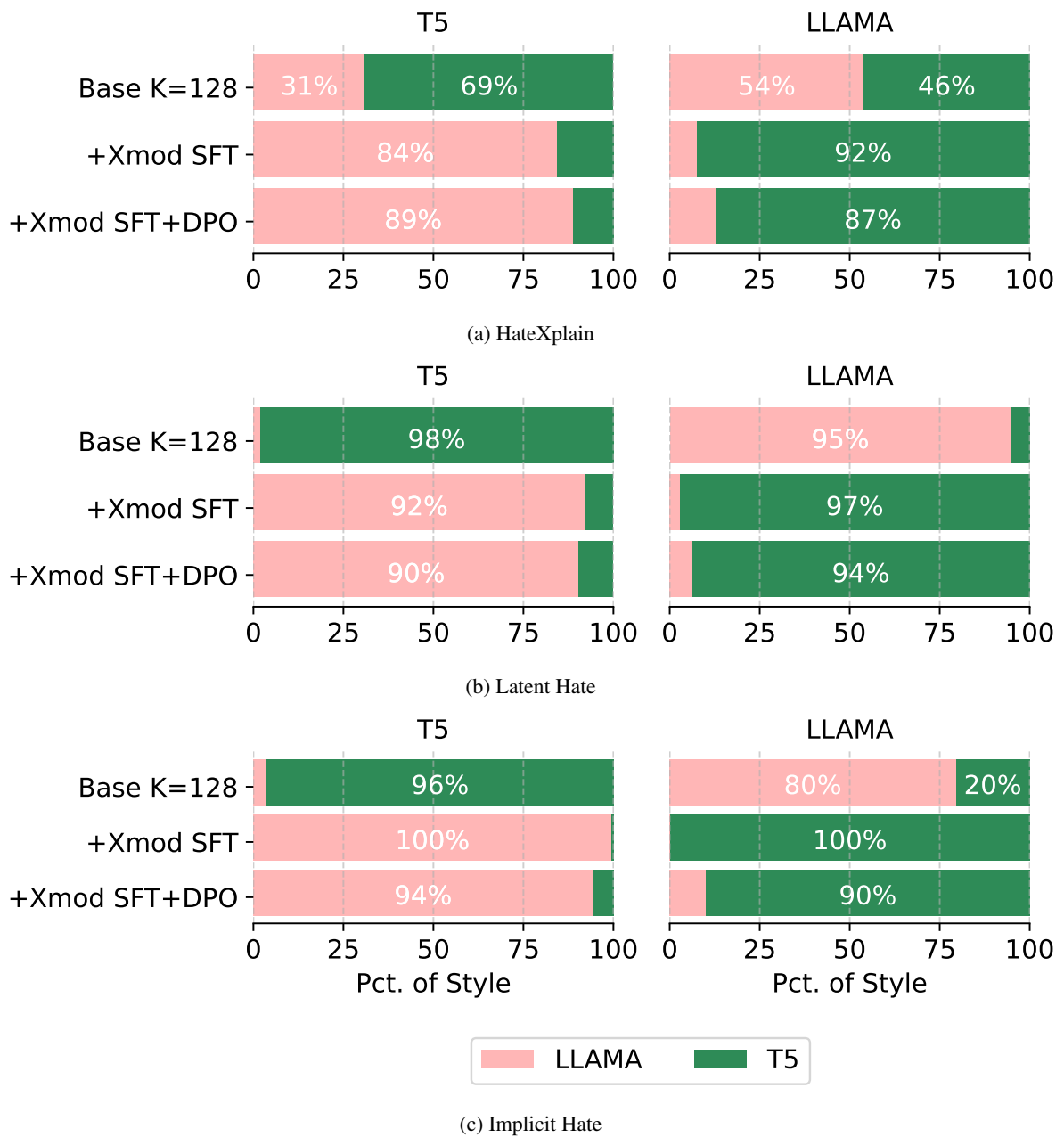


Figure 12: Percentage distribution of *Llama*- and *T5*-style explanations on test sets by BERT style classifiers, pre (*Base*) and post cross-model (*Xmod*) refined.

Dataset	Definition
HateXplain	<ul style="list-style-type: none"> <li>◇ <b>Normal:</b> neither hate speech nor offensive.</li> <li>◇ <b>Hate Speech:</b> 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.</li> <li>◇ <b>Offensive Language:</b> usage of rude, hurtful, derogatory, obscene or insulting language to upset or embarrass people.</li> </ul>
Latent Hate	<ul style="list-style-type: none"> <li>◇ <b>Not Hate:</b> speech or actions that do not involve any form of hatred, prejudice, or discrimination toward individuals or groups based on their characteristics.</li> <li>◇ <b>Explicit Hate:</b> openly expressed, direct forms of hatred and prejudice toward individuals or groups based on their characteristics.</li> <li>◇ <b>Implicit Hate:</b> coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity.</li> </ul>
Implicit Hate	<ul style="list-style-type: none"> <li>◇ <b>White Grievance:</b> 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.</li> <li>◇ <b>Inferiority Language:</b> 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.</li> <li>◇ <b>Irony:</b> refers to the use of sarcasm , humor, and satire to attack or demean a protected class or individual.</li> <li>◇ <b>Stereotypes And Misinformation:</b> 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.</li> <li>◇ <b>Threatening And Intimidation:</b> 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.</li> <li>◇ <b>Incitement to Violence:</b> includes flaunting in group unity and power or elevating known hate groups and ideologies</li> </ul>

Table 10: Definitions of toxic label categories and definitions (collected from their original publications) across datasets.

**POST:**

what kind of brzen hapless \*\*\* uses an expression like "there's no big \*\*\* there"

**LABEL:** HATE

(a) A sample post in *HateXplain* with gold label. Offensive content partially redacted.

**VARIANT:** T5 DPO @ K=128

**EXPLANATION:** The post targets the group of homosexuals by using a derogatory term "\*\*\*" to insult and demean them. This language is offensive as it is rude, hurtful, and insulting. Therefore, the post is labeled as Offensive.

**LABEL:** Offensive

**VARIANT:** T5 +XMOD SFT+DPO @ K= 256

**EXPLANATION:** The post targets the group of individuals who identify as "brzen" and use the derogatory term "\*\*\*" The language used in the post attacks and demeans this group based on their sexual orientation, which falls under the definition of hate speech.

**LABEL:** Hate

**VARIANT:** LLAMA DPO @ K=128

**EXPLANATION:** The post targets the homosexual community by using a derogatory slur "\*\*\*" to demean and insult them. This language is offensive as it is rude, hurtful, and insulting towards individuals based on their sexual orientation. Therefore, the post is labeled as Offensive.

**LABEL:** Offensive

**VARIANT:** LLAMA +XMOD SFT+DPO @ K=256

**EXPLANATION:** The post targets the homosexual community by using a derogatory term "\*\*\*" to insult and demean them. This language attacks and demeans a specific group based on their sexual orientation, which falls under the definition of hate speech. Therefore, the post is labeled as Hate.

**LABEL:** Hate

Figure 13: Examples of a post in *HateXplain* with gold label, along with the explanations of the *T5* and *Llama* self-augmented variants at K=128, and the cross-trained model using all K=256 in total. We observe that models trained with more data (+Xmod) gives the correct classification.

**POST:**

it means people aren't buying the lie! diversity k\*\*s now prove me wrong

**LABEL:** IMPLICIT HATE

(a) A sample post in *Latent Hate* with gold label. Offensive content partially redacted.

**VARIANT:** T5 DPO @ K=128

**EXPLANATION:** This post does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. It is a statement expressing skepticism towards the belief that diversity \*\*\*, but it does not disparage or express any form of hate towards any specific person or group. Therefore, it can be classified as "Not Hate."

**LABEL:** Not Hate

**VARIANT:** T5 +XMOD SFT+DPO @ K= 256

**EXPLANATION:** This post does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. The post simply states that people are not buying the lie that diversity \*\*\*, and asks for proof to prove the opposite. There is no indication of any negative or discriminatory language towards any specific group or entity. Therefore, this post should be classified as "Not Hate."

**LABEL:** Not Hate

**VARIANT:** LLAMA DPO @ K=128

**EXPLANATION:** The post uses coded or indirect language to disparage a person or group based on their race. Overall, the use of the phrase "diversity \*\*\*" and the challenge to "prove me wrong" indicate a negative view of diversity, suggesting a belief that diversity is harmful. This coded language indirectly disparages people of color, making it an example of implicit hate.

**LABEL:** Implicit Hate

**VARIANT:** LLAMA +XMOD SFT+DPO @ K=256

**EXPLANATION:** This post should be classified as "Not Hate" because it does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. The post simply expresses a belief that diversity is a lie and challenges the reader to prove the speaker wrong.

**LABEL:** Not Hate

Figure 14: Examples of a post in *Latent Hate* with gold label, along with the explanations of the *T5* and *Llama* self-augmented variants at K=128, and the cross-trained model using all K=256 in total. In this example, we observe that *T5* absorbs the verbosity of *Llama* after cross-training, yet still gets the wrong label. On the other hand, *Llama*'s explanation is more terse after cross-training, and also classifies the post incorrectly.

**POST:** hi can i get a couple of grammes of your \*\*\* power ?

**LABEL:** INCITEMENT

(a) A sample post in *Implicit Hate* with gold label. Offensive content partially redacted.

**VARIANT:** T5 DPO @ K=128

**EXPLANATION:** The post targets minorities by expressing frustration over their perceived privilege and casting the majority group (\*\*\*) as the real victims of racism. The post uses language that elevates \*\*\* power and unity, which is a known hate group ideology. The post also includes a sarcastic and ironic tone by asking for a "couple of grammes of your \*\*\* power," implying that minorities are inferior and need to be subjugated. This aligns with the definition of incitement, as it flaunts in-group unity and power while also elevating known hate groups and ideologies.

**LABEL:** Incitement

**VARIANT:** T5 +XMOD SFT+DPO @ K= 256

**EXPLANATION:** The post flaunts in-group unity and power by flaunting the idea of \*\*\* power. By using the term "\*\*\* power," the post elevates the idea of \*\*\* power and elevates the hate group ideology of \*\*\* supremacy. This aligns with the definition of incitement, which includes flaunting in-group unity and power or elevating known hate groups and ideologies.

**LABEL:** Incitement

**VARIANT:** LLAMA DPO @ K=128

**EXPLANATION:** This post targets \*\*\* people and their perceived desire for power and authority. By specifically asking for a couple of grammes of \*\*\* power, the post elevates and promotes the idea of \*\*\* supremacy, which is a hate group and ideology that advocates for the superiority of the \*\*\* race. This incitement of in-group unity and power aligns with the definition of incitement, as it promotes and supports a hate group and their extremist behavior.

**LABEL:** Incitement

**VARIANT:** LLAMA +XMOD SFT+DPO @ K=256

**EXPLANATION:** The post directly flaunts in-group unity and power by referencing \*\*\* power, which is a known hate group and ideology. The post specifically asks for a couple of grammes of \*\*\* power, which further promotes the idea of \*\*\* supremacy. This aligns with the definition of incitement, as it flaunts in-group unity and power and elevates known hate groups and ideologies.

**LABEL:** Incitement

Figure 15: Examples of a post in *Implicit Hate* with gold label, along with the explanations of the *T5* and *Llama* self-augmented variants at  $K=128$ , and the cross-trained model using all  $K=256$  in total. In this example, we also observe the cross-pollination of style in terms of verbosity, albeit to a lesser degree, but all models classify correctly.

#### **LLAMA - HATEXPLAIN**

EXPLANATION: The post targets the group of homosexuals by using the term "queer" in a derogatory manner. The language used attacks and demeans this group based on their sexual orientation, which falls under the definition of hate speech. Therefore, the post is labeled as "Normal."

LABEL: Normal

(a) Llama: The final label 'Normal' does not follow from the prior justification.

#### **T5 - LATENT HATE**

EXPLANATION: This post should be classified as "Not Hate" because it does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. The post simply mentions Alex Linder and white nationalism 1.0, without expressing any negative or discriminatory views towards any specific group or individual. Therefore, it does not meet the criteria for being classified as "Not Hate."

LABEL: Not Hate

(b) T5: The explanation starts with an asserted label yet later refutes this category.

#### **LLAMA - IMPLICIT HATE**

EXPLANATION: The post uses irony to attack and demean the idea of a wall, sarcastically suggesting that it should also prevent merpeople from the ocean. The use of humor and satire is used to mock the idea of a wall being a barrier to prevent all forms of transportation, including boats, tunnels, and airplanes, as well as the ocean.

LABEL: Threatening

(c) Llama: The explanation discusses factors directly relevant to 'Irony', yet the predicted label 'Threatening' is unrelated.

Figure 16: Examples of explanations that are judged to contain some forms of inconsistency with respect to the predicted label and/or its definition in our NLI-based analysis.