Treating Annotator Disagreement as a Controllable Latent Feature via Routed QLoRA Adapters

Jörn Boehnke Jacob Brophy Ashwin A Nair

Graduate School of Management, University of California, Davis

Abstract

Human annotators routinely disagree on hate speech labels, especially when interpreting sarcasm, coded language, or platform-specific norms. This disagreement reflects genuine differences in community standards, not annotation error. Most detection systems discard it. Conventional models often treat this inter-annotator disagreement as label noise. We propose a Mixture-of-Experts architecture in which three QLoRA adapters, trained on microblogging (HateXplain), collaborative editing (Jigsaw), and extremist forum (CrimeBB) corpora, are blended via a learned gating mechanism atop a Llama 2 7B backbone. A control parameter adjusts detection sensitivity at inference time, enabling more conservative or more permissive predictions without retraining. Each prediction includes a structured rationale identifying targeted groups, toxic spans, and harm mechanisms. Training requires approximately 10 GPU-hours per adapter on a single NVIDIA RTX 4090 with 4-bit quantization. Across three benchmarks, the default Logit Blending configuration maintains Macro-F1 within 5 percentage points of a monolithic QLoRA baseline trained on pooled data, and discrete top-3 routing matches or exceeds it in accuracy on the extremist-forum and collaborative-editing benchmarks. Sensitivity tuning yields per-group accuracy gains of up to 6 percentage points, and blinded GPT-4 evaluation ranks blended rationales above the monolithic baseline in logical consistency, completeness, and fluency. The result is a framework for hate speech detection that yields a transparent and computationally efficient alternative to monolithic classification.

Keywords: Artificial Intelligence, Hate Speech, Large Language Models, Llama, LoRA, Mixture of Experts, QLoRA

1 Introduction

The assessment of subjective content, particularly hate speech, is inherently challenging due to the sociolinguistic complexity of online discourse. Human annotators frequently disagree on labels, especially when interpreting sarcasm, analogy, coded slurs, or emoji cues, even when provided with detailed guidelines [1].

The practical cost is immediate. A classifier tuned for Twitter will over-flag on Wikipedia; one calibrated for Wikipedia will miss coded slurs on extremist forums. Retraining per deployment context is expensive, and maintaining separate models is unwieldy. Practitioners need a single system whose sensitivity can be adjusted at inference time, without retraining.

Treating annotator disagreement as label noise yields brittle systems. Monolithic models optimized for a single platform or annotator pool often transfer poorly across domains. Typical classifiers also return a single decision scalar with limited transparency regarding the factors that drove the prediction.

This paper presents such a system. We train three QLoRA adapters on corpora representing distinct moderation cultures – microblogging (HateXplain [2]), collaborative editing (Jigsaw [3]), and extremist forums (CrimeBB [4]) – and blend them via a learned gating mechanism [5] atop a Llama 2 7B backbone [6]. In Dynamic Routing mode, the gate weights adapters based on input semantics. In Policy Control mode, operators supply fixed blending weights, locking the model to a chosen sensitivity profile. Both modes generate structured rationales identifying targeted groups, toxic spans, and harm mechanisms in a single response.

We make three primary contributions:

1. A parameter-efficient Mixture-of-Experts framework that routes inputs to domain-specific QLoRA adapters, handling conflicting community norms without full fine-tuning.

2. An inference-time control parameter that adjusts detection sensitivity, including for specific demographic targets, without retraining or maintaining separate per-context models.

3. A structured natural-language rationale accompanying every prediction, identifying targeted groups, toxic spans, and the harm mechanism, supervised by GPT-4-generated chain-of-thought rationales.

1.1 Abbreviations and Acronyms

Table 1 details the primary acronyms and abbreviations utilized throughout this manuscript.

Table 1: List of Acronyms and Abbreviations

Acronym	Definition
LLM	Large Language Model
MoE	Mixture of Experts
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
QLoRA	Quantized Low-Rank Adaptation
NLP	Natural Language Processing
CoT	Chain-of-Thought
MLP	Multi-Layer Perceptron
ReLU	Rectified Linear Unit
GPU	Graphics Processing Unit

2 Related Work

This section reviews prior literature on (i) subjectivity and pragmatics in toxic language; (ii) explainable hate-speech detection; (iii) Mixture-of-Experts (MoE) architectures; and (iv) parameter-efficient fine-tuning (PEFT).

2.1 Subjectivity and Pragmatics

Hate-speech annotation exhibits substantial inter-annotator variability driven by pragmatics (sarcasm, analogy, implicature), annotator stance, and community norms. Treating this disagreement as stochastic noise often obscures systematic variation regarding what constitutes offensive content [1, 7–11]. This necessitates modeling rater heterogeneity rather than optimizing for a single consensus label. This perspectivist stance, in which annotator disagreement is modeled rather than averaged away, has become an active research line, including soft-label training that preserves annotation uncertainty and disagreement-aware classifiers for subjective tasks. Our work operationalizes this perspective by routing inputs among community-specialized adapters, thereby explicitly modeling the divergent judgments inherent in subjective tasks.

2.2 Explainable Toxic Speech Detection

Extensive research has established frameworks for hate speech detection using lexicon-based approaches [12, 13] and supervised learning on benchmark datasets [9, 14–20]. However, the majority of these studies deploy "black-box" classifiers that lack transparency regarding the decision boundary.

To address this opacity, explainable hate speech detection has emerged as a critical research area. Keyword-based explanation methods [2, 21–23] identify explicitly toxic tokens but often fail to capture implicit hate speech where no single word is profane. Subsequent approaches have utilized autoregressive language models trained on human-annotated rationales [21, 24, 25]; however, these models often suffer from hallucinations or inconsistent reasoning [26, 27].

Yang et al. [28] advanced this paradigm by distilling reasoning capabilities from larger models (e.g., GPT-3.5). We extend this line of inquiry by incorporating architectural modularity: our framework routes inputs to specific "expert" adapters based on platform or rater characteristics, ensuring consistency across heterogeneous data sources.

2.3 Mixture of Experts (MoE)

The Mixture of Experts (MoE) paradigm enables model scaling while maintaining constant inference costs by conditionally activating a subset of parameters [5, 29–31]. Unlike dense layers, an MoE layer comprises N independent feed-forward networks (experts) $\{E_i\}_{i=1}^{N}$ and a gating function $G(\cdot)$. The gate models a probability distribution over experts to determine activation for a given input representation x. The gating coefficient for expert E_i is typically computed via a softmax over the dot product of x and a trainable router matrix W_g:

$$G(x)_i = \frac{\exp(x \cdot W_{g,i})}{\sum_{j=1}^{N} \exp(x \cdot W_{g,j})} \tag{1}$$

The final output y is the weighted sum of the activated experts: $y = \sum_{i=1}^{N} G(x)_i E_i(x)$. We adapt this mechanism to weight parameter-efficient adapters rather than full feed-forward blocks.

2.4 Parameter-Efficient Fine-Tuning (QLoRA)

Low-Rank Adaptation (LoRA) [32] adapts Large Language Models (LLMs) to downstream tasks by freezing the pre-trained weights W_0 and optimizing low-rank decomposition matrices A and B. The update is formalized as $W = W_0 + \Delta W = W_0 + BA$, where the rank $r \ll \min(d_{in}, d_{out})$. Quantized LoRA (QLoRA) [33] further reduces memory overhead by backpropagating gradients through a 4-bit quantized version of W_0. This enables high-fidelity fine-tuning on consumer hardware, a constraint we adhere to in our system design.

2.5 Adapter Composition

Recent work has explored the arithmetic composition of LoRA adapters to facilitate multi-task generalization [34–36]. A parallel line treats multiple LoRA adapters as experts blended by a learned router, the Mixture-of-LoRA-Experts paradigm [37, 38], including parameter-efficient expert mixtures and soft expert merging [39, 40]. Our work differs in two respects. First, our router conditions on a frozen sentence-level embedding of the whole input rather than on token-level hidden states, so routing reflects the input's community context rather than per-token capacity needs. Second, we apply the mechanism to intra-task disagreement across moderation regimes, not to cross-task or cross-domain capacity scaling. In these frameworks, effective model weights $\hat{W}$ are derived via a linear combination of task-specific weight deltas:

$$\hat{W} = W_{base} + \sum_{i=1}^{N} w_i \cdot \Delta W_i \tag{2}$$

Our approach applies this composition strategy to *intra-task* variation: we blend adapters trained on distinct communities to model pragmatic disagreement, rather than to maximize generalization across disparate tasks.

To our knowledge, this is the first work to use a Mixture-of-LoRA-Experts router specifically to model inter-annotator disagreement as a controllable signal. Prior perspectivist and learning-with-disagreement work treats disagreement at the level of labels or annotators; prior Mixture-of-LoRA-Experts work targets capacity and multi-task transfer. We sit at their intersection: a routing mechanism whose blending weights are the operator's sensitivity control across moderation regimes.

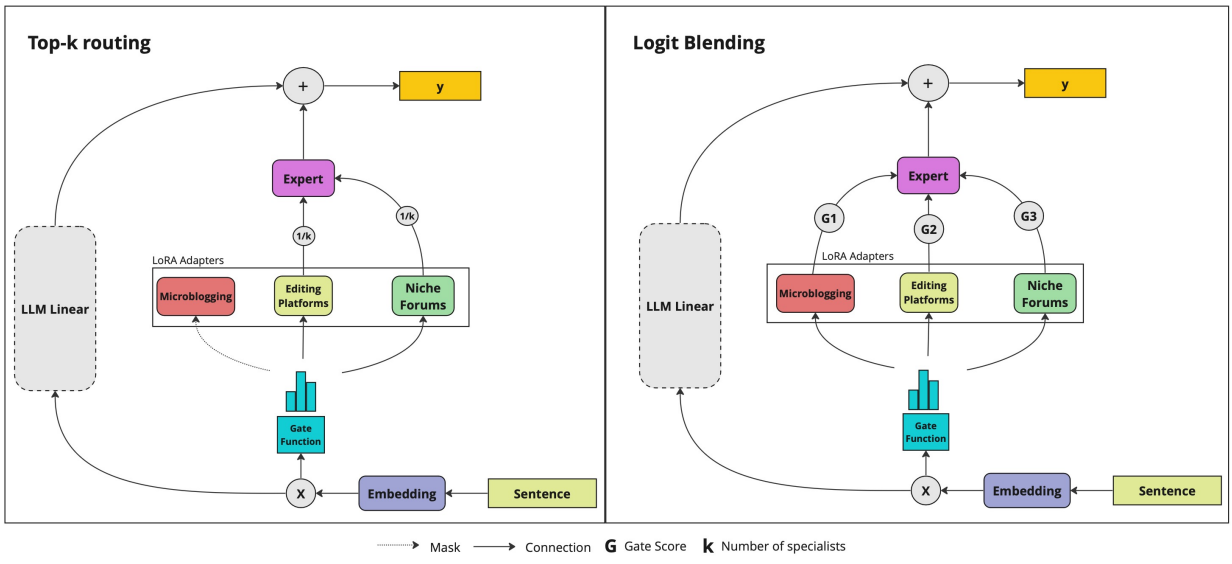

Figure 1: A learned router blends three community-specialized QLoRA adapters over a frozen Llama 2 7B backbone. Architecture of our hate speech detection framework. An input sequence is vectorized using an external sentence embedding model, and then passed to a gate function. In the case of top-k routing (top-2 or top-3), we form a weighted linear combination of the k adapters with the highest gate score with weight equal to $\frac{1}{k}$. In the case of Logit Blending, the raw softmax output scores are used to form a linear combination of all adapters.

3 Methodology

3.1 Data Preparation and Augmentation

Our approach leverages Quantized Low-Rank Adaptation (QLoRA) [33] to fine-tune foundational model weights via selective backpropagation into low-rank adapters. This strategy approximates the performance of full 16-bit fine-tuning while reducing memory requirements by approximately 10×.

Because annotator-level metadata is rarely available in large hate-speech datasets, we approximate heterogeneous annotation regimes using distinct datasets. Each corpus was constructed with different annotator pools, instructions, and moderation guidelines, which produce systematically different labeling behavior.

We therefore treat each dataset as an empirical realization of a particular moderation regime. Expert adapters are trained on each regime, and routing allows the model to interpolate between these regimes at inference time.

3.1.1 Data Sources

We curate a heterogeneous dataset spanning three distinct sociolinguistic domains. Each adapter is trained independently on one of the following corpora:

- **Microblogging:** A combination of the HateXplain benchmark [2] and the automated hate speech dataset collected by Davidson et al. [15], representing short-form text from Twitter and Gab. The Davidson et al. labels (hate, offensive, neither) map directly onto our taxonomy (Hateful, Offensive, Neutral) and are pooled with HateXplain only for adapter training; all microblogging evaluation, including the label distribution in Table 2 and the per-group analyses, uses the held-out HateXplain test split alone.
- **Collaborative Editing:** Wikipedia talk page comments labeled by Jigsaw/Conversation AI [3], representing moderation norms in public collaboration platforms.
- **Extremist Forums (CrimeBB):** Long-form posts sourced from the CrimeBB dataset [4], covering niche communities such as Stormfront and Inceldom. We refer to this domain as Extremist Forums throughout, including in all tables and figures.

3.1.2 Label Normalization and Rationale Generation

All dataset labels were mapped to a unified taxonomy: `{Neutral, Offensive, Hateful}`. To supervise the model's explainability capabilities, we augmented a subset of the data with synthetic rationales generated by GPT-4 [41].

We employed Chain-of-Thought (CoT) prompting [42] to enforce a structured output schema, ensuring consistent supervision signals for the adapters. The target output structure includes:

- **Holistic Rationale:** A concise justification (max. 150 words) detailing the reasoning behind the classification.
- **Classification:** The mapped ground-truth label.
- **Targeted Groups:** An enumeration of demographic groups referenced in the text.
- **Toxic Spans:** Extraction of specific lexical triggers (harmful words/phrases).
- **Harm Mechanism:** A brief explanation of *why* the extracted spans are harmful to the targeted group.

3.2 Adapter Routing and Composition

To model the heterogeneous nature of hate speech, we employ a routing mechanism that dynamically blends the weights of the N=3 community-specific QLoRA experts $\{E_i\}_{i=1}^{N}$.

3.2.1 Gating Mechanism

Unlike standard MoE layers that route based on internal hidden states, our router $R(\cdot)$ conditions on the semantic context of the entire input sequence. We utilize a frozen sentence-embedding encoder [43] to generate an input representation $u \in \mathbb{R}^{d_{emb}}$. Concretely, where the generic MoE gate above takes the hidden state x as input, our router conditions on u, so the routing decision reflects the whole input's community context rather than per-token features.

The router consists of a two-layer Multi-Layer Perceptron (MLP) with a hidden dimension of 8 and a Rectified Linear Unit (ReLU) activation. The output is normalized via a softmax function to produce a blending coefficient vector $s \in \Delta^{N-1}$:

$$h = \sigma(W_1 u + b_1) \tag{3}$$

$$s = \text{softmax}(W_2 h + b_2) \tag{4}$$

where $W_1 \in \mathbb{R}^{8 \times d_{emb}}$, $W_2 \in \mathbb{R}^{N \times 8}$, and σ denotes the ReLU activation.

The final effective adapter weight ΔW_{mix} applied to the base model is a weighted summation of the expert adapters:

$$\Delta W_{mix} = \sum_{i=1}^{N} s_i \cdot \Delta W_i \tag{5}$$

3.2.2 Inference Strategies

Our framework supports three distinct inference modes to balance automation and operator control:

1. **Logit Blending (Default):** The model utilizes the raw probability vector s from Eq. 4 to perform a weighted average of all experts, capturing nuanced disagreement where multiple community norms may apply simultaneously.

2. **Dynamic Sparse Routing (Top-k):** To reduce computational overhead, we activate only the top-k experts (where $k \in \{2, 3\}$). The weights are renormalized such that $s_i = 1/k$ for the selected experts and 0 otherwise. Using uniform rather than softmax-renormalized weights makes the top-k conditions differ from Logit Blending only in the number of active experts, so the sensitivity analysis in Section 5 isolates the effect of expert count from the gate's confidence values.

3. **Policy Control Mode:** To enforce specific moderation standards (e.g., "Strict Wikipedia Compliance"), operators may bypass the router and manually supply a fixed weight vector s^*, effectively locking the model into a specific sensitivity profile.

3.3 Implementation Details

We utilized the Llama-2 7B backbone [6]. All adapters were trained on a single NVIDIA RTX 4090 GPU (24GB VRAM). Training required approximately 10 GPU-hours per adapter.

3.3.1 Hyperparameters

We employed a standard QLoRA configuration with the following specifications:

- **Rank (r):** 64
- **Scaling (α):** 16
- **Dropout:** 0.1
- **Optimizer:** AdamW with a cosine learning rate scheduler (initial LR $= 2 \times 10^{-4}$).
- **Target Modules:** Adapters were injected into the query (W_q), key (W_k), value (W_v), and output (W_o) projection layers of the self-attention mechanism, as well as the gate, up, and down projections of the MLP blocks.

The framework structure is visualized in Fig. 1.

4 Experimental Setup

4.1 Evaluation Metrics

We assess framework performance across three dimensions to balance classification accuracy with interpretability:

1. **Classification Performance:** We report Accuracy and Macro-F1 scores. Macro-F1 is prioritized to account for the class imbalances detailed in Table 2.
2. **Target Identification:** We measure the model's ability to correctly identify targeted demographic groups (e.g., "African," "Jewish") using exact-match accuracy on the HateXplain test set.
3. **Rationale Quality (LLM-as-a-Judge):** To evaluate the semantic coherence and persuasive quality of generated explanations, we employ GPT-4 as an independent evaluator following the protocol of [44]. Crucially, the evaluator is **blinded to the ground-truth labels** to prevent bias. The judge ranks the rationales generated by the competing models based solely on their logical consistency with the input text, completeness of the argument, and linguistic fluency.

4.2 Datasets and Label Harmonization

We evaluate on the held-out test splits of the three corpora described in Section 3. To enable unified evaluation, source-specific labels were mapped to a common taxonomy $\mathcal{L} = \{\text{Neutral, Offensive, Hateful}\}$:

- **Jigsaw (Collaborative Editing):** Labels tagged *toxic, insult, profanity* were mapped to **Offensive**; *identity hate* was mapped to **Hateful**.
- **CrimeBB (Extremist Forums):** Labels tagged *toxic, severe toxic, threat* were mapped to **Offensive**; *identity attack* was mapped to **Hateful**.

Table 2 summarizes the label distribution for each evaluation set. Figure 2 details the distribution of targeted groups within the HateXplain subset.

Table 2: Distribution of labels for each evaluation dataset. Microblogging uses the held-out test set from [2]; collaborative editing uses the Jigsaw Toxic Comment Classification Challenge [3]; and extremist forums draw on CrimeBB forum data labeled by [4].

Platform	Neutral	Offensive	Hateful
Extremist Forums	1,443 (75%)	329 (17%)	152 (8%)
Collab. Edit	1,742 (91%)	157 (8%)	16 (1%)
Microblog	782 (41%)	548 (28%)	594 (31%)

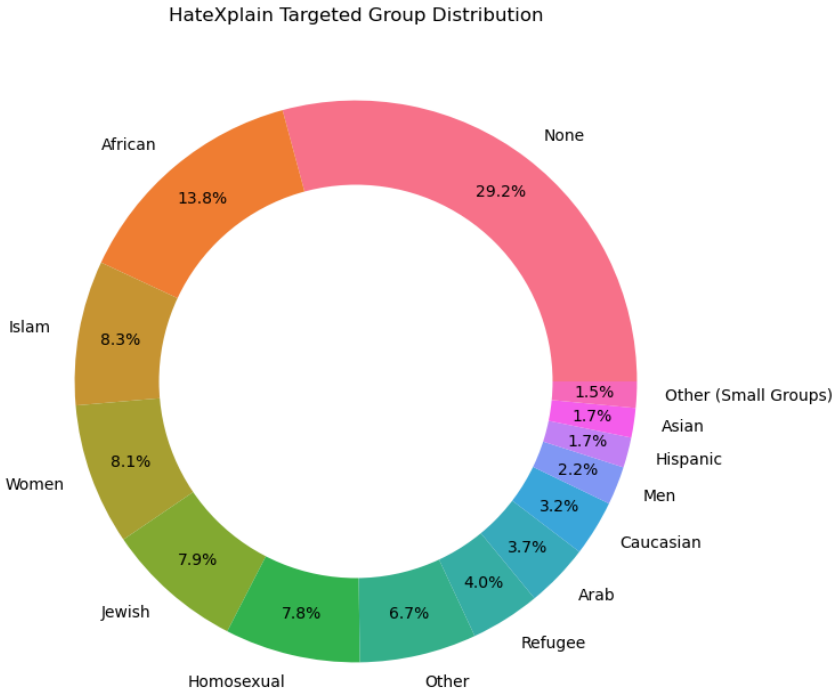

Figure 2: Five groups – African, Islam, Women, Jewish, and Homosexual – account for roughly two thirds of all targeted-group annotations. Targeted class distribution within the HateXplain evaluation dataset. The data labeling process of HateXplain involved 3 annotators for each post, as a result this distribution reflects the opinions of each annotator rather than only the most commonly listed targeted category among all 3 annotators. Note that the 'Other' category is a valid label given as an option to annotators, while the 'Other (Small Groups)' category is an aggregation of multiple sub-groups that were targeted in less than 1% of the posts.

4.3 Baselines and Comparisons

We compare our proposed **Logit Blending** and **Discrete Top-k** routing ($k = 2, 3$) against three control conditions:

- **BERT-Large (Fine-Tuned):** A discriminative baseline [45] trained on the pooled dataset. This serves as a non-generative benchmark for classification accuracy.
- **Monolithic QLoRA:** A single Llama-2 7B adapter trained on the pooled concatenation of all three training corpora. This baseline isolates the impact of the MoE architecture by removing the routing mechanism.
- **Reference Domain Routing (k=1):** A control condition where test inputs are manually routed to their corresponding domain adapter (e.g., Wikipedia comments are sent directly to the Collaborative Editing adapter). This quantifies the performance gap between our automated router and perfect domain knowledge. Throughout the results, k=1 denotes this reference condition, a single, manually domain-matched adapter, whereas Discrete Top-k routing with $k \in \{2, 3\}$ denotes the learned router restricted to its k highest-scoring experts; the two mechanisms are distinct and share only the notation k.

5 Results

We organize results around five claims. Each subsection states the claim, then presents the evidence.

5.1 Classification Performance

Splitting a single adapter into three routed experts does not degrade aggregate performance and, on two of three benchmarks, improves it.

Table 3 summarizes performance across the three evaluation domains. On Extremist Forums, Logit Blending achieves the highest accuracy (97.5%) and Macro-F1 (0.93), surpassing both the Monolithic QLoRA (96.5%, 0.90) and BERT (97.0%, 0.93). On Collaborative Editing, it matches BERT (94.1% vs. 94.0%) and narrows the F1 gap with the Monolithic baseline (0.70 vs. 0.75). HateXplain is the most challenging benchmark, with a roughly balanced three-class split (Table 2). Here, Logit Blending scores 63.2% accuracy and 0.61 F1 — modestly below the Monolithic baseline (66.0%, 0.62) and BERT (66.0%, 0.64). The k=3 configuration exceeds the Monolithic baseline on accuracy for both Forums and Editing, and on Macro-F1 for Forums (0.92 vs. 0.90); on Editing its Macro-F1 (0.64) trails the pooled baseline (0.75), so the gain there is in accuracy rather than balanced F1. Overall, domain-specialized routing mitigates the negative transfer inherent in training a single model on corpora with conflicting norms.

5.2 Sensitivity Scales Monotonically with Blending Density

Operators can shift the model's decision boundary in a predictable direction by adjusting a single parameter.

Table 4 and Fig. 3 quantify this effect. As active experts increase from k=1 to k=3, Hate predictions rise 47.3% relative to the ground-truth distribution while Neutral predictions fall 36.4%. Logit Blending occupies the middle ground: Hate +17.5%, Neutral +2.5%.

The monotonic relationship maps directly to deployment policy. At k=1, the model aligns closely with the Neutral distribution (+13.0%), reducing false positives for platforms prioritizing free

Table 3: Model accuracy for hate speech detection across varied datasets, representing microblogging platforms (HateXplain), collaborative editing platforms (Wikipedia), and extremist forums (CrimeBB). This table compares Monolithic QLoRA performance with various routing configurations.

Method	Extremist Forums		Collaborative Editing		HateXplain	
	Acc.↑	Macro F1↑	Acc.↑	Macro F1↑	Acc.↑	Macro F1↑
Monolithic QLoRA	96.5%	0.90	93.8%	0.75	**66.0%**	0.62
Discrete k=1	96.7%	0.91	92.3%	0.61	65.0%	0.63
Discrete k=2	93.5%	0.84	92.5%	0.66	64.1%	0.62
Discrete k=3	97.2%	0.92	94.1%	0.64	60.7%	0.59
Logit Blending	**97.5%**	**0.93**	**94.1%**	**0.70**	63.2%	0.61
BERT Baseline	97.0%	0.93	94.0%	0.70	66.0%	**0.64**

expression. At k=3, it maximizes Hate detection (+47.3%), suitable for child-safe or high-moderation environments. Logit Blending provides a balanced default with minimal Neutral distortion.

Qualitative analysis supports this. For pragmatically ambiguous statements like *"<user> i f*cking hate you yzm [face with tears of joy]"*, the Monolithic QLoRA predicts **Neutral** (focusing on the emoji), while the Logit Blending and Top-2 configurations predict **Offensive** (focusing on the text). This demonstrates that blending captures the dual interpretations inherent in the data.

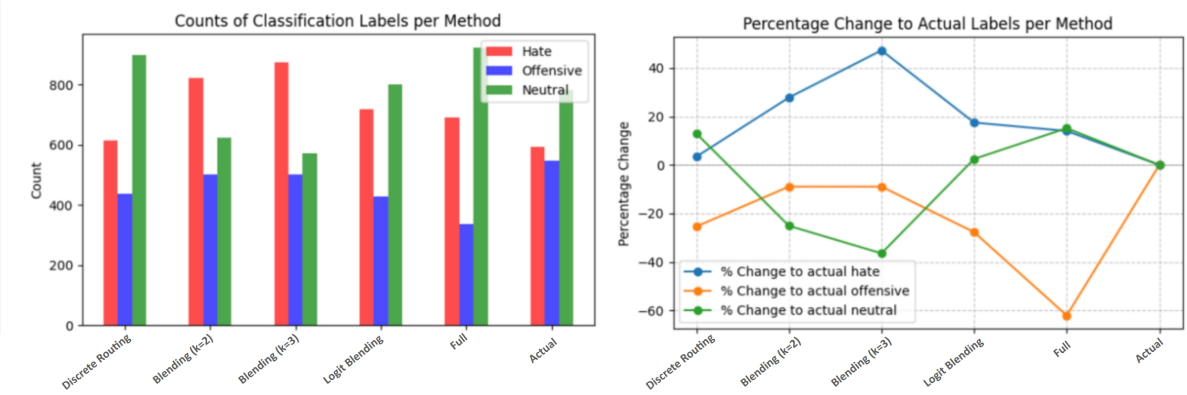

Figure 3: Detection sensitivity scales monotonically with blending density. Predicted label counts (Hate, Offensive, Neutral; bars) and percentage changes relative to the empirical HateXplain label distribution (lines), for discrete routing ($k = 1, 2, 3$) and Logit Blending. As active experts increase, Hate predictions rise and Neutral predictions fall.

5.3 Group-Specific Sensitivity

Different routing configurations benefit different demographic groups, with per-group accuracy gains of up to 6 percentage points over the Monolithic baseline.

Fig. 4 reports accuracy across the five most targeted groups in HateXplain. Logit Blending achieves 92.8% concordance with human annotators on Antisemitic content and 81.7% on LGBTQ+ content. For anti-Black hate speech, k=3 reaches 88.8%, outperforming the Monolithic baseline by 2.6 points. No single configuration dominates all demographics — precisely the outcome the framework is designed to address.

These results indicate that different routing strategies benefit different demographic groups, reinforcing the need for tunable sensitivity rather than a single global decision threshold.

Table 4: Sensitivity analysis of hate speech classification using various QLoRA adapter configurations. The table reports predicted label counts (Hate, Offensive, Neutral) and percentage changes relative to the empirical label distribution in the HateXplain test dataset for discrete routing (k=1, k=2, k=3), Logit Blending, and a monolithic QLoRA baseline.

Method	Hate	Off.	Neutral	% Change vs. Actual		
				Hate	Off.	Neutral
Monolithic QLoRA	691	338	922	+14.03	-62.13	+15.18
Discrete k=1	615	437	899	+3.41	-25.40	+13.01
Discrete k=2	823	503	625	+27.82	-8.94	-25.12
Discrete k=3	875	503	573	+47.30	-8.94	-36.47
Logit Blending	720	429	802	+17.50	-27.73	+2.49
Actual	594	548	782	0	0	0

Fig. 5 confirms that sensitivity adjustments do not erode target identification: in all configurations, the model and human annotators agree on which groups are targeted in the dominant quadrant (χ^2 test with Bonferroni correction across five groups, corrected $\alpha = 0.01$).

Agreement analysis (Tables 5 and 6) reveals a consistent pattern across methods. Complete agreement with human annotators is highest for Hateful content (60.0%) and lowest for Offensive content (13.9%). Partial disagreements cluster around the Offensive category (43.0%), reflecting its inherent ambiguity as the boundary between Neutral and Hateful. This asymmetry is stable across all routing configurations, indicating that the disagreement stems from the task itself, not from the architecture.

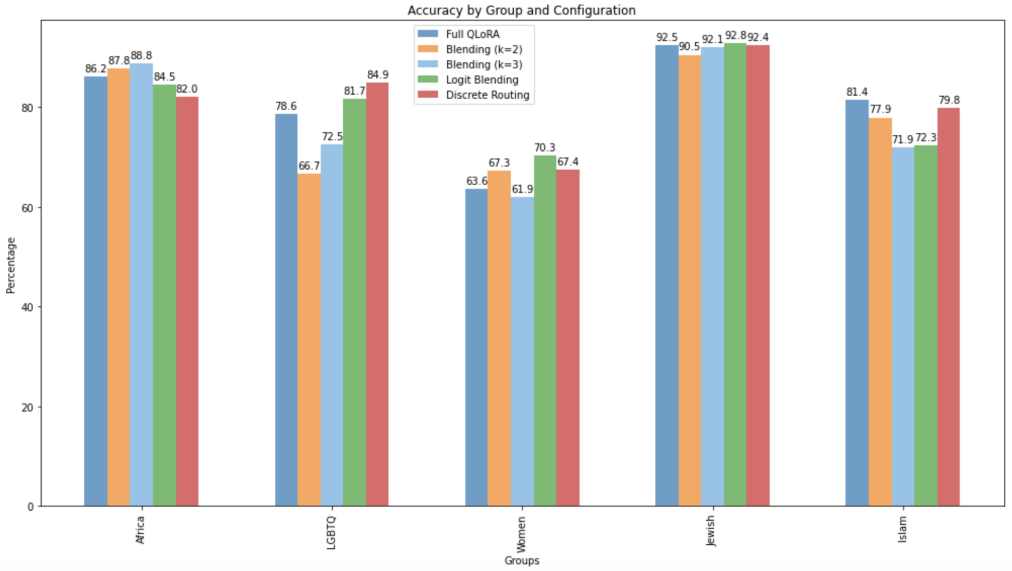

Figure 4: No single routing configuration dominates across demographic groups. Accuracy across the five most commonly targeted demographic groups in the HateXplain evaluation set, measured on instances where model predictions align with human annotations. Results are reported for Logit Blending and discrete routing configurations.

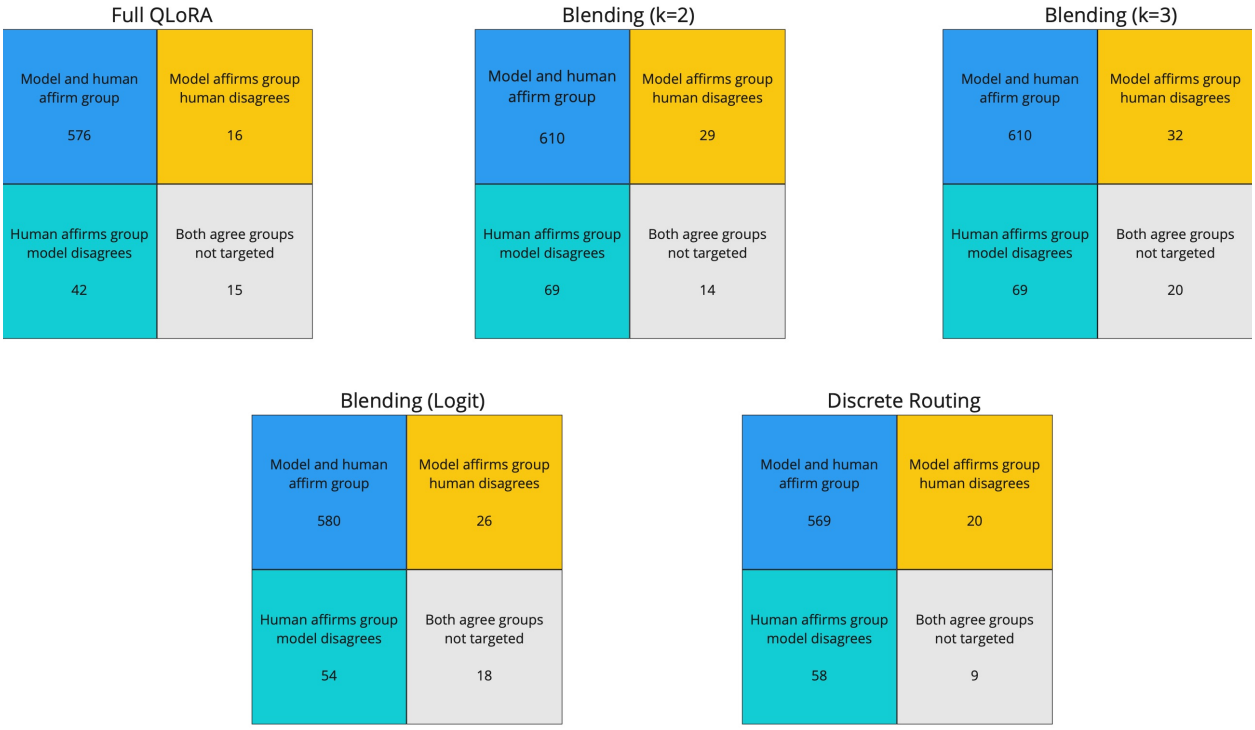

Figure 5: Sensitivity adjustments do not erode target identification. Concordance between human and model annotations for the five most targeted groups in HateXplain. Each 2×2 matrix reports counts of agreement and disagreement on whether a top-5 group was targeted. All configurations show significant concordance with human annotators (χ^2 test with Bonferroni correction across five groups, corrected $\alpha = 0.01$)

Table 5: This table shows the breakdown of the number of posts every blending method agreed on (and also aligned with the human annotators). Label represents which category was assigned to the post (neutral, offensive, hate), Number of Agreements represents the raw number of times the models agreed with each other and the human annotation, and % of total displays the percentage of values in every category that the models agreed on. Agreements occur at a much higher rate for hate speech than for either offensive or neutral statements, with the lowest agreement by far occurring for offensive statements.

Complete Agreement with Humans		
Label	**Num. of Agreements**	**% of Total**
Neutral	371	45.74%
Offensive	76	13.89%
Hateful	356	60.03%

5.4 Blended Adapters Produce More Coherent Rationales

Routing improves not only what the model predicts, but also the explanations it provides for its decisions.

We evaluate rationale quality using GPT-4 as a blinded judge following the protocol of [44], ranking each method on logical consistency, completeness, and fluency (Fig. 6). Because GPT-4 also generated the training rationales, the judge shares a prior with the supervision signal; we treat the ranking as indicative rather than definitive and revisit this circularity in the Limitations section.

Table 6: This table presents the number of instances where at least two methods disagree, but at least one matches the human annotators' response. We see that method disagreements occur at a higher rate for neutral and offensive statements (32.05% and 42.96% respectively) than for hateful statements (only 20.40% disagreement). As an inverse to Table 5, we also see that offensive statements had the highest rate of disagreement.

Partial Agreement with Humans		
Label	**Num. of Agreements**	**% of Total**
Neutral	260	32.05%
Offensive	235	42.96%
Hateful	121	20.40%

Blending with k=2 ranks first most frequently, with k=3 close behind. The Monolithic QLoRA, despite competitive classification accuracy on some splits, consistently ranks last in explanation quality.

This pattern reveals a "right answer, wrong reason" phenomenon. The monolithic model achieves label accuracy partly through spurious correlations, producing rationales that the blinded GPT-4 judge consistently ranks as less coherent and less complete. The k=2 configuration also shows the greatest rank stability, rarely falling below second place — suggesting that two community-specific perspectives per prediction strike the best balance between reasoning diversity and noise.

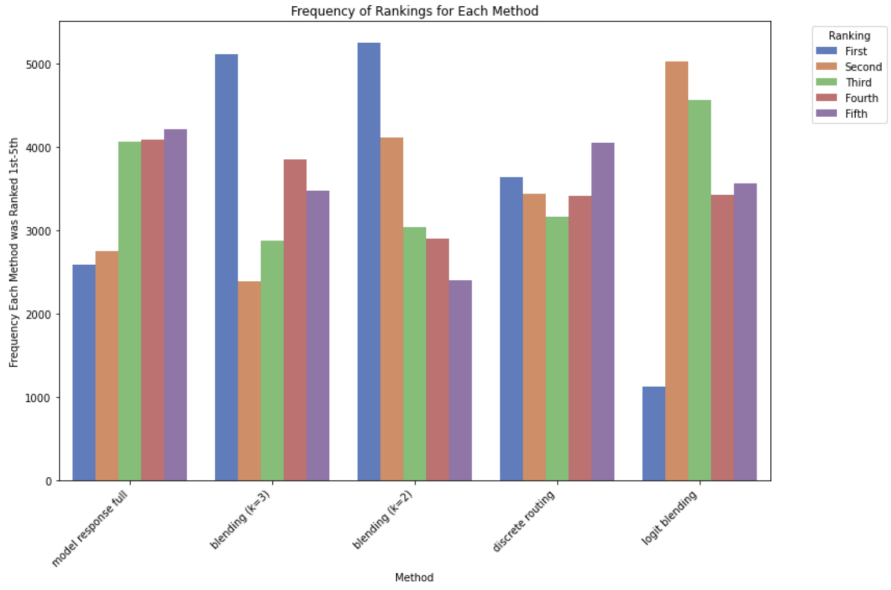

Figure 6: Top-k blending is ranked first most often, while the monolithic baseline is ranked last most often. Comparison of GPT-4 rankings for each method on the HateXplain evaluation set. The y-axis reports the number of times a method's classification and explanation was ranked 1st–5th by GPT-4 across 11 independent rankings per post. Rankings in which all methods were scored equally are excluded.

5.5 Annotator Disagreement Validates the Design

Ground truth is often not fixed in hate speech detection. The same texts, relabeled by a different annotator, produce a different set of correct answers.

Two independent annotators relabeled 1,000 HateXplain test samples (Table 7). These accuracies are computed against the re-annotated labels, not the original HateXplain labels, so their levels are not directly comparable to the accuracies in Table 3; what matters here is the rank order across configurations, not the absolute values. Annotator 1's labels favor k=3 blending (63.77% accuracy); Annotator 2's labels favor Discrete Routing at k=1 (57.51%). No configuration satisfies both.

This is the empirical case for the framework's existence. A task where the ground truth itself shifts with the evaluator demands a system whose sensitivity is adjustable — not as a convenience, but as a structural response to a structural property of the problem.

Table 7: Breakdown of accuracy across alternative human-assigned labels. Annotators were given the task of assigning the labels hate, offensive, or neutral for 1000 observations in the HateXplain test dataset. We see a large difference from the initial base annotations, demonstrating why pure accuracy may not be a good measure of framework performance given human bias for annotator labels.

Annotator	**Monolithic QLoRA**	k**=1**	k**=2**	k**=3**	**Logit Blending**
1	53.27	50.15	61.55	63.77	55.80
2	53.88	57.51	54.28	53.58	54.28

6 Conclusion

Annotator disagreement in hate speech detection is not noise; it reflects how different communities draw the line between acceptable and harmful speech. Domain-specific QLoRA adapters capture these differences through a lightweight routing mechanism that recombines them at inference time.

Each of three adapters learns the moderation norms of its training corpus. A learned router blends their outputs based on input semantics, or operators bypass the router and impose fixed weights. Routing with k=2 produces the most coherent rationales as judged by blinded GPT-4 evaluation, while Logit Blending achieves the highest accuracy on the Extremist Forums dataset. Sensitivity tuning yields per-group accuracy gains of up to 6 percentage points, and blended adapters produce rationales that a blinded GPT-4 judge ranks above the monolithic baseline's in coherence. This is evidence that monolithic training partly relies on spurious correlations.

The practical upshot: one model, one parameter. A children's platform weights the collaborative-editing adapter heavily; an adult forum favors the microblogging adapter. No retraining, no separate model per context.

These results are bounded by English-language data from 2017–2018, three fixed community types, and the use of GPT-4 as both rationale generator and evaluator. Future work should expand to multilingual settings, explore dynamic adapter loading to reduce the $\sim$30 GB memory footprint, and replace GPT-4 evaluation with human judges.

The principle may extend beyond hate speech. Other tasks where annotators legitimately disagree, such as sentiment or content rating, are natural candidates for treating that disagreement as structure rather than error, though we leave that evaluation to future work.

7 Limitations

We evaluate explanation quality using GPT-4 rather than human annotators due to the potentially offensive nature of hate speech content, though GPT-4's performance aligns with human agreement levels. Our router architecture assumes three distinct community types (microblogging, collaborative

editing, extremist forums) and may require adaptation for platforms with hybrid norms. Although the routing formulation is stated for a general number of experts N, all experiments use $N{=}3$; scaling behavior in the number of community adapters is not evaluated here. The 8-dimensional hidden layer was chosen empirically without exhaustive architecture search.

Our evaluation is limited to English-language datasets from 2017–2018 (HateXplain, Jigsaw, CrimeBB), which may not capture evolving discourse patterns or transfer to other languages with different cultural contexts for offensive speech. While our approach reduces parameters by $10\times$ compared to full fine-tuning, deployment still requires $\sim$30 GB memory for maintaining separate adapters. Future work should explore dynamic adapter loading and cross-lingual transfer.

Data and Code Availability

The HateXplain and Jigsaw training data and the corresponding GPT-4 rationale annotations, together with our adapter-training and routing code, are available from the authors on request. The CrimeBB data cannot be redistributed under the terms of our agreement with its providers; access requires a separate application to the CrimeBB maintainers at the Cambridge Cybercrime Centre. Llama 2 is used under Meta's community license, and GPT-4 outputs were obtained through the OpenAI API under its terms of service; these upstream licenses govern the model weights and outputs independently of the license covering this article.

Ethical Considerations

This article quotes verbatim examples of hateful and offensive language, including slurs, where necessary to illustrate model behavior; reader discretion is advised. The two annotators who relabeled 1,000 HateXplain test samples were volunteers who were informed in advance of the highly offensive nature of the material. All data analyzed in this work are pre-existing and publicly available, having been collected from public online forums and websites; the study involved no recruitment of or interaction with human participants, and no private or personally identifiable information was gathered. Because both the model evaluation and the two-annotator relabeling study are secondary analyses of publicly available data rather than research involving human subjects, institutional review board (IRB) approval was not required. Finally, the sensitivity control is intended for legitimate moderation contexts; the same mechanism could in principle be configured to under-enforce against particular groups, and we therefore recommend that deployments log and audit the sensitivity profiles they select.

Acknowledgment

The authors used OpenAI's GPT-4 in two roles: (i) to generate the structured natural-language rationales that supervise the adapters (the rationale-generation step of the Methodology), and (ii) as a blinded judge to rank rationale quality (the rationale-quality evaluation in the Results). All experimental design, analysis, and the scientific content of this article are the authors' own. No section of the manuscript was generated by AI; OpenAI's ChatGPT was used only for minor language and grammar editing, and the authors reviewed all such edits and take full responsibility for the final text.

References

[1] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[2] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection, 2022. arXiv:2012.10289.

[3] Cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, Nithum, and Will Cukierski. Toxic comment classification challenge, 2017.

[4] Sergio Pastrana, Daniel R Thomas, Alice Hutchings, and Richard Clayton. Crimebb: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 2018 World Wide Web Conference*, pages 1845–1854, 2018.

[5] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. arXiv:1701.06538.

[6] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[7] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015.

[8] Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[9] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[10] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[11] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, 2023.

[12] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words – a feature-based approach. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[13] Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *Italian Conference on Computational Linguistics*, 2018.

[14] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[15] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language, 2017. arXiv:1703.04009.

[16] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior, 2018. arXiv:1802.00393.

[17] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on Twitter. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

[18] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

[19] Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. Detect all abuse! toward universal abusive language detection models. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[20] Ju-Hyoung Lee, Jun-U Park, Jeong-Won Cha, and Yo-Sub Han. Detecting context abusiveness using hierarchical deep learning. In Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 10–19, Hong Kong, China, November 2019. Association for Computational Linguistics.

[21] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics.

[22] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics.

[23] Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[24] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[25] Fan Huang, Haewoon Kwak, and Jisun An. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pages 90–93, New York, NY, USA, 2023. Association for Computing Machinery.

[26] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online, August 2021. Association for Computational Linguistics.

[27] Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. Investigating the benefits of free-form rationales. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5867–5882, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[28] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-young Yun. Hare: Explainable hate speech detection with step-by-step reasoning. *arXiv preprint arXiv:2311.00321*, 2023.

[29] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. arXiv:2101.03961.

[30] Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Cross-token modeling with conditional computation, 2022. arXiv:2109.02008.

[31] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts, 2022. arXiv:2206.02770.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[33] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. arXiv:2305.14314.

[34] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition, 2024. arXiv:2307.13269.

[35] Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations, 2023. arXiv:2306.14870.

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. arXiv:2208.12242.

[37] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.

[38] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023.

[39] Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*, 2023.

[40] Mohammed Muqeeth, Haokun Liu, and Colin Raffel. Soft merging of experts with adaptive routing, 2023. arXiv:2306.03745.

[41] OpenAI. Gpt-4 technical report, 2023. arXiv:2303.08774.

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

[43] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. arXiv:1908.10084.

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. arXiv:2306.05685.

[45] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.