AWARE-HATE: Joint Sense–Subjectivity for Word-Level Hate Speech Detection

Anonymous ACL submission

Abstract

Word-level hate speech detection requires modeling both contextual meaning and annotator perspectives, yet current methods often overlook definitional sense and annotator subjectivity. We propose AWARE-HATE, a framework integrating dictionary definitions and annotator profiles. Our two-stage training establishes classification capability through initial supervised learning, then refines predictions via RL-based alignment. Experimental results demonstrate superior performance over fine-tuned LLMs, with ablations verifying that joint modeling of lexical sense and annotator subjectivity enhances detection efficacy.

Content warning: This paper contains potentially offensive language for illustration.

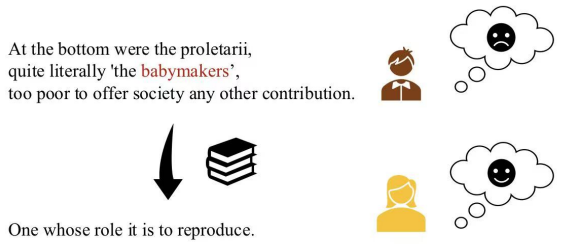

Figure 1: Word-level connotative variation. In the same context, the term *babymakers* elicits divergent hatefulness judgments across annotators—even when given the same dictionary gloss ("one whose role is to reproduce").

1 Introduction

Hate speech detection is a critical task in natural language processing (NLP) with substantial societal impact. While substantial progress has been made at the sentence and document levels (Gandhi et al., 2024), *word-level* detection remains underexplored despite its distinctive challenges. The meanings of candidate terms are rarely fixed; they are highly context-dependent and interpretation-sensitive, complicating both theoretical accounts and computational models of hate speech (Sayeed, 2013). Despite growing interest in Research on hate speech detection (HSD), there is limited systematic study of the sense and pragmatic mechanisms that govern how hateful word meanings are construed. Moreover, perceived hatefulness is inherently subjective: annotators frequently disagree even under identical contexts due to differing backgrounds and perspectives (Plank, 2022b), and models tend to underperform precisely on such contested cases. This observation aligns with a broader shift in NLP from relying on aggregated judgments to explicitly modeling inter-annotator variation (Uma et al., 2021; Basile et al., 2021).

As shown in Figure 1, in an identical context the term *babymakers* triggers divergent hatefulness judgments across annotators—even when they share the same dictionary gloss. This yields two core challenges for current HSD models: (i) dynamically reconciling definitional signals with rapidly evolving, context-dependent usage; and (ii) systematically accounting for the full spectrum of annotator perspectives to reduce coverage gaps and mitigate potential demographic bias. Meanwhile, efforts to model variation in subjective language interpretation have largely overlooked the *word-meaning* level. Prior work often foregrounds annotator features while underexamining how lexical senses and linguistic context interact to shape subjective judgments of hatefulness. (Hoeken et al., 2025).

We present AWARE-HATE, a two-stage framework for word-level hate speech detection that jointly models word sense and annotator subjectivity. In the supervised stage, the architecture employs DEFFUSION to fuse definitional with contextual senses, which substantially mitigates misclassification of hate-heterogeneous and pol-

ysemous target terms, and ANNOGATE to adaptively modulate annotator influences, improving generalization across annotators and reducing errors on high-disagreement subsets. trained with a multi-task objective that couples cross-entropy with contrastive regularization to capture interactions between objective sense and subjective perspectives. The backbone is trained with a multi-task objective that combines cross-entropy with contrastive regularization, encouraging representations to reconcile objective definitional cues with subjective perspectives. Building on this foundation, we apply GRPO as a second-stage distribution correction mechanism that targets difficult cases, leading to pronounced gains on disagreement and hate-heterogeneous subsets while leaving easy instances largely unchanged. Extensive experiments on the HATEWIC benchmark (Hoeken et al., 2024) show consistent improvements in macro-F_1 and enhanced robustness to inter-annotator disagreement, validating the effectiveness of joint sense-subjectivity modeling and our two-stage training approach.

2 Related work

2.1 Hate speech Detection

HSD has been dominated by transformer-based language models. Initial efforts focused on fine-tuning encoder-based models like BERT for classification (Caselli et al., 2021; Sarkar et al., 2021). Subsequent research has shifted towards prompt-based approaches with large language models (LLMs), leveraging their strong zero- and few-shot capabilities, particularly in instruction-tuned variants (Chiu et al., 2022; Plaza-del arco et al., 2023; Ronghao Pan, 2024). Building on the success of LLMs, recent work has begun applying preference optimization techniques to related tasks, such as sexism detection (Riahi Samani et al., 2025; Loftus et al., 2025) and counter-speech generation (Wadhwa et al., 2025).

However, most studies remain at the sequence level, underexploring word-level detection and fine-grained modeling of hateful connotation (cf. (Hoeken et al., 2024)). We address this gap by coupling word-meaning modeling with annotator-aware subjectivity modeling, enabling token-level judgments of a term's hateful connotation in context.

2.2 Word Meaning Modeling

Word meaning is context-dependent rather than fixed (Pustejovsky, 1991; Schütze, 1998; Haber and Poesio, 2024). In NLP, three task families operationalize such variation: Word sense Disambiguation (WSD) (Loureiro et al., 2021), Word sense Induction (WSI) (Eyal et al., 2022), and Lexical sense Change Detection (LSCD) (Periti and Montanelli, 2024). Most approaches rely on encoder-based contextual embeddings and sense-similarity metrics (Blevins and Zettlemoyer, 2020; Cassotti et al., 2023; Zhang et al., 2024), trained on general-purpose corpora. These models work well for standard sense resolution but are not tailored to hateful connotation or to fine-grained subjective context.

In hate-speech settings, however, connotation is central. Sequence-level studies leverage lexical features for detection (Koufakou et al., 2020; Zampieri et al., 2022) and analyze subtle devices such as dog whistles (Kruk et al., 2024; Mendelsohn et al., 2023), while other work focuses on swear words and slurs (Pamungkas et al., 2023; Hoeken et al., 2023), which are often treated as relatively stable across contexts (Frigerio and Tenchini, 1). More ambiguous pejoratives remain challenging (Dinu et al., 2021). Much of these studies frames hate identification as sequence classification, with limited explicit modeling of *intra-word* variation across contexts or individuals.

To address this gap, HateWiC (Hoeken et al., 2024) introduces a word-in-context benchmark that integrates dictionary definitions with annotator subjectivity; follow-up work further probes how definitional and linguistic cues interact with annotator perspectives (Hoeken et al., 2025). Yet current pipelines still (i) rarely ground hateful *connotation* in dictionary definitions and (ii) under-model subjective context. Accordingly, we introduce AWARE-HATEthat dynamically aligns definitional meaning with context while explicitly modeling annotator and reader subjectivity.

2.3 Subjectivity Perspectives Modeling

Traditional majority-vote aggregation can obscure individual differences (Davidson et al., 2019; Sap et al., 2021) and risk cultural biases (Elsafoury, 2022). A paradigmatic shift reconceptualizes annotator disagreement not as noise to be suppressed, but as an informative signal of underlying subjectivity (Larimore et al., 2021; Plank, 2022a; Fleisig et al., 2024). This perspective is especially

consequential for sentiment analysis and related tasks, where interpretive judgments are intrinsically shaped by individuals' experiences and social identities.

Research increasingly examines how annotator identity influences offensiveness perceptions. While some studies highlight sociodemographic factors like gender and age (Kocoń et al., 2021; Sang and Stanton, 2022), others emphasize ideological orientation (Sap et al., 2022) or moral values (Davani et al., 2024). Methodologically, subjectivity-aware modeling has incorporated annotator-specific signals in several ways, including embeddings from demographics, IDs, or labeling histories (Yin et al., 2023; Fleisig et al., 2023), contrastive learning of behavioral differences (Mokhberian et al., 2023), and more recently, gaze data (Alacam et al., 2024) and personalized LLM prompts (Orlikowski et al., 2025). Giorgi et al. (2025) further revealed the causal relationships between social attributes and labeling decisions. However, generalization to unseen annotators remains challenging. Our approach models free-form demographic descriptions and employs contrastive learning to capture both individual and group-level patterns, enhancing generalization.

2.4 Reinforcement Learning for Hate Speech Detection

Reinforcement learning from human feedback (RLHF) aligns language models with human preferences through supervised fine-tuning (SFT) followed by policy optimization, often using Proximal Policy Optimization (PPO) (Schulman et al., 2017). While effective in summarization (Stiennon et al., 2020) and instruction-following tasks (Ouyang et al., 2022), PPO requires substantial resources for value function training and large-batch updates.

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) offers an efficient alternative by replacing the explicit critic with group-wise reward statistics from multiple responses. This approach reduces computational costs while maintaining stability, as shown in reasoning tasks (Guo et al., 2025).

For hate speech detection, GRPO's group-wise comparison naturally handles annotation subjectivity. We adopt GRPO as a second-stage optimization after SFT to better capture nuanced annotator perspectives while ensuring training efficiency. The method's reward normalization effectively models the diverse judgments in hate speech labeling.

3 Method

This section details the architecture and training methodology of the proposed model, which tackles sense drift and subjectivity in word-level HSD by modeling the interaction between definitional and contextual senses and annotator representations, as illustrated in Figure 2.

3.1 Task Definition

Given a target term t and its specific context c, we retrieve k definitions of t from Wiktionary Dictionary, forming the set $D = \{def_1, def_2, ..., def_k\}$. The annotator description a captures variations in individual backgrounds.

The model input is formalized as a quadruple (t, c, D, a), and the output is a hate sentiment label y chosen from a label space $y \in \{\text{hate}, \text{not}\}$.

3.2 Method Overview

As shown in Fig. 2, AWARE-HATE consists of DEFFUSION, ANNOGATE, and a classifier, trained in two stages: supervised fine-tuning followed by preference-aligned RL. DEFFUSION integrates contextual and definitional signals to build the target-term representation, while ANNOGATE injects annotator-specific subjectivity via a gating mechanism before final classification. Starting from the SFT checkpoint, we apply GRPO-style RL by sampling multiple candidates per input, scoring them with a reward model, and optimizing a KL-regularized objective to better align model outputs with human preferences.

3.3 DefFusion

To address the polysemy and sense drift of target terms across contexts, we adopt a bi-encoder architecture following Blevins and Zettlemoyer (2020). Two independent encoders to represent the surrounding context and the candidate definitions. Concretely, we encode the context sentence c and extracts the target representation $\boldsymbol{E}_{\text{t}}$ at the position of term t:

$$\boldsymbol{E}_{\text{t}} = \text{ContextEnc}(c)[\text{idx}(t, c)] \in \mathbb{R}^{L_t \times d}, \quad (1)$$

where $\text{idx}(t, c)$ denotes the position of t in c; L_t denotes the number of subword tokens comprising t; and d is the embedding dimensionality. Each definition $def_i \in D$ is encoded and mean-pooled independently to obtain a per-definition vector $\mathbf{h}_i^{def}$:

$$\mathbf{h}_i^{def} = \text{Mean}\big(\text{DefEnc}(def_i)\big) \in \mathbb{R}^d, \quad (2)$$

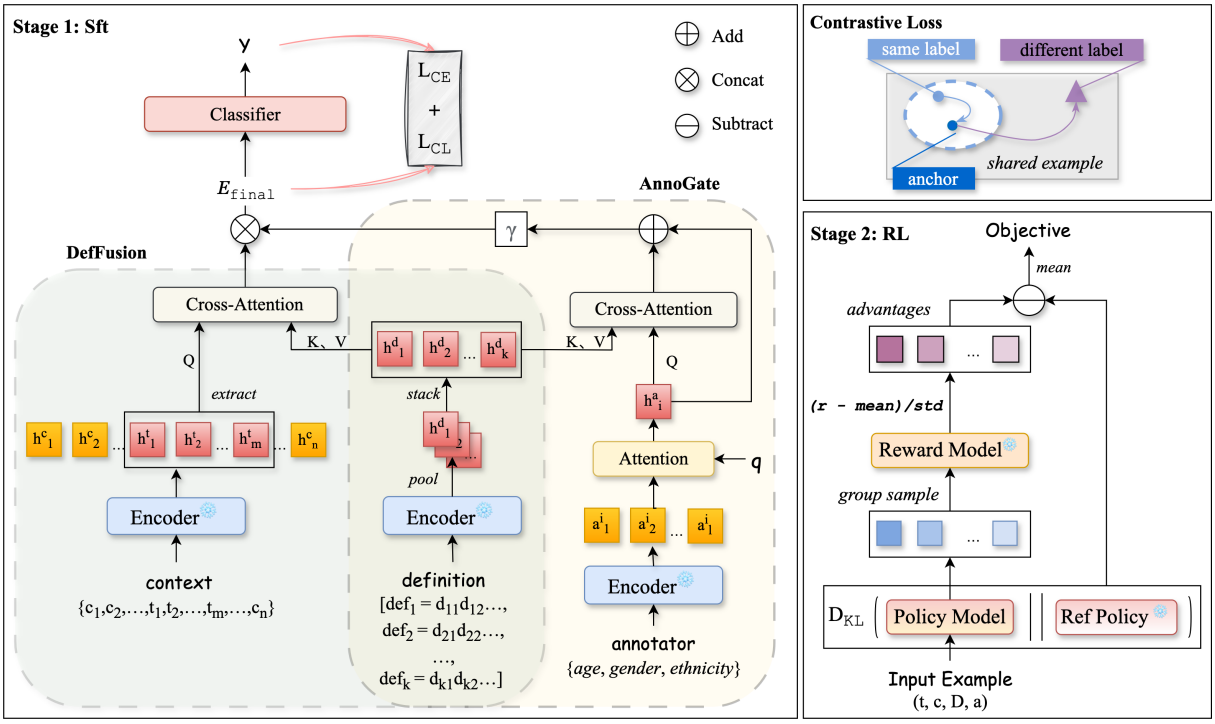

Figure 2: Overview of the proposed AWARE-HATE. **Stage 1 (Sft, left):** A context encoder produces token representations; the target term vector is extracted and fused with definition and annotator features via DEFFUSION and ANNOGATE. The resulting vectors are combined (addition⊕ / concatenation⊗ / subtraction⊖) to form $\mathbf{E}_{\text{final}}$, trained with $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{CL}}$ (top right). **Stage 2 (RL, bottom right):** The policy and reference models generate a group of outputs, which are scored by a reward model; group-normalized advantages are used to optimize an RL objective with a KL penalty.

where $i \in \{1, \ldots, k\}$ indexes the definitions in D, and $\text{Mean}(\cdot)$ denotes token-wise averaging.

Unlike prior work that ranks candidate definitions using cosine similarity, we retain all definitions and aggregate them with the target term via cross-attention. Firstly, $\{\mathbf{h}_i^{def}\}$ are stacked to form the definitions matrix:

$$\boldsymbol{E}_{\text{def}} = \text{Stack}_{i=1}^{k}\ \mathbf{h}_i^{def} = \begin{bmatrix} \mathbf{h}_1^{def} \\ \vdots \\ \mathbf{h}_k^{def} \end{bmatrix} \in \mathbb{R}^{k \times d}. \quad (3)$$

Then we perform context–definition cross-attention, using $\boldsymbol{E}_{\text{t}}$ as the query and $\boldsymbol{E}_{\text{def}}$ as keys and values, to obtain a senses-enhanced target embedding $\boldsymbol{E}_{\text{w}}$:

$$\boldsymbol{E}_{\text{w}} = \text{CrossAttn}(\boldsymbol{E}_{\text{t}},\ \boldsymbol{E}_{\text{def}}) \in \mathbb{R}^{L_t \times d}. \quad (4)$$

Instead of score-based single-definition selection, the module refines definitions with contextual cues and fuses all definitions via cross-attention, yielding a more complete senses representation that improves hate speech detection.

3.4 AnnoGate

Given the subjectivity of hate speech detection and the diversity of annotator preferences, we incorporate an annotator description text a. Following Hoeken et al. (2024), we linearize the annotator's demographic attributes into a short natural-language template (Reader is *AGE*, *GENDER*, and *ETHNICITY*), and encode it with an annotator encoder to obtain the annotator embeddings:

$$\boldsymbol{E}_{\text{a}} = \text{AnnEnc}(a) \in \mathbb{R}^{L_a \times d}, \quad (5)$$

where L_a denotes the length of a.

Rather than directly applying mean pooling to $\boldsymbol{E}_{\text{a}}$ and then concatenate it to the target vector as in prior work, we introduce a learnable query $q_a \in \mathbb{R}^d$ and perform single-query cross-attention over $\boldsymbol{E}_{\text{a}}$ to obtain an annotator embedding $\boldsymbol{h}_{\text{a}}$ that emphasizes cues most predictive for HSD:

$$\boldsymbol{h}_{\text{a}} = \text{Attn}(q_a,\ \boldsymbol{E}_{\text{a}}) \in \mathbb{R}^{1 \times d}. \quad (6)$$

To personalize senses interpretation with annotator background, we let $\boldsymbol{h}_{\text{a}}$ query the definition set $\boldsymbol{E}_{\text{def}}$ (Sec. 3.3) via cross-attention and add the returned signal back to the annotator vector to obtain

an updated annotator embedding:

$$\tilde{\boldsymbol{h}}_{\text{a}} \;=\; \text{Attn}(\boldsymbol{h}_{\text{a}},\; \boldsymbol{E}_{\text{def}}) + \boldsymbol{h}_{\text{a}} \in \mathbb{R}^{1\times d}. \quad (7)$$

A learnable gate $\boldsymbol{\gamma}_a \in \mathbb{R}$ then modulates the annotator signal by globally scaling the updated representation $\tilde{\boldsymbol{h}}_{\text{a}}$, allowing the model to regulate the annotator influence:

$$\boldsymbol{H}_{\text{a}} \;=\; \gamma_a\, \tilde{\boldsymbol{h}}_{\text{a}} \in \mathbb{R}^{1\times d}, \quad (8)$$

where $\boldsymbol{H}_{\text{a}}$ is the final subjectivity-aware annotator embedding.

We then construct the fused representation by concatenating the target word senses $\boldsymbol{E}_{\text{w}}$ with annotator $\boldsymbol{H}_{\text{a}}$, and adding type embeddings $\boldsymbol{e}_{\text{type}}$ to mark the source of each position in the sequence:

$$\boldsymbol{E}_{\text{final}} = [\boldsymbol{E}_{\text{w}} \,\|\, \boldsymbol{H}_{\text{a}}] + \boldsymbol{e}_{\text{type}} \in \mathbb{R}^{(L_t+1)\times d}, \quad (9)$$

where $[\cdot \,\|\, \cdot]$ denotes concatenation. And let $\boldsymbol{E}_{\text{type}} \in \mathbb{R}^{2\times d}$ be a lookup table with vectors for the target tokens and the annotator token. we broadcast $\boldsymbol{e}^{(0)}_{\text{type}}$ to the first L_t rows (target term) and $\boldsymbol{e}^{(1)}_{\text{type}}$ to the last row (annotator) before the element-wise addition. This role encoding helps the model distinguish lexical content from annotator information during fusion.

To model annotator disagreement, we apply contrastive learning to examples from the same context. For each example i, we obtain a normalized representation $\mathbf{z}_i$ from $\boldsymbol{E}_{\text{final}}$. Within the local group S_i sharing the same context, we define positive samples P_i (same label) and negatives N_i (different labels). The contrastive loss is:

$$\mathcal{L}_{\text{CL}} \;=\; -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \log \frac{\sum_{j\in P_i} \exp\!\left(\mathbf{z}_i^{\top}\mathbf{z}_j/\tau\right)}{\sum_{k\in S_i} \exp\!\left(\mathbf{z}_i^{\top}\mathbf{z}_k/\tau\right)}, \quad (10)$$

where $\tau = 0.1$ is the temperature, and $\mathcal{V}$ contains samples with at least one positive.

3.5 Training Strategy

Our training procedure consists of two stages: supervised fine-tuning followed by reinforcement learning with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). This approach ensures the model first acquires robust classification capabilities, then further aligns its behavior with task objectives while maintaining stability.

Stage 1: Supervised Fine-Tuning. The model is first trained using a combined objective of cross-entropy loss and contrastive loss:

$$\mathcal{L}_{\text{SFT}} = (1-\alpha)\,\mathcal{L}_{\text{CE}} + \alpha\,\mathcal{L}_{\text{CL}}, \quad (11)$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss, and $\mathcal{L}_{\text{CL}}$ is the contrastive loss defined in Section 3.4. The hyper-parameter $\alpha \in [0,1]$ balances between label prediction and representation learning, promoting both classification performance and discriminative embeddings.

Stage 2: Reinforcement Learning. Stage 1 already yields a robust supervised model that captures both sense information and annotator-specific subjectivity. Its training signal, however, remains instance-level: while cross-entropy and contrastive losses promote higher probability for the gold label and clustering of similar examples, they do not explicitly reason over *sets* of plausible outputs. In practice, many word-level hate instances in HATEWIC are ambiguous or noisy. For such cases, we aim to adjust the model's predictions without retraining the whole classifier.

To this end, Stage 2 treats the Stage 1 model as a reference policy $\pi_\theta(y \mid x)$ and applies GRPO as a lightweight second-stage refinement. For each input x, the current policy samples a small group of G candidate outputs at temperature T,

$$y^{(k)} \sim \pi_\theta(\cdot \mid x; T), \quad k = 1, \ldots, G,$$

and each candidate receives a sparse reward

$$r^{(k)} = \mathbb{I}[y^{(k)} = y^{\star}],$$

where $y^{\star}$ is the gold label. Group-wise normalization turns these rewards into advantages $A^{(k)}$ that encode how much better or worse a candidate is relative to its peers. A PPO-style clipped surrogate objective then adjusts the policy so that higher-advantage candidates become more likely and lower-advantage ones less likely, while a KL penalty against a frozen Stage 1 reference policy and the original supervised losses act as regularizers:

$$\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{clip}} + \beta\,\mathcal{L}_{\text{KL}} + \gamma_1\,\mathcal{L}_{\text{CE}} + \gamma_2\,\mathcal{L}_{\text{CL}}. \quad (12)$$

Conceptually, this stage does not change the supervision signal, but introduces a different optimization objective. GRPO works with groups of sampled outputs and their relative rewards, providing a sequence-level, preference-style update that

complements pointwise cross-entropy. In experiments, this yields consistent gains, especially on disagreement-prone and hate-heterogeneous subsets, without affecting performance on clear, high-agreement examples.

4 Experiments

4.1 Implementation Details

We implement all models in PyTorch using the HuggingFace Transformers library and optimize with AdamW and a linear learning rate schedule, with a base learning rate of 1×10^{-5}. In Stage 1 (supervised fine-tuning), the contrastive loss weight is set to $\alpha = 0.2$. In Stage 2 (GRPO), we use a group size of $G = 6$, sampling temperature $T = 1.2$, KL penalty coefficient $\beta = 0.05$, PPO clipping parameter $\epsilon = 0.2$, and supervised regularization weights $\gamma_1 = 0.5$ (cross-entropy) and $\gamma_2 = 0.2$ (contrastive loss). All encoders are based on `bert-base-uncased`, and the final classifier is a two-layer MLP with hidden size 768 and ReLU activation.

Experiments are run on a single NVIDIA GeForce RTX 2080 Ti GPU (11GB VRAM) on a Linux server.

4.2 Dataset

We conduct experiments on the HATEWIC dataset (Hoeken et al., 2024), a word-in-context benchmark for hate speech with annotator subjectivity. The corpus contains 4,027 unique instances and 12,442 annotations, obtained from three annotators per instance. Each example consists of a target term t, its context sentence c, a set of dictionary definitions D from Wiktionary, and annotator demographic attributes (age, gender, ethnicity).

For our main experiments, we merge *Strongly hateful* and *Weakly hateful* into a single *hate* class, discard *Cannot decide* instances, and keep *Not hateful* as *not*. Since *not* is more frequent than *hate*, we use macro-F_1 and class-wise F_1 as primary evaluation metrics.

4.3 Evaluation Metrics

We report macro-F_1 on the canonical label set as the main metric and additionally provide per-class F_1 to analyze class-specific behavior. Unless otherwise noted, results are averaged over ten-fold cross-validation, where each fold uses an 80/10/10 split for train/dev/test.

To assess generalization, we consider three data partitioning regimes: (i) *Random*, which splits by example to assess robustness to unseen contexts; (ii) *Out-of-Vocabulary (OoV)*, which splits by target term to test zero-shot generalization to unseen terms; and (iii) *Annotator*, which trains on a subset of annotators and evaluates on held-out annotators to examine subjectivity generalization. All metrics are computed on canonicalized labels for comparability across settings.

4.4 Baselines

To benchmark AWARE-HATE, we follow Hoeken et al. (2024) and construct three BERT-based inputs per instance: a word in context, a definition set, and the annotator description. We encode the sentence with BERT and extract the hidden state at the target subword position(s), averaging over subwords when the target term subwords. We encode the dictionary definitions and apply mean pooling to obtain a definition representation. For annotators, we encode the demographic template *"Reader is [AGE], [GENDER] and [ETHNICITY]."* with the same encoder and mean-pool to obtain the annotator embedding. These representations are concatenated and then fed into an MLP classifier. All baselines share the same preprocessing and train/dev/test splits for fair comparison.

4.5 LLM Baselines

We further include LLM baselines, all evaluated under the Random split (see Table 1). First, we train preference-optimized models with Direct Preference Optimization (DPO) starting from two Llama-3 checkpoints: `Sft` (supervised fine-tuned)[1] and `Instruct` (SFT followed by RLHF)[2] (Loftus et al., 2025). The prompt is adapted to our setting by including the context sentence, target term, its definition(s), and the current annotator profile; the full template is given in Appendix A.

Second, we evaluate in-context few-shot baselines. Specifically, we test Qwen3-4B-Instruct-2507[3] (reported as `Qwen3-instruct`), Qwen3-8B[4] (reported as `Qwen3`), and DeepSeek-LLM-7B-Base[5] (reported as `deepseek-base`). For all models, we use the same 4-shot prompt (Appendix A) with the four fields (Context, Target, Definition(s),

[1] huggingface.co/OpenRLHF/Llama-3-8b-sft-mixture
[2] huggingface.co/OpenRLHF/Llama-3-8b-rlhf-100k
[3] huggingface.co/Qwen/Qwen3-4B-Instruct-2507
[4] huggingface.co/Qwen/Qwen3-8B
[5] huggingface.co/deepseek-ai/deepseek-llm-7b-base

Model	Split / Protocol	$F1_{hate}$	$F1_{not}$	Macro-F_1
Panel A: Neural models				
BASELINE	Random	0.74	0.77	0.75
	OoV	0.74	0.76	0.73
	Ann	0.70	0.75	0.72
AWARE-HATE	Random	**0.78**	**0.80**	**0.79**
	OoV	0.77	0.78	0.78
	Ann	0.76	0.78	0.77
Panel B: LLM baselines				
Llama-3-Sft	DPO†	0.72	0.76	0.74
Llama-3-rlhf	DPO†	0.54	0.76	0.65
Qwen3-Instruct	Few-shot (ICL)	0.73	0.76	0.74
Qwen3	Few-shot (ICL)	0.69	0.77	0.74
DeepSeek-Base	Few-shot (ICL)	0.57	0.70	0.74

Table 1: Main results across data splits (per-class F_1 and macro-F_1).

Notes: Random/OoV/Ann denote the three split settings. Panel A reports neural models on all splits. Panel B reports LLM baselines under the Random split only. † DPO follows Loftus et al. (2025). **Bold** indicates the best score in each column.

Annotator profile) and require the model to output strictly one of {*hate*, *not*}. The outputs are post-processed using WordNet-based synonym expansion and lemmatization for robust label matching.

4.6 Results and Analyses

This section reports the main results, ablations, and qualitative analyses for AWARE-HATE. As detailed in Section 4.3, we evaluate our framework across multiple data-splitting protocols to assess model performance; we further conduct ablation studies to quantify the contribution of the key modules.

4.6.1 Main Results

Table 1 reports main results across splits. The baseline attains 0.75 Macro-F_1 on random but drops to 0.73 on both OoV and annotator splits, revealing limited robustness to unseen vocabulary and annotator distributions.

Our proposed model (AWARE-HATE) demonstrates robust performance across all settings, attaining the highest Macro-F_1 of 0.79 on the random split while maintaining strong results of 0.78 and 0.77 on OoV and annotator splits respectively. Notably, AWARE-HATE achieves the best per-class F_1 scores for both hate (0.78) and non-hate (0.80) classes in the random split setting.

For LLMs, both DPO-tuned Llama-3 variants, underperform compared to supervised method. We ascribe this mainly to objective mismatch (preference alignment rather than calibrated classification) and heightened prompt/format sensitivity. The SFT variant reaches 0.74 Macro-F_1 on random split, while the instruction-tuned version shows more significant degradation (0.65 Macro-F_1), particularly in the hate class (0.54 F_1).

Across all models, we observe consistent performance gaps between hate and non-hate classes, reflecting the dataset's inherent class imbalance. Furthermore, the performance degradation under OoV splits reveals a vocabulary generalization bottleneck, while the drops in annotator splits highlight challenges in generalizing across different subjective perspectives.

These findings underscore the importance of developing specialized architectures for hate speech detection that can better handle vocabulary variation and annotator subjectivity, rather than relying solely on general-purpose LLMs.

4.6.2 Ablation

We conduct comprehensive ablation studies under the random split configuration to systematically evaluate the contribution of each component in AWARE-HATE. The results are summarized in Table 3.

Optimization. Removing the contrastive loss $\mathcal{L}_{CL}$ leads to a Macro-F_1 drop from 0.79 to 0.78 ($\Delta = -0.01$), while disabling the RL stage $\mathcal{L}_{RL}$ further reduces performance to 0.77 ($\Delta = -0.02$). These results indicate that while the cross-entropy loss serves as the primary supervisory signal, the contrastive loss effectively captures inter-annotator subjective perspectives, and the RL further performs a distributional calibration through group-relative policy optimization on challenging samples. (Further analysis of RL is provided in the

Context (target underlined)	Definition	Annotator	Gold	Baseline	Ours
Anyone who would make a claim like this on national TV is a real asshat.	An obnoxiously ignorant person; a fool.	35-X-Black	HATE	**HATE**	**HATE**
We were ape over the new look.	Wild; crazy.	21-M-Black	HATE	NOT	**HATE**
These poor bastards started out life probably in bad or broken homes.	A person deserving of pity.	—	HATE	NOT	**HATE**

Table 2: Qualitative case studies illustrating definitional cues and annotator metadata. Targets are underlined. Annotator metadata follows *age–gender–ethnicity* ("X" denotes undisclosed). Gold and predictions are word-level labels. Correct predictions (matching Gold) are in **bold**. For details on the baseline, see Section 4.4.

data	variant	Macro-F_1	Δ
Full	Full AWARE-HATE	**0.79**	**0.00**
	w/o $\mathcal{L}_{CL}$	0.78	-0.01
	w/o $\mathcal{L}_{RL}$	0.77	-0.02
w/o Ann	w/o DefFusion	0.76	-0.03
	DEFFUSION only	0.77	-0.02
w/o Def	w/o ANNOGATE	0.74	-0.05
	ANNOGATE only	0.76	-0.03

Table 3: Ablation study analyzing model performance under different data settings and component configurations. $\mathbf{\Delta}$ represents the performance difference relative to the full model (0.79).

appendix. A)

Architecture. When annotator information is unavailable, removing DEFFUSION reduces Macro-F_1 to 0.76 ($\Delta = -0.03$), demonstrating that DEFFUSION effectively facilitates interaction between contextual and definitional information. Conversely, in the absence of definitional information, ablating ANNOGATE leads to the most significant performance reduction (0.74, $\Delta = -0.05$), indicating that ANNOGATE successfully models annotator-specific understanding of the target word in contexts.

The ablation results collectively confirm that: (i) both the contrastive loss and RL provide complementary gains beyond cross-entropy training; (ii) DEFFUSION and ANNOGATE effectively integrate objective sense information with subjective perspectives, each contributing substantially to the model's performance.

4.7 Qualitative Analysis

To better understand AWARE-HATE's behavior, we present three representative cases in Table 2.

In the first case, both the baseline and AWARE-HATE correctly classify the target word ("asshat") as hateful given clear contextual and definitional evidence. In the second case, AWARE-HATE outperforms the baseline by leveraging annotator information to capture a hateful interpretation of "ape" that is missed without modeling subjectivity. In the third case, despite unanimous labels across annotators, the baseline misclassifies "bastards" as non-hateful, whereas AWARE-HATE correctly recognizes its pejorative usage, suggesting a more effective balance between contextual cues and dictionary definitions.

Overall, these examples illustrate that AWARE-HATE improves robustness on challenging terms whose surface meanings can be misleading, by jointly integrating definitional knowledge with annotator modeling.

Limitations

Our work presents several limitations that warrant discussion. (i) Data Scope. All experiments are conducted on HateWiC, an English-language dataset. The framework's efficacy in cross-lingual, code-mixed, or domain-shifted scenarios remains unverified. (ii) Annotator Representation. We operationalize annotator identity via short, template-based demographic descriptors. This method offers a tractable but potentially coarse or incomplete approximation, which may not capture identity complexity and raises sensitivity concerns. (iii) Lexical Resource Quality. Definitions are sourced from crowd-sourced dictionaries, whose coverage is inherently uneven and may reflect community-specific biases. (iv) Modeling Granularity. While our model conditions on static demographic proxies, it does not incorporate dynamic annotator behaviors, richer interaction histories, or longitudinal preference drift. (v) Calibration and Fairness. Although we evaluate robustness to annotator disagreement, we do not conduct comprehensive subgroup fairness audits or detailed uncertainty calibration. Future work should include structured

fairness evaluations and human-model alignment studies. (vi) Computational Overhead. The introduction of DefFusion and AnnoGate modules adds 7M parameters to the frozen BERT backbone (6% of the total; see Appendix). We consider this a reasonable trade-off given performance gains on ambiguous content, yet the throughput–accuracy balance is not fully optimized. Exploration of model distillation or lighter architectures is deferred to future work. (vii) Ethical Considerations. The explicit modeling of sensitive attributes, while intended for transparency, risks reinforcing or proxying for protected characteristics. Any real-world deployment must be preceded by stringent data governance protocols and explicit consent mechanisms.

References

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes don't lie: Subjective hate annotation and detection with gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 187–205, Miami, Florida, USA. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

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, Online. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *Preprint*, arXiv:2109.00904.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2022. Detecting hate speech with gpt-3. *Preprint*, arXiv:2103.12407.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

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.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fatma Elsafoury. 2022. Darkness can not drive out darkness: Investigating bias in hate speechdetection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 31–43.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Aldo Frigerio and Maria Paola Tenchini. 1. Pejoratives: a classification of the connoted terms. *Rivista Italiana di Filosofia del Linguaggio*, 13(1).

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 653–670.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Sanne Hoeken, Ozge Alaçam, Dong Nguyen, Massimo Poesio, and Sina Zarrieß. 2025. Not just who or what: Modeling the interaction of linguistic and annotator variation in hateful word interpretation. In *16th International Conference on Computational Semantics*, page 73.

Sanne Hoeken, Sina Zarrieß, and Ozge Alacam. 2023. Identifying slurs and lexical hate speech via lightweight dimension projection in embedding space. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 278–289, Toronto, Canada. Association for Computational Linguistics.

Sanne Hoeken, Sina Zarrieß, and Özge Alaçam. 2024. Hateful word in context classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 172–186.

Yova Kementchedjhieva and Ilias Chalkidis. 2023. An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. *arXiv preprint arXiv:2305.05627*.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. Silent signals, loud impact: LLMs for word-sense disambiguation of coded dog whistles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12493–12509, Bangkok, Thailand. Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Frederick Liu, Terry Huang, Shihang Lyu, Siamak Shakeri, Hongkun Yu, and Jing Li. 2022. Enct5: A framework for fine-tuning t5 as non-autoregressive models. *Preprint*, arXiv:2110.08426.

Sebastian Loftus, Adrian Mülthaler, Sanne Hoeken, Sina Zarrieß, and Özge Alaçam. 2025. Using llms and preference optimization for agreement-aware hatewic classification. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 538–547.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *Preprint*, arXiv:2502.20897.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 57(1):155–188.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Barbara Plank. 2022a. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank. 2022b. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ali Riahi Samani, Tianhao Wang, Kangshuo Li, and Feng Chen. 2025. Large language models with reinforcement learning from human feedback approach for enhancing explainable sexism detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6230–6243, Abu Dhabi, UAE. Association for Computational Linguistics.

Rafael Valencia-García Ronghao Pan, José Antonio García-Díaz. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *Computer Modeling in Engineering & Sciences*, 140(3):2849–2868.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *Information for a Better World: Shaping the Global Future*, pages 425–444, Cham. Springer International Publishing.

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*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asad Sayeed. 2013. An opinion about opinions about opinions: subjectivity and the aggregate reader. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 691–696, Atlanta, Georgia. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *Preprint*, arXiv:1703.05175.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Sahil Wadhwa, Chengtian Xu, Haoming Chen, Aakash Mahalingam, Akankshya Kar, and Divya Chaudhary. 2025. Northeastern uni at multilingual counterspeech generation: Enhancing counter speech generation with LLM alignment through direct preference optimization. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 19–28, Abu Dhabi, UAE. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 902–913.

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. 2022. Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202–210, Marseille, France. European Language Resources Association.

Junwei Zhang, Ruifang He, Fengyu Guo, and Chang Liu. 2024. Quantum interference model for semantic biases of glosses in word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19551–19559.

A RL on Challenging Cases

Our second-stage training employs Group Relative Policy Optimization (GRPO) to refine model behavior on disagreement-prone and hate-heterogeneous cases. We evaluate the full AWARE-HATE('Full'), its supervised-only counterpart without RL ('w/o RL'), and the best prior baseline ('Baseline') on two types of challenging subsets from HateWiC: (i) instances where an annotator's label disagrees with the majority label (*Agreement w/ Majority = FALSE*); and (ii) examples whose senses definitions are *hate-heterogeneous*, i.e., the same gloss appears with both hateful and non-hateful majority labels in the dataset.

Table 4 reports accuracy under three data partitions, further divided into "easy" (majority-agreed or hate-homogeneous) and "hard" (disagreement or hate-heterogeneous) regions.

On easy cases with clear agreement and homogeneous senses, RL introduces only minor changes relative to the supervised-only model (e.g., from 0.83 to 0.86 on the random split), indicating that GRPO largely preserves performance on unambiguous inputs. In contrast, on disagreement-prone instances, the full model yields consistent though modest gains over the supervised variant, with the most notable improvement on the annotator split (0.27 to 0.36), suggesting that the preference-based objective helps align predictions with majority preferences precisely where annotator disagreement is highest. For hate-heterogeneous senses, RL generally brings small improvements (+0.02–0.04) on the hard subset while maintaining comparable performance on homogeneous senses, consistent with our goal of focusing RL updates on ambiguous lexical phenomena.

Overall, these results confirm that the RL stage is not the primary driver of performance—most of the macro-F_1 gains already arise from our sense–subjectivity architecture and multi-task supervised training. Instead, GRPO acts as a targeted refinement mechanism: it slightly improves agreement with annotator majorities on hard cases and helps the model better handle hate-heterogeneous senses, while leaving easy regions of the input space essentially unchanged. We view this as an

initial step toward preference-based tuning for subjective lexical phenomena, and we leave more sophisticated RL formulations and broader human preference signals to future work.

B Model Architectures and Extended Evaluation

B.1 Classification Architectures and Implementation

The broader literature distinguishes between *discriminative* architectures with fixed label spaces and *generative* text-to-text models with flexible output vocabularies, each offering complementary strengths (Kementchedjhieva and Chalkidis, 2023). Discriminative models typically yield strong in-domain accuracy, while generative approaches can better accommodate emerging labels and skewed class distributions (Liu et al., 2022). Encoder-only methods introduce inductive bias via prototype label vectors (Snell et al., 2017) or label-wise attention (LWAN) (Mullenbach et al., 2018), but their gains diminish once stronger encoders such as BERT or mT5 are used (Chalkidis et al., 2020, 2021; Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021). In contrast, encoder–decoder (text-to-text) models exploit label verbalizers and pretrained senses and often improve few- and zero-shot behavior (Raffel et al., 2020; Chung et al., 2022); for example, a non-autoregressive T5 variant that derives label-wise representations from the decoder surpasses encoder-only baselines on EURLEX-57K (Chalkidis et al., 2019).

Motivated by these trends, we perform a controlled comparison between the two paradigms for word-level hate detection under class imbalance and annotation noise. Prior work in hate speech detection predominantly adopts encoder–MLP classifiers. Here, we instead implement both (i) an encoder–decoder *generative* variant and (ii) an encoder–MLP *discriminative* variant within the same AWARE-HATEframework, so that architectural differences can be analyzed under identical training objectives and data splits.

Encoder–Decoder (Generative Variant). For the generative instantiation, we use T5 (Raffel et al., 2020) as the backbone. The decoder conditions on the fused representation $\mathbf{E}_{\text{final}}$ and generates a canonical label sequence autoregressively. Training uses teacher forcing with token-level cross-entropy:

$$\mathcal{L}_{\text{CE}}^{\text{gen}} = -\frac{1}{T_y}\sum_{t=1}^{T_y} \log p_\theta\big(y_t \mid y_{<t}, \mathbf{E}_{\text{final}}\big), \quad (13)$$

where $y_{1:T_y}$ denotes the canonicalized label string. At inference time, we apply beam search and map decoded sequences back to canonical labels via a synonym-normalization procedure.

Encoder–MLP (Discriminative Variant). For the discriminative instantiation, we follow standard practice and use a BERT encoder (Devlin et al., 2019) to produce $\mathbf{E}_{\text{final}}$, followed by a multilayer perceptron for classification. The model is trained with instance-level cross-entropy:

$$\mathcal{L}_{\text{CE}}^{\text{disc}} = -\log p_\theta\big(y \mid \mathbf{E}_{\text{final}}\big), \quad (14)$$

and predictions are obtained via `argmax` over the class logits.

Both variants are integrated into the same AWARE-HATEarchitecture and trained with the same multi-task and RL objectives as described in the main paper, enabling a fair comparison of their behavior under label skew and noisy, subjective annotations.

B.2 Extended Experimental Results

This section reports extended results for the binary, three-way, and four-way label configurations under three data partitioning regimes (random, OoV, and annotator splits), using only the Stage-1 supervised models without RL refinement. In all tables, **Baseline** denotes the best-performing encoder–MLP system from prior work, while **Gen** and **Disc** refer to our generative and discriminative instantiations of AWARE-HATE, respectively.

B.2.1 Binary Classification

Table 5 reports results for the binary setting. Overall, the discriminative variant (**Disc**) achieves the strongest aggregate performance, reaching 0.77 accuracy and 0.77 macro-F_1 on the random split and maintaining competitive scores under the OoV and annotator splits (0.76 and 0.75 macro-F_1, respectively). The generative variant (**Gen**) is slightly weaker in this coarse-grained setting but remains on par with the Baseline on most splits, indicating that it can still capture the main hate/non-hate distinction despite its more flexible label generation mechanism.

Model	Split	Agreement w/ Majority		Hate-Homogeneous	
		TRUE	FALSE	TRUE	FALSE
Baseline	Random	0.82	0.38	0.89	0.67
	OoV	0.83	0.37	0.88	0.67
	Annotator	0.80	0.33	0.88	0.63
w/o RL (SFT only)	Random	0.83	0.38	0.89	0.69
	OoV	0.82	0.39	0.88	0.68
	Annotator	0.84	0.27	0.92	0.65
Full (AWARE-HATE)	Random	**0.86** (+0.03)	0.38 (+0.00)	**0.93** (+0.04)	**0.71** (+0.02)
	OoV	**0.84** (+0.02)	0.39 (+0.00)	0.89 (+0.01)	**0.71** (+0.03)
	Annotator	**0.85** (+0.01)	**0.36** (+0.09)	0.91 (-0.01)	**0.69** (+0.04)

Table 4: Accuracy on challenging subsets with and without RL fine-tuning. We distinguish annotator-disagreement cases (*Agreement with Majority = FALSE*) and hate-heterogeneous cases (*Hate-Homogeneous = FALSE*). Numbers in parentheses denote absolute gains of AWARE-HATE-Full over the SFT-only variant.

model	split by	hate/1	not/0	Acc	macro-F_1
Baseline	random	0.74	0.77	0.75	0.75
	OoV	0.74	0.76	0.75	0.73
	annotator	0.70	0.75	0.73	0.72
Gen	random	0.73	0.77	0.75	0.75
	OoV	0.73	0.74	0.73	0.73
	annotator	0.71	0.74	0.72	0.72
Disc	random	0.75	0.78	0.77	0.77
	OoV	0.74	0.77	0.76	0.76
	annotator	0.74	0.76	0.75	0.75

Table 5: Binary classification results across different data splits. We report per-class F_1, overall accuracy (Acc), and macro-F_1.

B.2.2 Three-way Classification

Table 6 summarizes results for the three-way setting. All models find the *Weakly hateful* class challenging (0.32–0.40 F_1), reflecting the inherent ambiguity of borderline cases. Our discriminative variant (**Disc**) attains the strongest overall performance on the random split (0.65 accuracy, 0.59 macro-F_1) and remains robust under OoV and annotator splits (0.56 and 0.54 macro-F_1). The generative variant (**Gen**) yields slightly lower overall macro-F_1 (0.55 on random) but reaches the highest F_1 on *Weakly hateful* (0.40), suggesting that sequence generation can better accommodate minority classes and nuanced degrees of hatefulness.

B.2.3 Four-way Classification

The four-way configuration in Table 7 reintroduces the *Cannot decide* class, yielding a more realistic but substantially noisier scenario. Across models, macro-F_1 drops by roughly 0.16–0.19 compared to the three-way setting, confirming that fine-grained labels and noisy judgments substantially complicate the task. The discriminative variant (**Disc**) again achieves the strongest overall performance on the random split (0.62 accuracy, 0.43 macro-F_1), indicating that direct decision-boundary optimization is comparatively robust under label noise. At the same time, the generative model (**Gen**) exhibits a more balanced distribution of per-class F_1 scores and slightly higher performance on the minority and noisy *Cannot decide* class, highlighting its potential advantages when modeling rare categories and ambiguous cases.

Taken together, these extended results show that our framework can be instantiated with either generative or discriminative backbones: the discriminative variant excels in overall accuracy and macro-F_1, while the generative variant offers complementary strengths in handling label sparsity, ambiguous categories, and noisy real-world annotations.

C Computational Overhead of AWARE-HATE

We here quantify the additional computational cost introduced by the DEFFUSION and ANNOGATE modules, relative to the baseline (see Section 4.4).

Parameter counts. Table 8 details the parameter allocation in our implementation. The frozen BERT backbone constitutes 109M parameters, representing the majority of both parameters and computational load. The baseline model adds only 0.08M trainable parameters to this backbone. In comparison, AWARE-HATE incorporates three multi-head attention modules, two for DEFFUSION and one for ANNOGATE, along with supplementary LayerNorm and embedding layers, bringing the total to 7.2M trainable parameters and 116M overall.

Each cross-attention module contributes approx-

model	split by	Sh	Wh	Nh	Acc	macro-F_1
Baseline	random	0.47	0.32	0.76	0.61	0.52
	OoV	0.55	0.32	0.77	0.62	0.55
	annotator	0.50	0.31	0.77	0.62	0.53
Gen	random	0.53	**0.40**	0.77	0.62	0.55
	OoV	0.49	0.36	0.75	0.61	0.53
	annotator	0.47	0.35	0.75	0.60	0.51
Disc	random	**0.58**	0.39	**0.79**	**0.65**	**0.59**
	OoV	0.55	0.36	0.77	0.63	0.56
	annotator	0.52	0.33	0.77	0.62	0.54

Table 6: Three-way classification on random, Out-of-Vocabulary (OoV), and annotator splits. We report class-wise F_1 for *Strongly hateful* (Sh), *Weakly hateful* (Wh), and *Not hateful* (Nh), along with overall accuracy (Acc) and macro-F_1. Bold indicates the best value per column.

model	split by	Sh	Wh	N	Nh	Acc	macro-F_1
Baseline	random	0.40	0.31	0.00	0.72	0.55	0.36
	OoV	0.33	0.28	0.00	0.70	0.54	0.33
	annotator	0.49	0.31	0.00	0.74	0.59	0.39
Gen	random	0.45	0.41	**0.01**	0.74	0.59	0.40
	OoV	0.43	0.42	**0.01**	0.73	0.58	0.40
	annotator	0.37	0.38	0.00	0.72	0.56	0.37
Disc	random	**0.56**	0.39	0.00	**0.77**	**0.62**	**0.43**
	OoV	0.55	0.36	0.00	0.76	0.61	0.42
	annotator	0.53	0.31	0.00	0.76	0.60	0.40

Table 7: Four-way classification on random, Out-of-Vocabulary (OoV), and annotator splits. We report class-wise F_1 for *Strongly hateful* (Sh), *Weakly hateful* (Wh), *Cannot decide* (N), and *Not hateful* (Nh), along with overall accuracy (Acc) and macro-F_1. This setting preserves the original label space, thereby increasing class imbalance and treating *Cannot decide* as real-world noise. Bold indicates the best value per column.

Component	#Params	Share of total
BERT backbone (`BertModel`)	109M	$\sim 94\%$
Classifier MLP	82.1K	$< 0.1\%$
DEFFUSION cross-attn (2x)	$\sim$ 4.8M	$\sim 4.1\%$
ANNOGATE attention	$\sim$ 2.4M	$\sim 2.1\%$
LayerNorm + embeddings (aux)	$\ll$ 0.5M	$< 0.5\%$
Total (AWARE-HATE)	**116M**	**100%**

Table 8: Parameter breakdown for AWARE-HATE. The additional DEFFUSION and ANNOGATE modules contribute only a small fraction of total parameters relative to the frozen BERT backbone.

imately 2.4M parameters, resulting in 4.8M from DEFFUSION and 2.4M from ANNOGATE. Collectively, these components account for 6–7% of total parameters, while the backbone remains responsible for over 90% of both parameters and computation.

Training time. Under our 10-fold cross-validation setup, the first-stage supervised fine-tuning of AWARE-HATEamounted to approximately 2 hours and 40 minutes in total across all folds. The subsequent second-stage training based on GRPO required an additional 1 hour and 20 minutes. As a result, the reinforcement learning stage introduces roughly 1.3 hours of wall-clock time, accounting for about 50% of the first-stage training cost and approximately one-third of the total training time. Given that the BERT encoder remains the primary computational bottleneck, we consider this overhead to be acceptable in light of the substantial improvements observed on ambiguous and highly subjective instances.

A Prompt for LLM Baselines

A.1 DPO-based Preference Optimization

For the DPO-based preference-optimized LLM baselines (Section 4.5), we condition the model on the context sentence, target term, its definition, and the current annotator profile. The full prompt template is given below.

Instruction:
Given the following sentence that mentions a particular term, and the definition of the term, classify whether the current

annotator considers the meaning of that term expresses hate toward a person or group within that specific sentence by giving one of the following corresponding labels:
"hateful"
"not hateful"

Input:
Sentence: "[*SENTENCE*]"
Term: "[*TERM*]"
Definition: "[*DEFINITION*]"
Annotator: "[*ANNOTATOR*]"

Response:

Except for this prompt design, all other training details for the preference-optimized LLM baselines follow prior DPO-based configurations.

A.2 In-Context Few-Shot Classification

For the instruction-tuned LLM baselines (Qwen2.5 and DeepSeek) evaluated in-context without fine-tuning, we use the following 4-shot prompt. The model is required to output strictly one of {*hate*, *not*}.

You are a strict classifier. Output ONLY one of: hate or not.

Example 1:
Context: The bugger's given me the wrong change.
Target: bugger
Definition: A foolish or worthless person or thing; a despicable person.
Annotator: Reader is 30, Female and White
Label: hate

Example 2:
Context: He's a silly bugger for losing his keys.
Target: bugger
Definition: A foolish or worthless person or thing; a despicable person.
Annotator: Reader is 25, Female and Black
Label: not

Example 3:
Context: They look at those massive tummies and ample asses on their fellow big men with a leering, lusty eye, the same way your average club kid might go all aflutter when he spots an Aberzombie in the house.
Target: Aberzombie
Definition: An unthinking conformist who wears fashions from or in the style of the brand Abercrombie & Fitch.
Annotator: Reader is 35, Female and White
Label: not

Example 4:
Context: This is no huge overgrown abbey-lubber.
Target: abbey-lubber
Definition: An able-bodied idler who grew sleek and fat from the charity of religious houses; a lazy monk.
Annotator: Reader is 22, Male and White
Label: hate

Now, classify the following sample strictly by returning one of ['hate', 'not']:
Context: "[*SENTENCE*]"
Target: "[*TERM*]"
Definition: "[*DEFINITION*]"
Annotator: "[*ANNOTATOR*]"
Label: