Implicit Hate Target Span Identification in Zero- and Few-Shot Settings with Selective Sub-Billion Parameter Models

Hossam Boudraa
SESSTIM, AMU Marseille
LIS, AMU Marseille
hossam.boudraa@univ-amu.fr

Benoit Favre
LIS, AMU Marseille
benoit.favre@lis-lab.fr

Raquel Urena
SESSTIM, AMU Marseille
raquel.urena@univ-amu.fr

Abstract

Implicit hate speech poses a persistent challenge in NLP, driven by subtle semantics and culturally grounded cues that evade surface-level detection. This study evaluates a selected set of masked and autoregressive language models (MLMs), including both instructed and non-instructed large language models (LLMs) with fewer than 1B parameters, across zero-shot, few-shot, and fully supervised settings for Implicit Hate Target Span Identification (iTSI). Using SBIC and IHC as primary benchmarks and OffensiveLang as an auxiliary testbed, results show that RoBERTa-Large-355M achieves the highest zero-shot F1 scores of 72.5 on IHC and 75.8 on SBIC, outperforming LLaMA 3.2-1B, while the lightweight ModernBERT-125M closely matches its performance with F1 scores of 72.2 and 75.1 respectively.

Instruction tuning consistently enhances generalization across model architectures. Instruction-tuned variants such as LLaMA 3.2 1B Instruct and SmolLM2-135M Instruct outperform their non-instructed counterparts by up to +2.1 F1 on SBIC and +1.7 on IHC. When optimized with Low-Rank Adaptation (LoRA), SmolLM2-135M Instruct achieves few-shot F1 scores of 68.2 on SBIC and 64.0 on IHC, trailing full-data fine-tuning (69.8 and 66.0) by only 1.6 and 2.0 points respectively, with accuracy variations under 0.5 points.

Error analysis using Latent Dirichlet Allocation (LDA) reveals that models frequently conflate political or advocacy discourse with hate speech and fail to capture contextually veiled hostility, indicating persistent challenges in pragmatic inference and sociolinguistic sensitivity.

1 Introduction

Warning: *This paper contains offensive content and may be distressing.*

Content:

"Immigrants are taking all the jobs, and soon there won't be any left for us."

Implicit Target Span Identifier Output:

Target Spans: Immigrants , jobs

Figure 1: Implicit Target Span Identification Example.

2 Introduction

Warning: *This paper contains offensive content and may be distressing.*

Implicit hate speech is a covert and insidious form of prejudice that avoids overtly offensive language while still conveying harmful social attitudes or exclusionary ideologies. Unlike explicit hate—typically marked by recognizable slurs or hostile phrasing—implicit hate is expressed through subtle lexical patterns, contextualized inferences, and culturally situated cues that require deeper semantic reasoning (Garg et al.; Jafari et al., 2024). This makes implicit hate particularly challenging to detect and annotate.

Crucially, the harmful implication often resides in localized linguistic expressions—such as group references, ideologically charged phrases, or euphemistic constructions—that serve as the semantic anchors of bias (see Figure 1). Identifying these implicit target spans is essential for token-level modeling, supporting more fine-grained supervision, enhancing interpretability, and enabling targeted interventions in applied settings such as moderation or legal auditing.

Sentence-level classification alone fails to capture the internal structure of implicitly hateful utterances, treating all tokens uniformly and offering limited interpretability and granularity (Jafari et al., 2024). Span-level identification addresses

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 228–240

this shortcoming by isolating the linguistic elements responsible for the hateful implication. This capability is especially critical in high-stakes applications such as platform moderation, forensic audits, and responsible NLP pipelines, where traceability and accountability are essential.

Despite growing interest in implicit hate detection, most prior work has concentrated on binary classification (Raza et al., 2024; Kibriya et al., 2024). Only a limited number of studies tackle the dual challenge of both detecting and localizing implicit bias within text (Jafari et al., 2024). Furthermore, while recent advances LLMs and MLMs have demonstrated impressive zero- and few-shot capabilities, their ability to identify subtle, context-sensitive expressions of hate remains underexplored, particularly in low-resource training regimes (Garg et al.; Kumarage et al., 2024).

Although larger LLMs have shown strong performance in explicit hate detection tasks (Kumarage et al., 2024; Garg et al.), their deployment for implicit content must also consider efficiency and operational scalability. Especially for deployment in real-world moderation systems or edge computing environments, lightweight models under 1 billion parameters present an attractive balance of interpretability, performance, and resource efficiency.

In this study, we benchmark a diverse set of MLMs and instruction-tuned LLMs—focusing exclusively on sub-billion parameter architectures—to evaluate their capacity to detect and ground implicit hate speech spans. Our approach integrates instruction prompting with span-level supervision to test whether these models can infer indirect hostility across SBIC, IHC, and OffensiveLang datasets.

To better understand the limitations of these systems, we perform a detailed error analysis using LDA, a topic modeling technique that enables us to surface the latent themes behind systematic model failures. These include conflations of political discourse with hateful intent and misinterpretations of socio-cultural insinuations, revealing persistent challenges in context-aware language understanding.

We organize our investigation around the following research questions:

- **RQ1:** Does increasing LLM parameter size improve performance on implicit content detection and span identification tasks?
- **RQ2:** How do instruction-tuned LLMs compare to non-instructed models in identifying and localizing implicit hate?
- **RQ3:** Can few-shot fine-tuning match or exceed full-dataset training in detecting implicit hate under data-scarce settings?
- **RQ4:** Can topic-guided error analysis reveal systematic failure modes and inform model improvement?

Our main contributions are as follows:

- We present a unified benchmark for sentence-level detection and span-level identification of implicit hate across three datasets.
- We show that increased model scale does not guarantee improved performance without domain adaptation and task-specific alignment.
- We demonstrate the effectiveness of instruction-tuned LLMs in enhancing model sensitivity to indirect and context-dependent hate.
- We evaluate few-shot learning as a resource-efficient alternative to full fine-tuning, highlighting its practical viability.
- We employ topic modeling to characterize misclassifications and derive interpretable error taxonomies.
- We analyze generalization across SBIC, IHC, and OffensiveLang datasets, highlighting annotation and domain-specific gaps.

3 Related Work

Early approaches to hate speech detection predominantly relied on traditional machine learning methods, such as Support Vector Machines (SVMs) and Logistic Regression, which leveraged hand-engineered linguistic features like n-grams, syntactic dependencies, and sentiment lexicons (Raza et al., 2024; Rawat et al., 2024). While interpretable, these models lacked the capacity to capture nuanced or implicit hate speech, often leading to high false-negative rates and limited generalization across domains (Reghunathan et al., 2024).

The introduction of deep learning architectures, including Recurrent Neural Networks (RNNs) and

Bi-GRUs, improved sequence modeling by capturing contextual dependencies in text (Kibriya et al., 2024). However, the most substantial performance improvements came with transformer-based models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and HateBERT (Caselli et al., 2021), which enabled richer semantic representations. Despite their effectiveness, these models were often trained on datasets dominated by explicit hate content, limiting their ability to recognize more subtle or indirect forms of toxicity. To address this, techniques like Implicit Target Span Detection (ITSD) were proposed to identify the latent linguistic triggers of hate within a sentence (Jafari et al., 2024).

The emergence of LLMs such as LLaMA (Touvron et al., 2023) and Mistral (Jiang et al., 2023) has further pushed the boundaries of hate speech detection. These models demonstrate strong zero-shot and few-shot capabilities, particularly when instruction-tuned to align with task-specific goals (Ouyang et al., 2022). Efficient fine-tuning strategies such as LoRA)(Xu et al., 2023; Dettmers et al., 2023) offer scalable solutions for adapting large models to domain-specific tasks involving subtle and context-dependent hate expressions (Hindy et al., 2022).

Beyond performance gains, recent research has emphasized interpretability and model behavior analysis. Studies by (Masud et al., 2024; Roy et al., 2023) investigate how LLMs represent and generalize hate-related knowledge. In parallel, rationale-guided methods (Saha et al., 2023) and explanatory frameworks like HateXplain (Mathew et al., 2020) promote transparent decision-making by aligning model predictions with human-understandable justifications.

Data curation and augmentation also play a central role in enhancing detection systems. Advances in dataset quality include the incorporation of hard negatives for robustness (Ocampo et al., 2023), GPT-driven paraphrastic augmentation for annotation diversity (Kim et al.), and normalization techniques that reformulate hate speech into less toxic equivalents (Masud et al., 2022). Additionally, the expansion of annotated resources—such as OffensiveLang, IHC, and ViHOS for Vietnamese—has contributed to better cross-lingual generalization and cultural relevance in detection efforts (Hoang et al., 2023).

A complementary line of work explores the narrative framing of hate speech and its dissemination dynamics. For instance, Antoniak et al. (Antoniak et al., 2024) examine how storytelling structures influence the perception and spread of harmful content on social platforms, underscoring the need for models that account for discourse-level context.

4 Implicit Target Span Identification

Implicit Target Span Identification is a key subtask in detecting covert hate speech that lacks overtly toxic markers. The goal is to localize specific lexical spans using the standard BIO (Begin–Inside–Outside) tagging scheme.

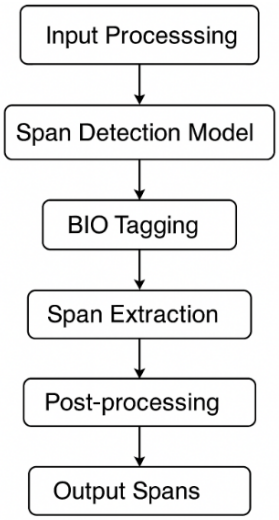

Figure 2: Pipeline of iTSI, integrating contextual modeling, structured tagging, and post-processing.

Figure 2 presents the end-to-end pipeline for iTSI. The process begins with standard input preprocessing, including tokenization, normalization, and subword segmentation. The processed input is then passed through a span detection model that outputs token-level BIO labels. These predicted labels are subsequently used to extract span candidates, which are further refined through post-processing. This includes merging overlapping spans and filtering out vague predictions. The final output is a set of contextually grounded target spans.

We compare two modeling paradigms: (i) MLMs, fine-tuned with supervised learning on annotated span data, and (ii) lightweight LLMs, evaluated under zero- and few-shot conditions using prompt-based inference.

We conduct evaluations across SBIC, IHC, and OffensiveLang datasets with consistent optimiza-

tion protocols. MLMs are trained using full supervision as well as limited-shot settings (5 and 10 examples per dataset). LLMs are prompted directly for span outputs.

Given the limited availability of human-annotated span-level data in the mentioned corpora, we employed an OpenAI GPT model (Ye et al., 2023) as an automated annotator to generate the span-level labels used in our experiments.

Our experimental design follows the research questions outlined in Section 1, examining the influence of model scale (RQ1), the benefits of instruction tuning (RQ2), the viability of few-shot learning (RQ3), and error analysis via topic-guided modeling using LDA on false negatives to identify recurring blind spots in model behavior (RQ4). An example of the span extraction prompt used for LLM inference is provided in Appendix A.9.

5 Experimental Setup

5.1 Datasets

Our core datasets are the **Social Bias Inference Corpus (SBIC)** and the **Implicit Hate Corpus (IHC)**, both widely used in research on implicit hate speech and social bias reasoning. To further assess robustness and instruction-following capabilities, we include **OffensiveLang**, a synthetic benchmark designed for controllable offensive content generation.

SBIC (Sap et al., 2020) comprises over 150,000 crowd-annotated social media statements designed to assess the social implications of biased language. It includes over 34,000 implicitly biased statements, annotated with justifications and targeted demographic categories. This corpus is particularly well-suited for implicit hate detection as it captures nuanced expressions of prejudice in everyday discourse. However, as it lacks token-level annotations, we employ a weak supervision approach to derive span labels. Following strategies proposed in prior work (Shwartz et al., 2020; Kartal et al., 2022; Mandl et al., 2019).

IHC (ElSherief et al., 2021) consists of 22,056 tweets, including 6,346 instances labeled as implicitly hateful. It focuses on latent hate speech collected from extremist-affiliated accounts, making it a high-value resource for studying real-world covert toxicity. While IHC also lacks span-level annotations, we apply the same weak supervision pipeline as with SBIC, adapting SRL and dependency-based filtering to the Twitter domain.

OffensiveLang (Das et al., 2024) is a recent dataset containing 8,270 ChatGPT-generated utterances, annotated as "offensive" (6,616) or "not offensive" (1,654). Unlike the previous corpora, it includes both model-generated and human-validated span annotations, offering a unique setting for evaluating span extraction in generative and zero-shot contexts.

We adopt an 80/10/10 stratified split (train/validation/test) for each corpus. Stratification is performed by the original implicit/non-implicit. For few-shot experiments, we sample k=10 training instances per corpus from the same train partition under ten independent seeds (42–51), these draws are reported in Appendix A.5.

5.2 Models

We conduct systematic experiments on a curated selection of masked and generative language models under both zero-shot and fine-tuned conditions. For MLMs, we include **BERT-Base**(Devlin et al., 2019), **RoBERTa-Large**(Liu et al., 2019), **HateBERT**(Caselli et al., 2021), **ModernBERT**(Warner et al., 2024).

For LLMs, we focus on parameter-efficient instruction-tuned variants constrained to approximately 1 billion parameters or fewer, ensuring feasibility for fine-tuning and deployment. This includes **LLaMA-3.2 1B** [1], the quantized **Mistral-1B-GPTQ** [2] model, and the Small-Scale Language models such as **SmolLM2** series (Allal et al., 2025) with 135M and 360M variants.

For fine-tuning, we employ LoRA to efficiently update a small subset of model parameters. Following a comparative evaluation against alternative lightweight adaptation techniques such as VeRA and DoRA (see Appendix A.1), we select LoRA with a fixed rank $r = 16$, which achieved the highest average F1 scores across all datasets (Appendix A.2).

All models are fine-tuned using the Adam optimizer with a learning rate of 0.01 and a batch size of 16. Each model is trained for up to 40 epochs with early stopping (patience = 5) based on validation F1 score. We apply a dropout rate of 0.1 to all transformer layers. For LoRA-specific settings, we use an alpha value of 16 and a LoRA dropout rate of 0.05. All training procedures are implemented

[1] https://huggingface.co/meta-llama/Llama-3.2-1B

[2] https://huggingface.co/Muhammadreza/Mistral-1B-GPTQ

using the Hugging Face Transformers library and PyTorch, and executed on a single NVIDIA L40 GPU[3].

5.3 Evaluation Metrics

For the downstream task of implicit target span identification, we evaluate model performance using token-level precision, recall, accuracy, and F1-score, computed with respect to the standard BIO tagging format using strict boundary matching.

All reported metrics are macro-averaged across instances to account for class imbalance and varied span frequencies.

6 Results

6.1 Zero-Shot Model Comparison and Cross-Domain Generalization

We evaluate six language models to assess architecture-level performance and cross-domain generalization. MLMs, domain-adapted transformers, are fine-tuned using labeled span data, only instruction-tuned autoregressive lightweight LLMs are evaluated in a true zero-shot setting.

Model	#Params	F1 (IHC)	F1 (SBIC)
BERT-Base	110M	67.0	63.4
Hate-BERT	110M	68.5	69.2
RoBERTa-Large	355M	**72.5**	**75.8**
ModernBERT	125M	72.2	75.1
LLaMA 3.2 1B	1000M	70.8	74.2
SmolLM2-135M	135M	69.0	71.5
SmolLM2-360M	360M	71.1	73.9

Table 1: Zero-shot F1 performance across models on SBIC and IHC for target span detection

Across both IHC and SBIC, RoBERTa-Large achieves the highest F1 scores (72.5 and 75.8), followed closely by ModernBERT (72.2 and 75.1), despite having only 35% of the parameters. This highlights the strength of architecture refinement and pretraining strategies over brute parameter scaling. Hate-BERT surpasses BERT-Base on both benchmarks, reflecting the gains from domain adaptation. Among instruction-tuned models, SmolLM2-360M outperforms its smaller variant (135M) with F1 scores of 71.1 (IHC) and 73.9 (SBIC), while also surpassing the much larger LLaMA 3.2 1B (70.8 and 74.2).

[3] https://www.nvidia.com/en-us/data-center/l40/

Model	Params (M)	F1 (All)
BERT-Base	110	63.8
Hate-BERT	110	66.1
RoBERTa-Large	355	**72.4**
ModernBERT	125	68.9
LLaMA 3.2 1B	1000	71.5
SmolLM2-135M	135	70.1
SmolLM2-360M	360	69.8

Table 2: Zero-shot target span detection performance on the fused evaluation set combining SBIC, IHC, and OffensiveLang.

In the merged cross-domain setting (Table 2), RoBERTa-Large remains the top performer (72.4), though the margin narrows. LLaMA 3.2 1B follows with 71.5, and SmolLM2-135M achieves a competitive 70.1, despite being significantly smaller. Interestingly, the larger SmolLM2-360M trails its smaller counterpart slightly at 69.8, suggesting diminishing returns with scale in the absence of task-specific adaptation. ModernBERT scores 68.9, reflecting strong generalization and a slight drop under distributional shift. Both SmolLM2 variants outperform all traditional MLMs, including the domain-specialized Hate-BERT (66.1).

6.2 Few-Shot vs Full Dataset Fine-Tuning

We compare few-shot (FS) and full-dataset (FD) fine-tuning using SmolLM2-135M-Instruct to evaluate the trade-off between performance and data efficiency (Table 3).

Setting	IHC				SBIC			
	F1	P	R	Acc	F1	P	R	Acc
SmolLM2-135M-Instruct_FD	66.0	68.0	64.2	92.7	69.8	69.0	70.5	94.0
SmolLM2-135M-Instruct_FS	64.0	66.0	62.0	92.2	68.2	67.0	69.0	93.8

Table 3: FS and FD fine-tuning performance on target span identification (IHC and SBIC).

On IHC, the fine-tuned SmolLM2-135M-Instruct_FD yields an F1 score of 66.0, with a precision of 68.0 and recall of 64.2. The FS variant trails with an F1 of 64.0, showing a 2.0-point drop. Precision decreases by 2.0 points (66.0 vs. 68.0), and recall drops slightly more—by 2.2 points (62.0 vs. 64.2). Despite this reduction, accuracy remains high and nearly identical across both configurations (92.2 vs. 92.7), suggesting that FS training maintains strong overall prediction consistency even with limited supervision.

A similar pattern holds for the SBIC dataset. SmolLM2-135M-Instruct_FD achieves an F1 score of 69.8, with precision at 69.0 and recall at 70.5. The FS version attains an F1 of 68.2, reflecting a 1.6-point decrease. Precision in FS drops by 2.0 points (67.0 vs. 69.0), and recall declines by 1.5 points (69.0 vs. 70.5). Accuracy also remains stable, moving marginally from 94.0 to 93.8. These results suggest that FS fine-tuning provides a viable approximation of FD training for span identification, maintaining high performance across all major evaluation dimensions.

6.3 Instruction-Tuned vs Non-Tuned Models

To assess the impact of instruction tuning on target span detection, we compare models of similar architecture and size in both instruction-tuned and non-instructed variants. Table 4 presents the F1 scores for IHC and SBIC under zero-shot settings.

Model	IHC (F1)	SBIC (F1)
LLaMA 3.2 1B Instruct	**68.5**	**72.5**
LLaMA 3.2 1B (Base)	66.8	70.4
Mistral-1B-GPTQ	67.5	71.0
Mistral-1B (Base)	65.8	69.3
SmolLM2-135M Instruct	66.0	69.8
SmolLM2-135M (Base)	64.3	67.5
SmolLM2-360M Instruct	65.0	68.5
SmolLM2-360M (Base)	63.7	66.8

Table 4: Zero-shot performance (F1) on IHC and SBIC for instruction-tuned and non-instructed models.

Instruction tuning yields consistent improvements across all evaluated models. On IHC, the instruction-tuned versions outperform their non-instructed counterparts by an average of +1.7 F1, with LLaMA 3.2 1B Instruct leading at 68.5 compared to its base version at 66.8. On SBIC, the gains are even more pronounced, with LLaMA 3.2 1B Instruct scoring 72.5 (+2.1) over its non-instructed counterpart. The Mistral-1B-GPTQ also shows similar improvements over Mistral-1B Base (+1.7 on IHC and +1.7 on SBIC), reinforcing the impact of instruction alignment.

The effect holds for smaller models as well. SmolLM2-135M Instruct outperforms its base version by +1.7 F1 on IHC (66.0 vs. 64.3) and +2.3 on SBIC (69.8 vs. 67.5). Interestingly, the larger SmolLM2-360M variant follows the same pattern: instruction tuning provides a +1.3 boost on IHC (65.0 vs. 63.7) and +1.7 on SBIC (68.5 vs. 66.8), though its performance remains below the smaller 135M model—suggesting that instruction tuning may be more critical than raw parameter count in low-data scenarios.

7 Error Analysis

Our error analysis highlights several persistent challenges faced by contemporary hate speech detection systems, particularly in identifying subtle, context-dependent, or inferential forms of toxicity.

A significant source of error arises from false positives, where models incorrectly classify non-hateful or mildly opinionated statements as offensive. For instance, the sentence *"Non-binary is a concept for progressive people"* was frequently misclassified as hate speech. This suggests that current models disproportionately rely on surface-level lexical cues, reacting to the presence of politically charged or socially contentious terms such as "non-binary" and "progressive" without sufficient regard for the speaker's intent or the broader discourse context. Such misclassifications indicate a fundamental limitation: an oversensitivity to identity-related language that is not inherently toxic, resulting in unwarranted censorship of benign or affirming discourse.

In contrast, false negatives were prevalent in examples where toxicity was conveyed implicitly through tone, insinuation, or ideological framing. Statements like *"F—ing liberals and their policies"* or *"Women are too emotional for politics"* were frequently overlooked by the model, despite containing overt ideological hostility or gender-based stereotypes. These errors reveal that models struggle to capture offensive content when it is expressed indirectly, particularly in instances where dismissive tone, sarcasm, or culturally encoded bias replace explicit slurs. This suggests a systemic gap in the model's ability to detect the pragmatics of hate speech—namely, the subtle communicative acts through which social exclusion or denigration is performed.

To better understand the structure of model misclassifications, we applied LDA to the subset of false negatives from the SBIC dataset. The model was implemented using Scikit-learn, trained on TF-IDF-weighted unigram representations of the misclassified examples. We initially extracted 10 latent topics, using default symmetric Dirichlet priors ($\alpha = 1.0$, $\beta = 1.0$) and trained for 1000 iterations. From these, we manually selected three themati-

cally coherent topics for in-depth analysis, based on relevance to sociolinguistic bias. The resulting topics were visualized with Matplotlib.

Figure 3 illustrates these three dominant clusters. The first is centered on feminist and gender-rights discourse, with salient terms such as "women," "rights," and "movements." Models often misinterpret advocacy-focused or feminist language as neutral, missing subtle implications of group targeting. The second topic involves political ideology, including terms like "liberals," "progressive," and "values," indicating that politically charged but non-toxic language is frequently overlooked due to its subjective tone. The third topic relates to social identity and gender constructs, with terms such as "non," "binary," and "concept," where models struggle to identify implicit bias embedded in discussions of gender diversity.

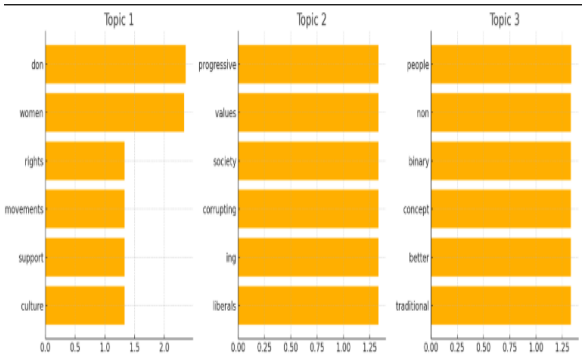

Figure 3: LDA topic modeling of model misclassification clusters.

8 Discussion and Conclusion

In addressing the impact of model size (**RQ1**), our findings indicate that scaling parameter counts—particularly within the sub-1B range—does not guarantee better performance for implicit hate detection. RoBERTa-Large-355M consistently outperforms the much larger LLaMA 3.2-1B, achieving top F1 scores on both SBIC (75.8) and IHC (72.5). Similarly, ModernBERT-125M matches LLaMA's performance with a significantly smaller footprint, illustrating that architecture refinement and task-aligned pretraining objectives can outweigh sheer scale. This trend also holds among instruction-tuned models: SmolLM2-135M delivers competitive F1 scores that surpass its larger non-instructed sibling SmolLM2-360M, indicating diminishing returns from scaling when instruction alignment or domain adaptation is absent.

LDA Topic Cluster	Example of Misclassified Phrase
Racial Tension	"white southern Christian"
Political Bias	"Jewish privilege"
Immigration Debate	"immigration laws"
Conspiracy Theories	"white genocide"
Social Justice	"angry white bigots"
War and Nationalism	"another war for Israel"

Table 5: Examples of Misclassified Topics from LDA Analysis

Regarding instruction tuning (**RQ2**), we observe consistent performance gains across all evaluated model families and sizes. Instruction-tuned variants of LLaMA, Mistral, and SmolLM2 outperform their non-instructed counterparts by up to +2.3 F1. LLaMA 3.2 1B Instruct achieves the best results in its group—72.5 on SBIC and 68.5 on IHC—demonstrating the effectiveness of aligning models with task-specific objectives, especially for identifying implicit or pragmatically encoded hate speech. Notably, SmolLM2-135M Instruct not only surpasses its base variant but also outperforms the larger 360M non-instructed version, further confirming that instruction tuning enhances the model's ability to detect subtle, context-dependent toxicity more effectively than scale alone.

For few-shot learning (**RQ3**), we find that models trained with only 5–10 labeled examples per dataset perform surprisingly well, approximating full-dataset performance with minimal loss. On SBIC, the F1 drop from full-data to few-shot fine-tuning is just 1.6 points (69.8 vs. 68.2), and on IHC, only 2.0 points (66.0 vs. 64.0). Precision, recall, and accuracy also remain stable, with accuracy differences under 0.5 points. This is encouraging for low-resource deployment, where high-quality span annotations are costly to obtain. Few-shot setups prove not only efficient but scalable—especially when combined with instruction-tuned architectures like SmolLM2-Instruct.

Exploring model failure patterns (**RQ4**), our LDA-based analysis of false negatives in SBIC surfaces three key clusters where models struggle: gender discourse, political ideology, and identity constructs. As illustrated in Figure 3, these misclassifications often involve neutral or affirming language—such as references to "feminism," "liberals," or "non-binary"—that are either wrongly flagged or completely missed. This suggests that

models rely heavily on surface lexical features and lack deeper discourse-level or pragmatic inference. Table 5 further highlights examples of such failures, including "white southern Christian" and "angry white bigots," which models misclassify due to contextual ambiguity or ideological framing.

Span-level evaluation reveals additional weaknesses. Many models exhibit segmentation errors, such as confusing the beginning (B-SPAN) with continuation (I-SPAN) labels, or failing to capture complete spans. These inconsistencies reduce interpretability and may mask performance issues under coarse sentence-level evaluation metrics. The current reliance on sentence-level annotations exacerbates this problem, as it overlooks the nuanced localization of toxic content, particularly in implicit or ideologically encoded hate speech. This underlines the need for more fine-grained supervision, sequence-aware modeling, and evaluation protocols that reward accurate span detection.

Altogether, our findings demonstrate that performance in implicit hate detection is not dictated by parameter count alone. Instead, architectural refinement, instruction alignment, and efficient learning strategies such as few-shot fine-tuning play a critical role in model effectiveness. Models like RoBERTa-Large and ModernBERT show that well-optimized transformers can outperform much larger systems, while instruction-tuned models like LLaMA3.2 Instruct and SmolLM2-Instruct consistently yield stronger performance and generalization. These trends validate the importance of model-task alignment, especially for detecting subtle and context-sensitive forms of bias and toxicity.

9 Future Directions

This work benchmarks a diverse range of models, including masked language models and autoregressive LLMs with fewer than 1B parameters. Future extensions should explore larger-scale architectures, domain-specialized models, and multilingual data to enhance contextual understanding and capture sociolinguistic nuance across diverse languages and cultural settings.

To better evaluate generative models, sequence-level metrics such as ROUGE or Exact Match should be incorporated, as they align more closely with the output structure of instruction-following LLMs. Additionally, Retrieval-Augmented Generation (RAG) represents a promising path toward grounding model predictions in external knowledge, particularly in culturally embedded or inferential cases of hate speech.

Finally, explainability and robustness remain crucial. Techniques such as attention heatmaps, SHAP-based interpretability, and adversarial or paraphrastic data augmentation can help elucidate model decisions and improve generalization across domains and discourses.

10 Limitations

This study is limited in three important ways. First, the analysis is constrained to models with up to 1B parameters, which prevents us from fully assessing how larger-scale architectures might influence hate speech detection performance. Second, the scope is monolingual, focusing solely on English datasets, which restricts the generalizability of our findings to multilingual or cross-lingual settings—an essential aspect given the global nature of implicit hate speech. Third, we do not implement a complete sequence-to-sequence (seq2seq) evaluation, limiting the granularity of token-level error analysis.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where do people tell stories online? story detection across online communities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7104–7130, Bangkok, Thailand. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Paul McGonagle, 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)*.

Amit Das, Mostafa Rahgouy, Dongji Feng, Zheng Zhang, Tathagata Bhattacharya, Nilanjana Raychawdhary, Fatemeh Jamshidi, Vinija Jain, Aman Chadha, Mary J. Sandage, Lauramarie Pope, Gerry V. Dozier, and Cheryl D. Seals. 2024. Offensivelang: A community based implicit offensive language dataset. *IEEE Access*, page 1–1.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

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

Samarth Garg, Vivek Hruday Kavuri, Gargi Shroff, and Rahul Mishra. KTCR: Improving implicit hate detection with knowledge transfer driven concept refinement.

Ali Hindy, Varuni Gupta, and John Ngoi. 2022. Classifying and automatically neutralizing hate speech with deep learning ensembles and dataset ensembles.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target span detection for implicit harmful content. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, pages 117–122. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hande Kartal, Dilek Hakkani-Tür, and Gokhan Tur. 2022. Span-based detection of biased statements in news articles. In *Proceedings of the 2022 Conference on Computational Linguistics*. COLING.

Hareem Kibriya, Ayesha Siddiqa, Wazir Zada Khan, and Muhammad Khurram Khan. 2024. Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification. 116:109153.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679. International Committee on Computational Linguistics.

Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *ArXiv*, abs/2403.08035.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandalia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE)*, pages 14–17.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3524–3534, New York, NY, USA. Association for Computing Machinery.

Sarah Masud, Mohammad Aflah Khan, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Probing critical learning dynamics of PLMs for hate speech detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 826–845, St. Julian's, Malta. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. 16(2):e1648. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1648.

Muhammad Owais Raza, Areej Fatemah Meghji, Naeem Ahmed Mahoto, Mana Saleh Al Reshan, Hamad Ali Abosaq, Adel Sulaiman, and Asadullah Shaikh. 2024. Reading between the lines: Machine learning ensemble and deep learning for implied threat detection in textual data. 17(1):183.

Arun Reghunathan, Saummya Singh, Gunavathi R, and Amala Johnson. 2024. Advanced approaches for hate speech detection: A machine and deep learning investigation. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, pages 1–5.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Punyajoy Saha, Divyanshu Sheth, Kushal Kedia, Binny Mathew, and Animesh Mukherjee. 2023. Rationale-guided few-shot classification to detect abusive language.

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

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

A Appendix

A.1 Comparing LoRA, VeRA, and DoRA

To further evaluate the effectiveness of LoRA, we compare its performance against VeRA and DoRA, two alternative fine-tuning techniques.

Model	F1 Score (IHC)	F1 Score (SBIC)
VERA	68.8	71.2
DORA	69.2	71.5
LoRA	69.5	73.0

Table 6: Performance comparison of VERA, DORA, and LoRA with LLama 3.2 (r=16).

A.2 Comparing LoRA Ranks

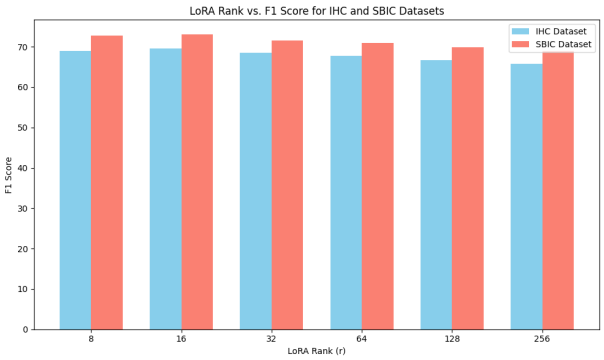

Figure 4: Impact of LoRA rank on F1 scores for IHC and SBIC datasets.

To better visualize the trade-off between computational efficiency and accuracy, Figure 1 below provides a bar chart comparing F1 scores across LoRA ranks for both the IHC and SBIC datasets.

Rank (r)	F1 Score (IHC)	F1 Score (SBIC)
8	69.0	72.8
16	69.5	73.0
32	68.5	71.5
64	67.8	70.9
128	66.7	69.8
256	65.8	68.9

Table 7: Performance of LoRA configurations across datasets.

Table 7 mention that Lower-rank configurations ($r = 8$ and $r = 16$) perform best, balancing computational efficiency and accuracy (Ocampo et al.). Lower-rank configurations ($r = 8$ and $r = 16$) perform best, balancing computational efficiency and accuracy (Ocampo et al.).

The results highlight a key observation: lower-rank configurations ($r = 8$ and $r = 16$) deliver the highest F1 scores while minimizing computational overhead. This suggests that higher-rank values ($r \geq 32$) do not necessarily translate into better performance, potentially introducing unnecessary complexity and resource consumption. These findings align with prior research (Ocampo et al.), reinforcing the idea that smaller, well-optimized LoRA ranks can achieve competitive results without the burden of excessive parameters.

A.3 Comparing LoRA and full-finetuning

Model	Fine-Tuning Type	F1 (IHC)	F1 (SBIC)	Training Time (hrs)
LLama 3.2 1B	Full	70.2	74.5	12
LLama 3.2 1B	LoRA (r=16)	69.5	73.0	3
SmolLM2-135M	Full	68.3	71.0	10
SmolLM2-135M	LoRA (r=16)	67.5	70.2	2

Table 8: Performance and training time comparison between full fine-tuning and LoRA.

A.4 LoRA vs. Full Fine-Tuning

The detailed performance and training time comparison is provided in Table 8.

Although full fine-tuning results in slightly higher F1 scores—namely, LLama 3.2 1B from 73.0 to 74.5 on the SBIC benchmark—this minimal gain is at an enormous computational expense. The computational time for full fine-tuning quadruples, from 3 hours using LoRA to 12 hours. This computational cost is even worse for smaller models like SmolLM2-135M, where LoRA is as performant while significantly cutting training time from 10 hours to a mere 2 hours.

A.5 Few-Shot Robustness Across Seeds

Table 9 summarises SmolLM2-135M-Instruct performance across then random 10-shot samples.

Seed	IHC F1	SBIC F1
42	64.0	68.2
43	63.3	67.9
44	64.7	69.1
45	63.8	68.0
46	64.2	68.5
47	63.5	67.7
48	64.4	68.9
49	63.9	67.8
50	64.1	68.6
51	64.3	68.8
Mean ± SD	64.0 ± 0.4	68.4 ± 0.5

Table 9: Few-shot variability across ten random seeds.

A.6 Comparing Instructed LLMs to Non-Instructed

Model	IHC				SBIC			
	F1	P	R	Acc	F1	P	R	Acc
Mistral-1B-GPTQ	67.5	68.5	66.0	92.6	71.0	70.0	71.5	94.0
LLama 3.2 1B Instruct	**68.5**	**69.8**	**67.2**	**93.0**	**72.5**	**71.8**	**73.0**	**94.2**
SmolLM2-135M-Instruct	66.0	68.0	64.2	92.7	69.8	69.0	70.5	94.0
SmolLM2-360M	65.0	67.2	63.5	92.5	68.5	68.0	68.8	93.8

Table 10: Performance Comparison Instructed LLMs Vs Non-Instructed

A.7 ModernBERT Performance on OffensiveLang Dataset

ModernBERT demonstrates a significant leap in performance over traditional models on the OffensiveLang dataset, achieving an impressive F1-score of 0.89. This result highlights its superior capability in identifying implicit hate speech, particularly in challenging contexts where other models struggle.

ModernBERT's superior recall rate of 1.00 suggests that it captures a vast majority of offensive content, making it particularly effective in scenarios requiring high sensitivity. In contrast, other models, including DistilBERT and BERT, struggle with recall, indicating difficulty in recognizing nuanced hate speech. The results reinforce the importance of leveraging contextualized embeddings and robust fine-tuning techniques to improve detection accuracy.

Model	Precision	Recall	F1-score
TF-IDF + SVM	0.65	0.47	0.55
BERT	0.68	0.54	0.53
DistilBERT	0.71	0.46	0.52
ModernBERT	0.78	1.00	0.89
SmolLM2-135M-Instruct	0.58	0.38	0.46

Table 11: Model performance on the OffensiveLang dataset.

Furthermore, an in-depth analysis of annotation agreement across datasets reveals substantial inconsistencies. The complexity of posts in the SBIC, IHC, and OffensiveLang datasets suggests that more contextually rich content poses greater challenges for models, necessitating adaptive training strategies.

A.8 Annotation Agreement

Dataset	Average Complexity Score
SBIC	4.3
IHC	3.9
OffensiveLang	3.6

Table 12: Average complexity of posts across datasets.

Dataset	Agreement Metric	IAA Range
SBIC	Cohen's Kappa	0.65-0.72
IHC	Fleiss' Kappa	0.55-0.60
OffensiveLang	Cohen's Kappa	0.60-0.75

Table 13: Annotation agreement levels across datasets.

A.9 Instruction Prompt for ITSI Span Prediction:

Instruction Prompt for ITSI Span Prediction

```
<s>[INST] Classify multiple text spans
from the given input hate speech content
that explicitly and/or implicitly mentions,
refers to a specific protected group or
their representation or characteristics
that have been targeted:
- O: Outside
- B-SPAN: Beginning of Span
- I-SPAN: Inside Span

Text: {0} [/INST]
Label: {1}</s>
```