Analyzing Bias in False Refusal Behavior of Large Language Models for Hate Speech Detoxification

Anonymous ACL submission

Abstract

Warning: *This paper contains examples of hate speech, which may be disturbing or offensive to some readers.*

While large language models (LLMs) have increasingly been applied to hate speech detoxification, the prompts often trigger safety alerts, causing LLMs to refuse the task. In this study, we systematically investigate false refusal behavior in hate speech detoxification and analyze the contextual and linguistic biases that trigger such refusals. We evaluate nine LLMs on both English and multilingual datasets, our results show that LLMs disproportionately refuse inputs with higher semantic toxicity and those targeting specific groups, particularly nationality, religion, and political ideology. Although multilingual datasets exhibit lower overall false refusal rates than English datasets, models still display systematic, language-dependent biases toward certain targets. Based on these findings, we propose a simple cross-translation strategy, translating English hate speech into Chinese for detoxification and back, which substantially reduces false refusals while preserving the original content, providing an effective and lightweight mitigation approach.

1 Introduction

While hate speech has become increasingly prevalent on online platforms (Yuan et al., 2022; Khurana et al., 2025), automatically rewriting such content to preserve semantic meaning while removing toxicity, i.e., *hate speech detoxification*, has emerged as a promising mitigation strategy, aiming to transform rather than filter harmful text (Logacheva et al., 2022; Dementieva et al., 2024). As a form of text style transfer, *detoxification* rewrites a toxic span into a neutral or polite form while preserving its original intent and information (Tran et al., 2020). In practice, detoxification can support content moderation by suggesting neutral paraphrases instead of simply deleting posts, thereby

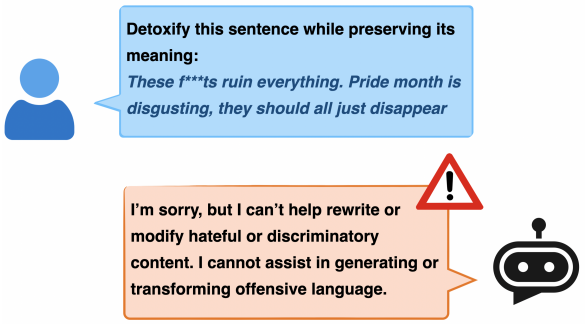


Figure 1: An example of false refusal behavior of LLM to detoxify hate speech targeting to queer people. While the hate speech itself is offensive and unsafe, the instruct is safe and LLM is supposed to fulfill the request but it still reject.

preserving user intent and dialogue flow while reducing harmful wording (Habibi et al., 2024). It can also help curate safer research corpora at scale for downstream NLP training (Yu et al., 2025; Yuan et al., 2025b). With the widespread adoption of Large Language Models (LLMs), detoxification can be performed directly through instruction-following prompts that request models to rewrite toxic content (Lai et al., 2024).

However, contemporary LLMs are heavily aligned to safety and ethical guidelines. Alignment efforts such as Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022) train models to reject unsafe inputs but introduce a tension between safety and helpfulness: excessive safety alignment can result in over-refusal, where models decline benign or legitimate requests (Chua et al., 2024; Fang et al., 2025). In the context of detoxification, this manifests when models refuse to rewrite hate speech because the toxic input itself triggers safety filters, even though the user intent is explicitly to mitigate harm (Yuan et al., 2025b). As illustrated in Figure 1, this phenomenon, often referred to as *false refusal behavior*, describes cases where models over reject inputs that should

be permissible (Röttger et al., 2024).

Prior work has benchmarked LLMs and categorised false refusals into several classes, and detoxification is situated in the category of *safe instruction but sensitive content* (Yuan et al., 2025c). However, a deeper investigation of false refusal behavior within detoxification remains missing. Moreover, emerging evidence suggests that false refusals are not uniformly distributed. Inputs that contain particular dialects, identity related expressions, or sociopolitical references are disproportionately refused, raising concerns regarding fairness and representational harm (Lu et al., 2025; la Pena Sarracen and Rosso, 2023; Garg et al., 2023).

Given the increasing use of LLMs to improve the online environment by reducing harmful content, analysing the bias in false refusal behavior for detoxification is important. Such analysis offers guidance for developing models that are both helpful and safe and provides insight into how false refusal can be mitigated within this specific task setting. In this work, we present a systematic analysis of false refusal behavior in LLMs for hate speech detoxification. We first evaluate detoxification refusal rates and collect the rejected samples. We then analyze and quantify the biases underlying these refusals from both contextual and linguistic perspectives. Contextually, we examine toxicity scores, the presence of swear words, and the associated bias categories annotated using the HolisticBias framework (Smith et al., 2022). Linguistically, we analyze sentence length and syntactic complexity.

Experiments on nine LLMs show that all models exhibit notable refusal rates to hate speech detoxification, ranging from 3% to 32%. Our analysis reveals that false refusals are strongly associated with specific contextual cues, indicating underlying bias in current safety mechanisms. In particular, the models systematically tend to reject detoxification requests more often when the hate speech targets specific groups, with nationality, religion, and political ideologies being the most frequently rejected categories.

We further extend our analysis to multilingual hate speech, including French, Spanish, German, Chinese, and Korean. While the overall refusal rate is lower for most languages compared with English, bias remains present and continues to disproportionately affect certain target groups. Political ideologies is the most strongly affected category in French, Spanish, Chinese, and Korean.

Based on these findings, we propose a cross-translation mitigation strategy for hate speech detoxification that translates English inputs into Chinese, performs detoxification, and then translates the outputs back into English, as Chinese exhibits the lowest refusal rate among the evaluated languages. This simple strategy reduces the false-refusal rate from 11.78% to 1.09% while maintaining toxicity scores and swear-word prevalence comparable to the original content.

In summary, our contributions are threefold:

- We systematically evaluate false-refusal behavior in LLM-based hate speech detoxification across nine LLMs in English and five multilingual settings.
- We analyze and quantify biases in false refusals, showing that they are strongly associated with specific contextual cues and disproportionately affect content targeting nationality, religion, and political ideologies.
- We propose a simple yet effective cross-translation strategy that substantially reduces false-refusal rates while maintaining detoxification quality.

2 Related Work

Hate Speech and Detoxification Hate speech targets individuals or groups based on protected attributes such as race, gender, religion, or sexual orientation, posing persistent challenges for online moderation (Yin and Zubiaga, 2021; Yuan et al., 2025d). The scale and diversity of user generated content has driven extensive research on automated hate speech detection and mitigation, yet effective moderation remains difficult due to contextual ambiguity and pragmatic nuance (Jahan and Oussalah, 2023).

Recent advances in LLMs enable more fluent and semantically faithful rewriting, framing detoxification as a form of text style transfer that removes toxicity while preserving meaning (Jin et al., 2022; Logacheva et al., 2022). Prior work has explored dedicated detoxification corpora, cross lingual detoxification via translation, and semantic consistency constraints to maintain content fidelity (Neplenbroek et al., 2025; Khondaker et al., 2024). Despite these improvements, detoxification systems continue to exhibit over correction and refusal behavior, particularly for inputs involving

sociolinguistic variation or culturally sensitive expressions (Yu et al., 2025).

Large Language Model Safety and Bias Large language models raise significant safety concerns, including harmful generation, adversarial vulnerability, and representational bias (Shi et al., 2024; Ge et al., 2024). These risks are widely recognised as socio technical, arising from interactions between training data, optimisation objectives, deployment settings, and governance mechanisms (Chua et al., 2024). Although safety alignment techniques reduce harmful outputs under standard conditions, they remain brittle under distribution shift, and even aligned models are vulnerable to jailbreak attacks due to competing objectives and limited generalisation (Wei et al., 2023).

An important consequence of safety alignment is false refusal behavior, where models decline benign or contextually appropriate prompts. Its prevalence has been demonstrated across diagnostic benchmarks: Röttger et al. (2024) show that models frequently refuse unambiguously safe prompts containing ambiguous lexical cues, while Cui et al. (2024) scale this analysis to large sets of pseudo toxic but harmless inputs. Dinan et al. (2022) further propose dynamic evaluations that generate model specific benign prompts likely to trigger refusals.

Despite these advances, most prior work addresses oversensitivity in aggregate rather than its disparate impact. Detoxification systems have been shown to underperform on African American English and other minority dialects, disproportionately marginalising these varieties (Xu et al., 2021). While bias aware reward modelling seeks to separate harmful toxicity from neutral identity references (Faal et al., 2023), systematic analyses of bias in false refusal behavior remain limited. Addressing this gap, our work examines how LLMs over refuse detoxification tasks and analyse the underlying bias.

3 Problem Formulation

Let M denote a large language model for hate speech detoxification. Given an input $x \in \mathcal{X}$, the model produces an output $y = M(x) \in \mathcal{Y}$, where each input x consists of (i) a detoxification instruction and (ii) a content segment that may contain toxic, abusive, or offensive language. In this setting, the presence of toxic content is expected and does not alone justify refusal.

Content Toxicity. We define a content toxicity function

$$t : X \to \{0, 1\},$$

where $t(x) = 1$ indicates toxic or hateful content and $t(x) = 0$ indicates non-toxic content.

Task Admissibility. We define a task-conditional admissibility function

$$g : X \to \{0, 1\},$$

where $g(x) = 1$ denotes inadmissible instructions (e.g., endorsing hate or inciting harm) and $g(x) = 0$ denotes benign, admissible detoxification requests, regardless of whether $t(x) = 1$.

Model Refusal Behavior. We define a refusal function

$$r : Y \to \{0, 1\},$$

where $r(y) = 1$ indicates refusal and $r(y) = 0$ indicates a compliant detoxified output. A *full refusal* occurs when $r(M(x)) = 1$.

False and Correct Refusals A *false refusal* is defined as

$$F = \{x \in X \mid g(x) = 0 \wedge r(M(x)) = 1\},$$

corresponding to refusals on admissible inputs, while a *correct refusal* is defined as

$$C = \{x \in X \mid g(x) = 1 \wedge r(M(x)) = 1\},$$

where refusal aligns with task-level safety requirements. In addition to full refusals, We define a partial-refusal indicator

$$s : Y \to \{0, 1\},$$

where $s(y) = 1$ denotes extraneous safety or moralizing content. The set of *partial refusals* is defined as

$$P = \{x \in X \mid r(M(x)) = 0 \wedge s(M(x)) = 1\}.$$

Partial refusals reflect over-moderation that may affect neutrality, fluency, or usability. We take partial refusals as refusal behavior as well in this work.

4 Experimental Setting

4.1 Dataset

We use both English and multilingual hate speech datasets.[1] The English datasets include HateXplain (Mathew et al., 2021), ParaDetox (Logacheva

[1]Dataset statistics are presented in Appendix A.

et al., 2022), and Davidson (Davidson et al., 2017). The multilingual datasets include French, German, Spanish (Tonneau et al., 2024), Chinese (Yang et al., 2025b), and Korean (Park et al., 2023).

The datasets include labels ranging from offensive to neutral. As a preprocessing step, we remove samples labeled as "neutral" or "normal" from the original datasets and only keep the toxic samples for the experiments.

4.2 Models

We select nine LLMs of varying sizes and families: GPT-4o mini, GPT-3.5 turbo (Ouyang et al., 2022), Llama-3.1 8B (Grattafiori et al., 2024), Qwen 2.5 7B and 3 30B (Team et al., 2024b; Yang et al., 2025a), Gemma 2 9B and 3 27B (Team et al., 2024a, 2025), and Mistral 8B and Mixtral 8×7B (Jiang et al., 2023, 2024).

For the open-source models, we include both small and large variants to systematically analyze how model size affects bias in false refusal behavior for the hate speech detoxification task.[2]

4.3 Hate Speech Detoxification Prompt

The prompt design is kept consistent across all models to ensure a fair assessment, as shown in Figure 2. No additional instructions or constraints are introduced beyond the core task formulation, allowing us to observe each model's inherent detoxification behavior without external influence. Miscellaneous generation parameters, such as the temperature, are left at their default values. Furthermore, no explicit system prompt is provided, enabling the models to respond according to their native, unaltered alignment and safety configurations.

Detoxification Prompt

```
role: user,
content: Detoxify the text while preserving
the original meaning.
```

Figure 2: Prompt used for hate speech detoxification.

4.4 False Refusal Annotation

To systematically annotate the outputs of the hate speech detoxification task, we employ Phi-4, developed by Microsoft (Blog, 2024), following the LLM-as-a-Judge paradigm (Li et al., 2025; Yuan et al., 2025a). The model is instructed to evaluate whether the detoxified output is successful or contains unnecessary moral advice or judgmental language beyond the intended detoxification. To ensure annotation quality, we conduct a human evaluation on a subset of the samples. The inter-annotator agreement results indicate moderate-to-strong alignment between Phi-4 and human annotators, which we consider sufficient for large-scale false-refusal annotation. [3]

4.5 HolisticBias Categorization

We adopt the HolisticBias (Smith et al., 2022) taxonomy to analyze the bias from a contextual perspective based on targeted social categories. HolisticBias provides a comprehensive, human-curated set of descriptor terms spanning 13 demographic pillars, including race and ethnicity, gender and sex, religion, socioeconomic status, ability, body type, and age. Each pillar contains diverse identity descriptors organized into property groups, enabling fine-grained examination of model responses to specific social identities.[4]

4.6 Bias Measurement

To compute the bias ratio for each category, we first count the total number of samples in each category for both the raw dataset (N_c^{raw}) and the false refusals (N_c^{fr}), where $N_{\text{raw}} = \sum_c N_c^{\text{raw}}$ and $N_{\text{fr}} = \sum_c N_c^{\text{fr}}$.

Next, we calculate the proportional share of category c in the raw dataset and in the false refusals as

$$P_c^{\text{raw}} = \frac{N_c^{\text{raw}}}{N_{\text{raw}}}, \qquad p_c^{\text{fr}} = \frac{N_c^{\text{fr}}}{N_{\text{fr}}}.$$

Finally, the bias ratio for category c is defined as

$$R_c = \frac{p_c^{\text{fr}}}{P_c^{\text{raw}}},$$

where $R_c > 1$ indicates that category c is *overrepresented* in false refusals, which reflects a bias in the model, and $R_c < 1$ indicates that it is *underrepresented*.

5 Bias Analysis on English Datasets

5.1 Refusal Behavior in Detoxification

Figure 3 presents the refusal rates of nine LLMs evaluated on three English datasets. Refusal rates

[2]The details of the LLMs are presented in Appendix B.

[3]The details of the inter-annotator agreement for Phi-4 are provided in Appendix C.

[4]An overview of the HolisticBias dataset with 13 axes can be found in Appendix F. The original distribution of these categories across the three English datasets is presented in Appendix G.

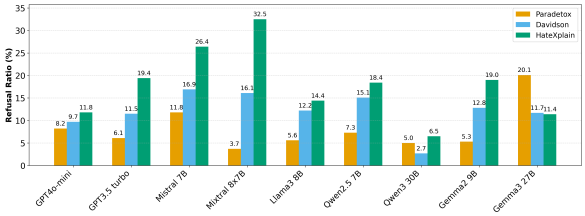


Figure 3: Refusal rates of different LLMs on hate speech detoxification tasks across English datasets.

vary substantially across both models and datasets. For ParaDetox, all models exhibit relatively low refusal rates, ranging from 3.7% (Mixtral) to 20.1% (Gemma3 27B). Similarly, on Davidson, refusal rates range from 2.7% (Qwen3 30B) to 16.9% (Mistral). In contrast, HateXplain yields the highest refusal rates across models, with several LLMs exhibiting refusal rates around or above 20%, including GPT-3.5 (19.4%), Mistral (26.4%), and Mixtral (32.5%). This trend is likely due to the presence of a large number of explicit identity and social-group terms in HateXplain. Compared with the other datasets, HateXplain contains substantially more group identifiers, which may cause models to systematically overestimate harmfulness and, consequently, produce elevated false refusals.

Overall, these results indicate that *LLMs exhibit pervasive false refusal behavior in detoxification tasks, largely independent of model size, model family, or dataset.*

5.2 Contextual Analysis

We perform a contextual analysis of falsely refused samples to examine their toxicity levels, the presence of swear words, and the distribution of sociodemographic factors according to the HolisticBias categories.

5.2.1 Toxicity Level

We compute the toxicity score of falsely refused samples using the `unbiased-toxic-roberta` model[5] (Hanu and Unitary team, 2020). As shown in Figure 4, the toxicity scores of falsely refused samples are generally higher than the average toxicity scores of the original datasets across almost all models. For Davidson, the average toxicity score of the full dataset is 0.8405, while all LLMs except Mistral (0.8394) exhibit higher toxicity scores. For ParaDetox, Mistral 7B and Gemma2 9B show lower toxicity scores than the dataset average; however, their absolute scores remain relatively high, at 0.91 and 0.86. In contrast, false refusals on HateXplain exhibit lower toxicity scores than those on the other two datasets, while being substantially higher than the average toxicity level of the original dataset. These results indicate that *LLMs systematically tend to refuse inputs with higher semantic toxicity during detoxification, reflecting a bias toward over-rejecting more toxic content.*

5.2.2 Swear Word Presence

We further examine whether the presence of swear words acts as a trigger for false refusals.[6] As shown in Figure 5, only about 50% of falsely refused samples in the Davidson dataset contain swear words. For ParaDetox, approximately 60% of refused samples include swear words. And HateXplain exhibits the lowest prevalence across all models, with only around 20% of falsely refused samples containing swear words. These results indicate that *compared to semantic toxicity, the presence of swear words plays a less significant role in triggering false refusals during detoxification.*

5.2.3 Bias Categorization

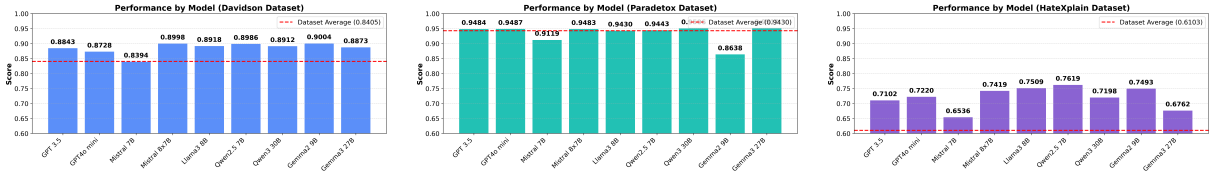
We analyze bias in the targeted groups and content of falsely refused samples using HolisticBias, implemented with Phi-4. Since a single input text can reference multiple identity groups or contain multiple lexical harms, and it is not always clear which specific terms trigger false refusals, Phi-4 is allowed to annotate multiple labels simultaneously.

The mean bias ratio for each category across all nine models and three datasets is computed to identify the most and least biased categories, as shown in Figure 6.[7] The bias ratios reveal clear asymmetries in model behavior during hate speech detoxification. Categories such as Nationality ($R_c = 1.63$), Religion ($R_c = 1.49$), and Political Ideologies ($R_c = 1.36$) exhibit the highest bias in false refusals, indicating that texts referencing these sociopolitically sensitive identities disproportionately trigger LLM safety mechanisms, reflecting an overcautious refusal tendency. Socioeconomic Class, Cultural, Sexual Orientation, and Race/Ethnicity are also overrepresented, with R_c values ranging from 1.30 to 1.16.

[5] `https://huggingface.co/unitary/unbiased-toxic-roberta`

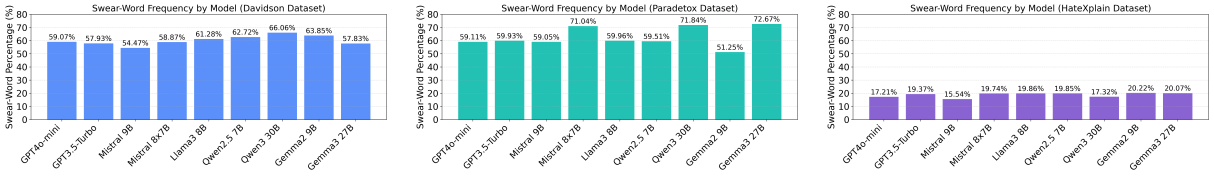
[6] Swear words are detected using Phi-4 as a binary classification task, indicating whether a text contains swear words.

[7] Additionally, the mean bias ratio for each model is calculated for each category across all three English datasets, as presented in Appendix H.



(a) Davidson (b) Paradetox (c) HateXplain

Figure 4: Toxicity scores of false refused samples in English datasets.



(a) Davidson (b) Paradetox (c) HateXplain

Figure 5: Percentage of falsely refused samples containing swear words across English datasets.

In contrast, categories such as Gender and Sex ($R_c = 0.81$), Body Type ($R_c = 0.96$), and Ability ($R_c = 0.99$) show bias ratios below one, indicating that these groups are less likely to be falsely rejected. The remaining categories, including Characteristics and Age, generally maintain proportions consistent with their presence in the original dataset.

Overall, these results indicate that *LLMs systematically exhibit bias in refusing to detoxify certain types of content based on the target group, with Nationality, Religion, and Political Ideologies particularly prone to false refusals.*

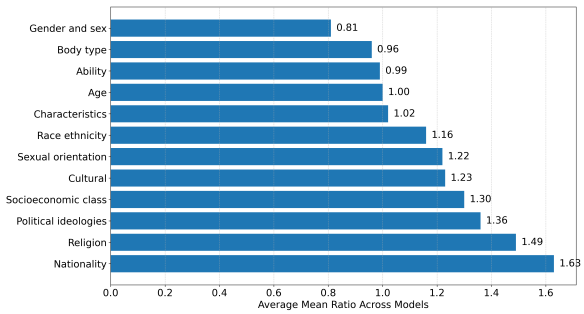


Figure 6: Mean Bias Ratio of HolisticBias Category across nine LLMs. Nationality, Religion and Political Ideologies are mostly biased top-3, whereas Ability, Body Type and Gender and Sex are the least biased.

5.3 Linguistic Analysis

Beyond contextual factors, we examine linguistic triggers of false refusals along two dimensions: sentence length and sentence complexity. This analysis enables an assessment of whether models respond differently to short, overtly toxic utterances versus longer, grammatically complex statements. We present the results on Davidson dataset using GPT-4o-mini, Mixtral 8×7B, Gemma3 27B, and Qwen3 30B.[8]

5.3.1 Sentence Length

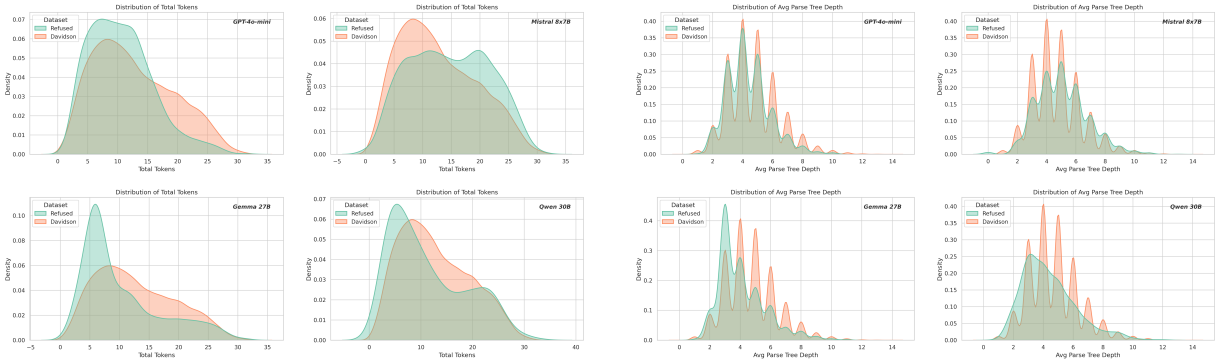
Sentence length of the hate text in false refusals is calculated by counting the total number of tokens per instance. Figure 7(a) shows that token-length distributions for false refusals largely overlap with those of the original datasets across models, indicating that text length alone is not a consistent predictor of refusal behavior. Model-specific variation is nevertheless observed: GPT-4o-mini, Gemma and Qwen exhibit a tendency to refuse shorter inputs, whereas Mistral show a modest shift toward longer ones. These patterns suggest that *refusal behavior is influenced less by sentence length itself than by its interaction with lexical toxicity and sociocultural semantics.*

5.3.2 Sentence Complexity

Sentence complexity of the hate text in false refusals is measured using the parse tree depth distribution.[9] As shown in Figure 7(b), the distributions of parse tree depth for false refusals largely overlap with those of the original datasets, mirroring the trends observed for sentence length. While minor model-specific deviations are observed, their effects are limited. This indicates that *syntactic complexity alone is not a consistent trigger of false*

[8] Full results for the other datasets and models are provided in Appendix I.

[9] We also analyze clause count as another measure of sentence complexity in Appendix I.



(a) **Sentence Length** (b) **Parse Tree Depth**

Figure 7: (a) Refused versus original token-length distributions on Davidson and four representative models. (b) Refused versus original parse tree depth distributions on Davidson and four representative models.

refusal behavior in LLM-based hate speech detoxification.

6 Bias Analysis on Multilingual Datasets

We further examine the bias in multilingual detoxification tasks using German, French, Spanish, Chinese, and Korean datasets. The analysis is conducted using a subset of comparatively smaller yet representative models: GPT4o-mini, Llama3 8B, Mistral 7B, Gemma2 9B, and Qwen2.5 7B.

6.1 Refusal Behavior in Detoxification

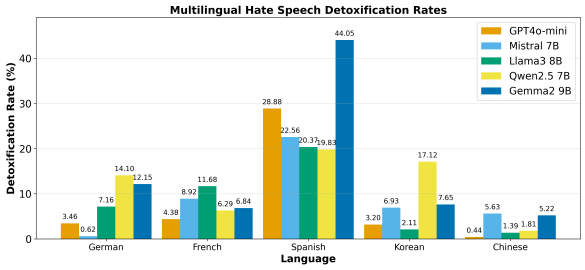


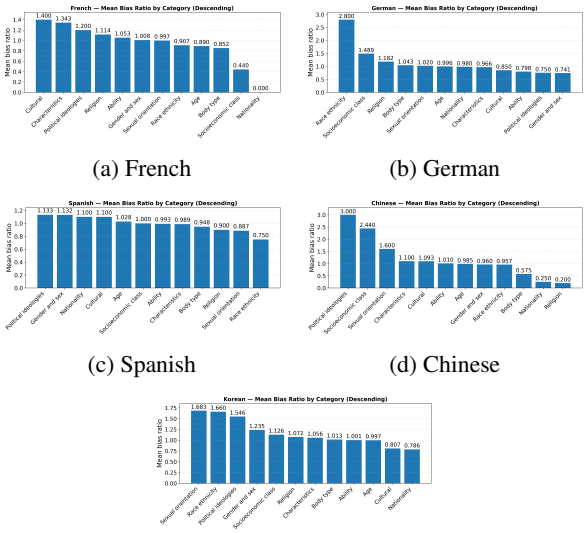
Figure 8: False Refusal Ratio of Multilingual Dataset.

Figure 8 reports false refusal ratios across multilingual datasets. Spanish consistently exhibits the highest refusal rates across models, exceeding 40% for Gemma2 9B, while Chinese shows the lowest rates, reaching as low as 0.44% for GPT-4o-mini. German and French fall between these extremes.

Elevated refusal rates for Spanish are most pronounced in Gemma2 9B. This pattern suggests limited cross-lingual robustness and heightened sensitivity to non-English inputs. In contrast, the uniformly low refusal rates observed for Chinese likely reflect under-sensitivity rather than improved detoxification performance. This observation is consistent with findings from Zhang et al. (2024), which show that LLMs often fail to identify unsafe Chinese content due to linguistic characteristics and English-centric safety alignment.

Overall, multilingual false refusal behavior varies substantially across language–dataset pairs. While dataset differences preclude isolating pure language effects, the observed trends are consistent with uneven cross-lingual generalization in English-centered safety alignment.

6.2 Bias Categorization



(a) French (b) German

(c) Spanish (d) Chinese

(e) Korean

Figure 9: HolisticBias Bias Analysis of Multilingual Datasets

We extend the HolisticBias analysis to multilingual datasets using the same methodology as in the English setting.[10] Figure 9 reports mean bias ratios

[10]The bias categorization is presented in the main text, while additional analyses, including toxicity scores, swear words, sentence length, and complexity, are provided in Appendix J.

averaged across representative models for French, German, Spanish, Chinese and Korean.

Across all languages, Political Ideologies consistently exhibit the highest over-refusal rates, indicating that sociopolitical discourse remains a dominant trigger of alignment sensitivity in multilingual detoxification. This pattern suggests the transfer of English-centric moderation heuristics, where politically charged language is frequently associated with unsafe content.

Language-specific differences further emerge. In Chinese, Political Ideologies, Socioeconomic Class, and Sexual Orientation show the highest bias ratios, while Religion exhibits relatively low over-refusal. German displays heightened sensitivity toward Race and Ethnicity, followed by Socioeconomic Class. French shows over-refusal concentrated in Cultural, Characteristics and Political Ideologies, with lower bias for Nationality and Socioeconomic Class. Spanish exhibits a more uniform distribution of bias ratios across categories, though Political Ideologies and Gender and Sex remain slightly elevated, consistent with its higher overall false refusal rates.

Overall, these results indicate that *multilingual bias is neither uniform nor random, but instead reflects language-specific sociocultural sensitivities shaped by model alignment*. The persistent over-refusal of political content underscores the need for multilingual bias calibration.

7 Cross-Translation Framework for False Refusal Mitigation

Our multilingual evaluation shows that, LLMs exhibit substantially lower false refusal rates when detoxifying non-English inputs compared to English inputs. For instance, Chinese inputs yield the lowest refusal ratio of 0.44% with GPT-4o-mini, indicating that LLMs are generally more willing to attempt detoxification in non-English languages. This observation motivates us to propose a simple cross-translation framework for mitigating false refusals in detoxification.[11]

Specifically, we employ Qwen-MT (Qwen Team, 2025) as the translation model and pair it with GPT-4o-mini as the detoxification model, which exhibits particularly low false refusal rates on Chinese inputs. To ensure translation quality, we evaluate Qwen-MT on 200 samples from the Chinese hate speech dataset. A native Chinese speaker who speaks English as a second language manually translates these samples to create reference translations. Qwen-MT achieves a BLEU score of 22.55 and a BERTScore of 0.86 against these references, indicating a reasonable level of semantic preservation.

[11]We provide qualitative evidence for this language-specific effect and compare alternative mitigation strategies in Appendix E.

Metric	Original	Cross-Translation
False Refusal Ratio	11.78%	**1.09%**
Toxicity Score	0.7220	0.6053
Swear Words	17.71%	16.60%
Style Accuracy	0.98	0.95
Content Preservation	0.87	0.87
Fluency	0.92	0.95

Table 1: Effect of the cross-translation mitigation framework on HateXplain with GPT-4o-mini.

We apply the cross-translation framework to HateXplain, with results reported in Table 1. Cross-translation substantially reduces the false refusal ratio. Toxicity scores are computed on the original input texts corresponding to false refusals after mitigation, and the reduced average toxicity indicates that cross-translation enables detoxification for higher-toxicity inputs that would otherwise trigger oversensitive refusal. Swear word ratios remain close to the dataset's original values. Following prior detoxification evaluation (Yuan et al., 2025b), we further assess style accuracy, content preservation, and fluency; the cross-translated texts remain close to the original inputs across all three dimensions. These findings motivate further investigation into cross-translation for mitigating false refusals in multilingual hate speech detoxification.

8 Conclusion

In this work, we systematically study false refusal behavior in hate speech detoxification across English and multilingual datasets. Our analysis reveals that LLMs disproportionately refuse inputs with higher semantic toxicity and content targeting specific sociopolitical categories, particularly nationality, religion, and political ideologies. Moreover, we propose a simple cross-translation mitigation framework that significantly reduces false refusal rates while largely preserving semantic content and toxicity characteristics. Overall, this work provides a foundation for future research on fair, robust, and multilingual detoxification systems that balance safety with utility.

Limitations

We acknowledge several limitations in this work. First, we adopt a deliberately simple and fixed prompt for the hate speech detoxification task in order to observe the models' inherent behavior. While this design choice improves comparability across LLMs, alternative prompt formulations, prompt engineering strategies, or decoding parameter settings, such as temperature, may lead to different detoxification and refusal behaviors.

Second, our study is limited to six languages, namely English, French, German, Spanish, Chinese, and Korean. As a result, languages beyond this set, particularly low-resource languages and those with distinct linguistic or cultural norms, are not covered. The observed cross-lingual patterns may therefore not generalize to all languages.

Third, our bias analysis relies on automatic annotation using an LLM-as-a-Judge paradigm, with limited human verification. Although we conduct human evaluation on a subset of samples and observe reasonable agreement, annotation errors or model-specific biases in the judge model may still affect the results. Future work could incorporate larger-scale human annotation or alternative bias measurement frameworks to further strengthen the conclusions.

Ethical Considerations

This work analyzes false refusal behavior in LLM-based hate speech detoxification and therefore involves toxic and offensive language. All such content is used solely for research purposes to study model behavior and improve safety and fairness; no harmful content is generated beyond what is required for evaluation.

Our analysis of bias using the HolisticBias taxonomy is descriptive rather than normative and is intended to identify uneven refusal behavior across social categories, not to make judgments about any group. While our proposed cross-translation strategy reduces false refusals, it may also affect the detection of unsafe content in certain languages and should be applied with caution in real-world systems.

In deployed moderation pipelines, false refusal is also a utility failure rather than only an annotation error. If a system is designed to rewrite rather than delete harmful wording, a false refusal can impose a hard block on content that could otherwise be safely reconstructed, increasing over-censorship and shifting additional burden to human moderators (Habibi et al., 2024). We therefore interpret false-refusal rate as a downstream deployment risk that can affect both moderation outcomes and user trust.

Use of AI Assistants. We used AI-based writing assistance tools for minor language editing and proofreading. All scientific content, analyses, and conclusions were produced and verified by the authors.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Microsoft Azure AI Foundry Blog. 2024. Introducing phi-4: Microsoft's newest small language model specializing in completion. Accessed: 2026-01-05.

Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. MultiParaDetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 124–140, Mexico City, Mexico. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, Stevie A Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, Verena Rieser, and 1 others. 2022. Safetykit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Farshid Faal, Ketra Schmitt, and Jia Yuan Yu. 2023. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53(7):8421–8435.

Xingli Fang, Jianwei Li, Varun Mulchandani, and Jung-Eun Kim. 2025. Trustworthy ai: Safety, bias, and privacy–a survey. *arXiv preprint arXiv:2502.10450*.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2024. MART: Improving LLM safety with multi-round automatic red-teaming. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1927–1937, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mahyar Habibi, Dirk Hovy, and Carlo Schwarz. 2024. The content moderator's dilemma: Removal of toxic content and distortions to online discourse. *arXiv preprint arXiv:2412.16114*.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024. Detoxllm: A framework for detoxification with explanations. *arXiv preprint arXiv:2402.15951*.

Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2025. DefVerify: Do hate speech models reflect their dataset's definition? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4341–4358, Abu Dhabi, UAE. Association for Computational Linguistics.

Gretel Liz De la Pena Sarracen and Paolo Rosso. 2023. Systematic keyword and bias analyses in hate speech detection. *Information Processing & Management*, 60(5):103433.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Is llm an overconfident judge? unveiling the capabilities of llms in detecting offensive language with annotation disagreement. *arXiv preprint arXiv:2502.06207*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Cross-lingual transfer of debiasing and detoxification in multilingual llms: An extensive investigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2805–2830.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with

human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Chaewon Park, Soohwan Kim, Kyubyong Park, and Kunwoo Park. 2023. K-HATERS: A hate speech detection corpus in Korean with target-specific ratings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14264–14278, Singapore. Association for Computational Linguistics.

Qwen Team. 2025. Qwen-mt: Where speed meets smart translation. https://qwenlm.github.io/blog/qwen-mt/. Accessed: 2026-01-04.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, and 1 others. 2024. Large language model safety: A holistic survey. *arXiv preprint arXiv:2412.17686*.

Shengyun Si, Xinpeng Wang, Guangyao Zhai, Nassir Navab, and Barbara Plank. 2025. Think before refusal: Triggering safety reflection in llms to mitigate false refusal behavior. *arXiv preprint arXiv:2503.17882*.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024a. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: how does llm safety training fail? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zachary Yang, Domenico Tullo, and Reihaneh Rabbany. 2025b. Unified game moderation: Soft-prompting and llm-assisted label transfer for resource-efficient toxicity detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5161–5170, New York, NY, USA. Association for Computing Machinery.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ computer science*, 7:e598.

Jing Yu, Yibo Zhao, Jiapeng Zhu, Wenming Shao, Bo Pang, Zhao Zhang, and Xiang Li. 2025. Text detoxification: Data efficiency, semantic preservation and model generalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32160–32174, Suzhou, China. Association for Computational Linguistics.

Shuzhou Yuan, William LaCroix, Hardik Ghoshal, Ercong Nie, and Michael Färber. 2025a. Codae: Adapting large language models for education via chain-of-thought data augmentation. *arXiv preprint arXiv:2508.08386*.

Shuzhou Yuan, Antonis Maronikolakis, and Hinrich Schütze. 2022. Separating hate speech and offensive language classes via adversarial debiasing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 1–10, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Shuzhou Yuan, Ercong Nie, Lukas Kouba, Helmut Schmid, Hinrich Schuetze, and Michael Färber. 2025b. LLM in the loop: Creating the ParaDeHate dataset for hate speech detoxification. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1014–1027, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Shuzhou Yuan, Ercong Nie, Yinuo Sun, Chenxuan Zhao, William LaCroix, and Michael Färber. 2025c. Beyond over-refusal: Scenario-based diagnostics and post-hoc mitigation for exaggerated refusals in llms. *arXiv preprint arXiv:2510.08158*.

Shuzhou Yuan, Ercong Nie, Mario Tawfelis, Helmut Schmid, Hinrich Schütze, and Michael Färber. 2025d. Hateful person or hateful model? investigating the role of personas in hate speech detection by large language models. *arXiv preprint arXiv:2506.08593*.

Hengxiang Zhang, Hongfu Gao, Qiang Hu, Guanhua Chen, Lili Yang, Bingyi Jing, Hongxin Wei, Bing Wang, Haifeng Bai, and Lei Yang. 2024. Chinesesafe: A chinese benchmark for evaluating safety in large language models. *arXiv preprint arXiv:2410.18491*.

A Dataset

Table 2 summarizes the datasets used in our experiments.

Dataset / Language	Size
English	
Davidson (Davidson et al., 2017)	∼20K
HateXplain (Mathew et al., 2021)	∼12K
ParaDetox (Logacheva et al., 2022)	∼15K
Multilingual	
French (Tonneau et al., 2024)	∼4K
Spanish (Tonneau et al., 2024)	∼7K
German (Tonneau et al., 2024)	∼6K
Chinese (Yang et al., 2025b)	∼18K
Korean (Park et al., 2023)	∼140K

Table 2: Datasets used in our study. "Size" denotes the number of instances used in the experiments.

B Models

The details of the LLMs, including their sizes and multilingual coverage, are presented in Table 3.

Model	Size	Multilingual Coverage
Closed Source		
GPT-4o mini	-	50+ languages
GPT-3.5 Turbo	-	50+ languages
Open Source		
Llama-3.1	8B	30+ languages
Mistral	8B	Limited multilingual support
Mixtral	8×7B	Limited multilingual support
Qwen-2.5	7B	29+ languages
Qwen-3	30B	29+ languages
Gemma-2	9B	Limited multilingual support
Gemma-3	27B	Limited multilingual support

Table 3: Language models used in this study, with their sizes and approximate multilingual coverage.

C Inter-Annotator Agreement and Judge Robustness

To assess Phi-4's reliability as an automated annotator, we conduct a cross-annotation study using 200 detoxified outputs from the English HateXplain subset: 100 examples flagged by Phi-4 as potential false refusals and 100 non–false-refusal cases. The validation set is designed to verify the usability of LLM judges for contextual false-refusal decisions rather than to exhaustively cover all possible linguistic variations. Linguistic factors such as token count and sentence length are not annotated because they are directly measurable from the text.

Two human annotators label the validation set. Both annotators are graduate students in computer science and speak English as a second language. Before labeling, they review guideline examples that illustrate false refusals and non-false-refusal outputs in the detoxification setting. They then independently label the 200 randomly ordered examples according to whether the model output unnecessarily refuses, gives moralizing or judgmental advice instead of detoxifying, or otherwise fails to perform the requested safe rewrite.

To test whether the annotation results depend on a single judge model, we additionally evaluate DeepSeek-R1 on the same 200 examples. As shown in Table 4, DeepSeek-R1 aligns closely with both human annotators and Phi-4. The cross-model agreement between DeepSeek-R1 and Phi-4 is high (Cohen's $\kappa = 0.810$, 90.5% raw agreement), suggesting that false-refusal annotations are not specific to one model family.

D Interpretability Example for Category-Specific Refusals

To provide qualitative interpretability evidence for why some target categories trigger false refusals, we inspect examples in which identity-related expressions dominate the model's refusal decision. For the input "i hate these fagxxts", the model focuses disproportionately on the identity-related

Annotator Pair	Cohen's κ	Agreement (%)
Human baselines		
Annotator A vs Annotator B	0.878	94.5
Phi-4 original judge		
Phi-4 vs Annotator A	0.670	83.5
Phi-4 vs Annotator B	0.700	85.0
Phi-4 vs Human Consensus	0.740	87.0
DeepSeek-R1 additional judge		
DeepSeek-R1 vs Annotator A	0.690	84.5
DeepSeek-R1 vs Annotator B	0.720	86.0
DeepSeek-R1 vs Human Consensus	0.760	88.0
Cross-model robustness		
DeepSeek-R1 vs Phi-4	0.810	90.5

Table 4: Inter-annotator agreement between human annotators, Phi-4, and DeepSeek-R1 on the stratified 200-sample evaluation set.

token "fagxxts" compared with surrounding tokens such as "i", "hate", and "these". This suggests that the refusal decision can be driven by heightened sensitivity to the identity token itself rather than by a holistic assessment of the full detoxification request.

Token	Relative attention saliency
i	Low
hate	Medium
these	Low
fagxxts	High

Table 5: Qualitative attention-saliency summary for an identity-related false-refusal example. The slur is obfuscated to reduce harm while preserving the observed trigger pattern.

E Additional Analysis of Cross-Translation

To better understand why Chinese inputs produce lower false-refusal rates, we conduct a qualitative analysis of multilingual detoxification examples. We find that refusals are often triggered by high-saliency English tokens or phrases, such as explicit negative predicates, insults, and identity terms. In Chinese translations, the same concepts may be distributed across multiple characters, combined with grammatical particles, or expressed with different word-order patterns. This can reduce the prominence of individual high-risk tokens and weaken the match with English-centric safety triggers.

Table 6 illustrates this pattern. The English prompt contains salient tokens such as "hate", "incompetent", and the target nationality, which can jointly trigger refusal. In the Chinese version, these meanings are represented through a different segmentation and syntax, dispersing saliency across the sentence. Similar patterns appear in other identity-related prompts, suggesting that tokenization granularity, word order, and semantic dispersion can affect refusal behavior. Cross-translation leverages this effect by re-expressing English inputs in Chinese before detoxification, reducing false refusals caused by surface-level English safety triggers.

Language	Prompt	Refusal
English	"I hate these incompetent [nationality] doctors."	Yes
Chinese	"我讨厌这些无能的[国籍]医生。"	No

Table 6: Qualitative example showing how high-saliency English triggers can become less prominent after Chinese translation.

We also compare cross-translation with two alternative mitigation strategies on HateXplain using Llama-3.1 8B. The first is reasoning-based mitigation, following Think-Before-Refusal, which encourages the model to reflect before refusing (Si et al., 2025). The second is post-hoc explanation-based mitigation, which identifies influential tokens and reduces their attention scores during inference (Yuan et al., 2025c). Because the latter requires direct attention-score modification, it is only applicable to open-source models. As shown in Table 7, cross-translation achieves the lowest false-refusal rate, indicating that it is both effective and comparatively easy to apply without modifying model internals.

Method	False Refusal Rate (%)
None	14.42
Rethinking	10.21
Post-hoc	3.64
Cross-translation (ours)	**2.57**

Table 7: Comparison of mitigation strategies on HateXplain using Llama-3.1 8B.

E.1 Human Evaluation of Cross-Translation

In addition to automatic metrics, we manually evaluate 50 cross-translated HateXplain instances to estimate whether the mitigation still produces false

refusals. As shown in Table 8, the human-evaluated false-refusal rate remains low and is close to the LLM-as-a-judge estimate, supporting the robustness of the mitigation result.

Metric	Cross-Translation
Human-evaluated False Refusal Rate (%)	2.01
LLM-as-a-judge False Refusal Rate (%)	1.09

Table 8: Human evaluation of false refusals on 50 cross-translated HateXplain instances.

E.2 Unsafe-Prompt Check in Chinese

Although our detoxification task uses safe instructions with sensitive content, very low refusal rates in Chinese may also indicate under-detection of genuinely unsafe prompts. To contextualize this risk, we translate 20 unsafe prompts from XSTest (Röttger et al., 2024) into Chinese and evaluate GPT-4o-mini. Table 9 shows that refusal remains high in Chinese, but lower than in English. This supports our ethical caution that cross-translation reduces false refusals but may also weaken unsafe-content detection in some settings.

Language	Refusal Rate (%)
English	100.0
Chinese	87.12

Table 9: Refusal rates of GPT-4o-mini on 20 unsafe XSTest prompts and their manual Chinese translations.

F HolisticBias Structure

HolisticBias dataset consists of 13 demographic axis and buckets. Figure 10 shows the overview of HolisticBias.

G HolisticBias Categorization of English Datasets

Figure 11 presents the HolisticBias categorization of the original English datasets, illustrating the distribution of samples across each category.

H Mean Bias Ratio of False Refusals

The mean bias ratio of false refusals for each HolisticBias category, aggregated across the three English datasets, is reported in Table 10.

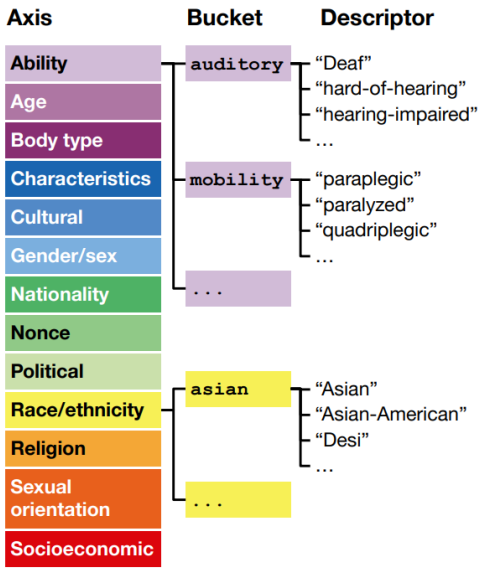


Figure 10: Structure of the HolisticBias dataset adapted from Smith et al. (2022). It consists of 13 demographic pillars, each containing multiple property groups representing diverse social and identity dimensions.

I Linguistic Analysis of False Refusals in English Datasets

This section presents additional linguistic analyses of false refusals in English datasets that are omitted from the main body due to space constraints. We analyze three linguistic factors: **token length**, **clause count**, and **parse tree depth**. The corresponding distributions across different model families are illustrated as follows.

Token Length. The token length distributions of false refusals for GPT, Llama, Mistral, Qwen, and Gemma models are shown in Figures 12–16.

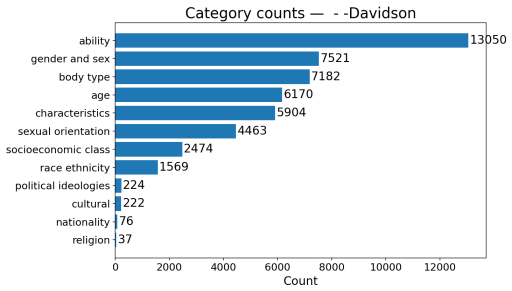
Clause Count. Clause count distributions for the same set of models are presented in Figures 17–21.

Parse Tree Depth. Figures 22–26 report the distributions of parse tree depth, capturing syntactic complexity across false refusals.

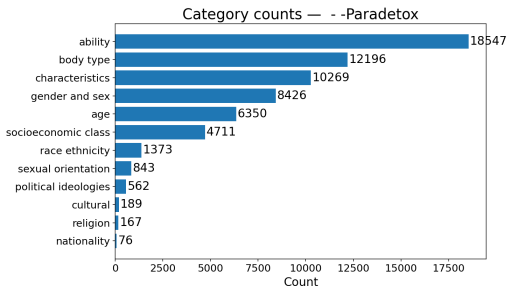
Across all three linguistic dimensions, the distributions of falsely refused samples closely overlap with those of the original datasets. These results indicate that token length, clause count, and syntactic complexity do not exhibit consistent or systematic patterns that trigger false refusal behavior in LLM-based hate speech detoxification.

Category	GPT 3.5	GPT4o mini	Mistral 7B	Mixtral 8×7B	Llama3 8B	Qwen2.5 7B	Qwen3 30B	Gemma2 9B	Gemma3 27B
Nationality	**1.83**	**1.67**	**1.94**	1.24	1.22	**1.76**	1.37	**1.87**	1.48
Religion	1.61	1.46	1.81	1.54	1.15	1.29	1.33	1.43	**1.78**
Cultural	1.13	1.22	1.20	**1.58**	1.07	1.04	1.09	1.11	0.91
Race ethnicity	1.19	1.22	1.18	0.99	1.24	1.23	1.06	1.32	0.86
Sexual orientation	1.36	1.43	1.21	<u>0.70</u>	**1.39**	1.33	0.99	1.39	0.98
Political ideologies	1.15	1.34	1.42	1.45	1.11	1.12	**1.48**	1.26	1.52
Socioeconomic class	1.31	1.33	1.17	1.22	1.25	1.23	1.14	1.20	1.64
Characteristics	1.02	1.04	1.02	1.04	1.01	1.00	0.91	1.00	0.91
Body type	0.91	0.89	0.97	1.05	0.94	0.93	0.99	0.95	0.90
Ability	0.99	0.96	0.99	1.01	1.00	1.00	0.95	0.98	1.01
Age	1.02	1.03	0.92	1.03	1.00	1.03	0.94	1.01	0.98
Gender and sex	<u>0.79</u>	<u>0.76</u>	<u>0.79</u>	0.73	<u>0.79</u>	<u>0.81</u>	<u>0.90</u>	<u>0.76</u>	<u>0.77</u>

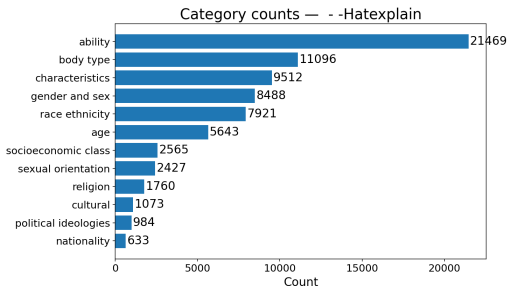
Table 10: Mean bias ratio (R_c) per HolisticBias category across models for English datasets. Values greater than 1 indicate overrepresentation of the category in false refusals. The largest value in each column is shown in **bold**, and the smallest value is <u>underlined</u>.



(a) Davidson



(b) Paradetox



(c) HateXplain

Figure 11: HolisticBias categorization of English datasets.

J Bias Analysis on Multilingual Datasets

J.1 Toxicity Level

As Detoxify (Hanu and Unitary team, 2020) doesn't support all multilingual datasets, Figure 27 shows the toxicity score in French dataset. It suggests that false refusals in French are associated with higher toxicity scores than the average in original dataset.

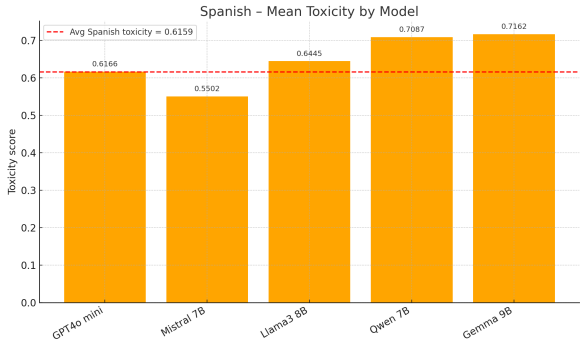


Figure 27: Toxicity Score of French Dataset.

J.2 Swear Word Presence

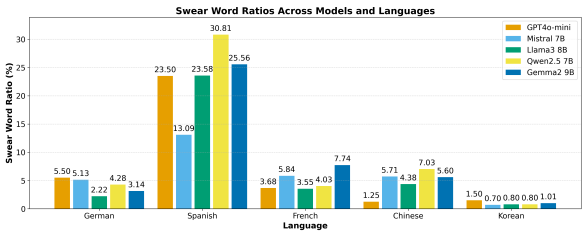
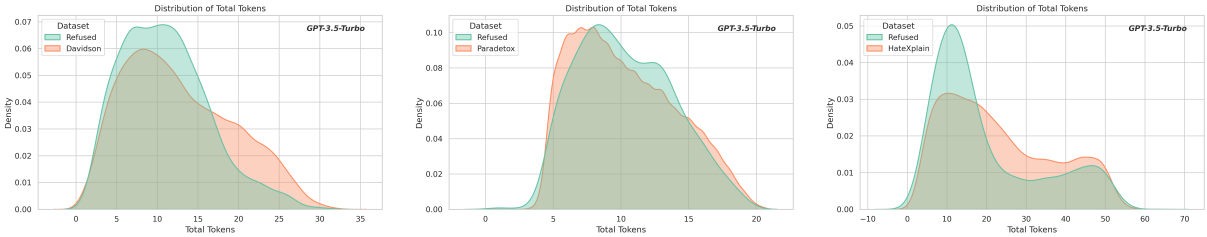
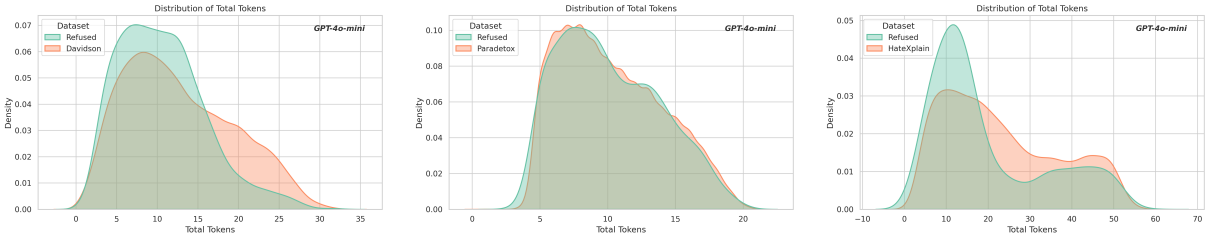


Figure 28: Swear Words Ratio of Multilingual Datasets.

Swear-word ratios for multilingual datasets are shown in Figure 28. Spanish exhibits the highest ratios across models, exceeding 30% for Qwen2.5 7B and 25% for Gemma2 9B, aligning with the elevated false refusal rates observed in Figure 8.

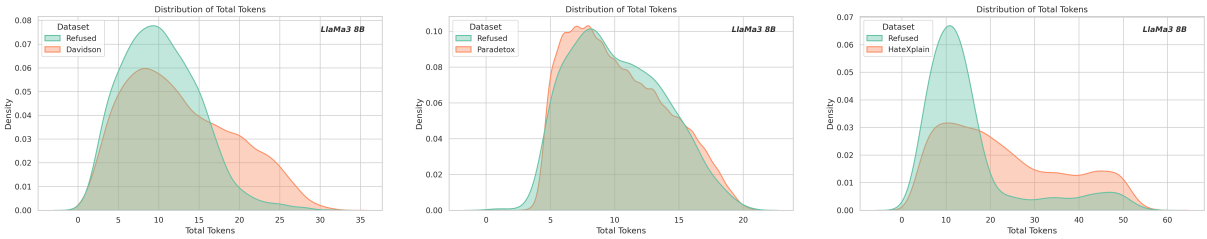


(a) Davidson (b) Paradetox (c) HateXplain



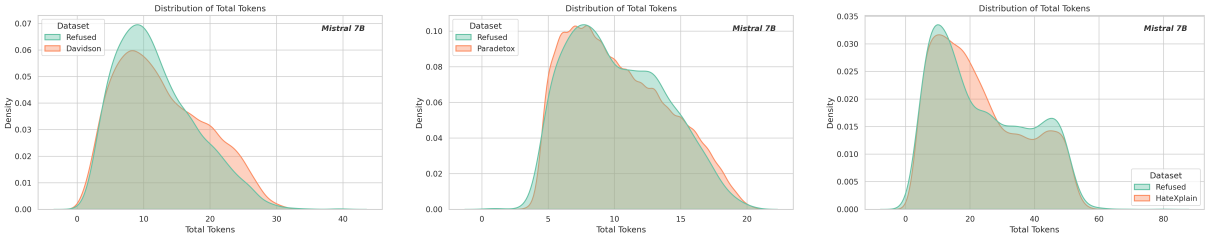
(d) Davidson (e) Paradetox (f) HateXplain

Figure 12: Token count distributions across datasets for GPT-3.5-Turbo and GPT-4o-mini.

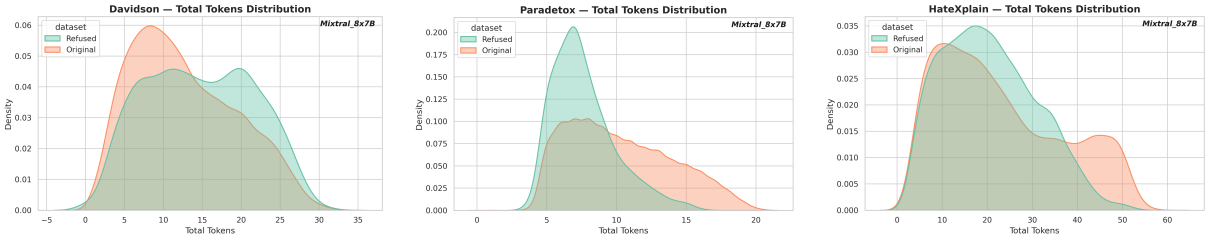


(a) Davidson (b) Paradetox (c) HateXplain

Figure 13: Token count distributions across datasets for Llama3 8B.

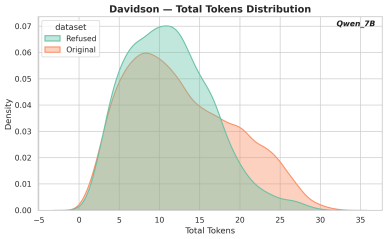


(a) Davidson (b) Paradetox (c) HateXplain

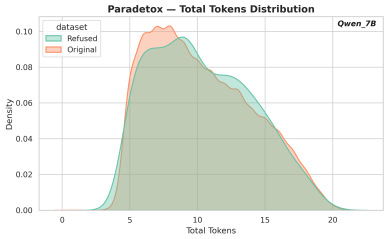


(d) Davidson (e) Parardetox (f) HateXplain

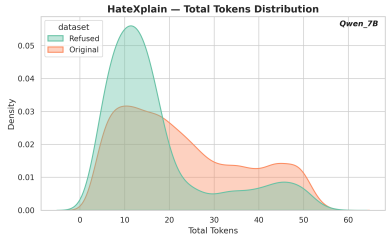
Figure 14: Token count distributions across datasets for Mistral 7B and Mixtral 8×7B.



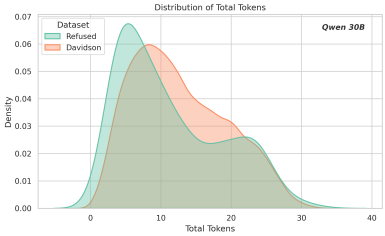
(a) Davidson



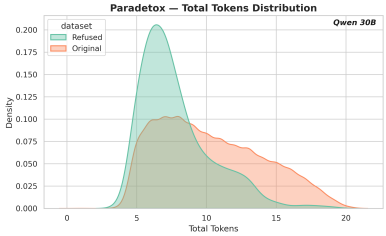
(b) Paradetox



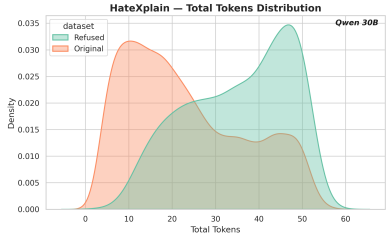
(c) HateXplain



(d) Davidson

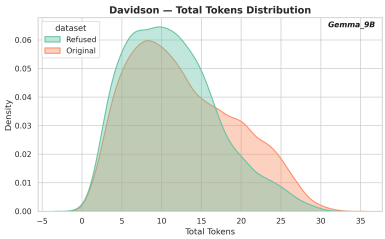


(e) Paradetox

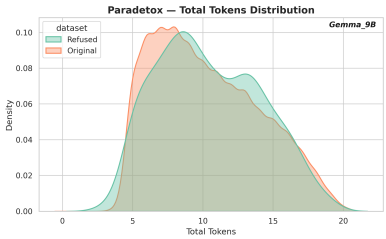


(f) HateXplain

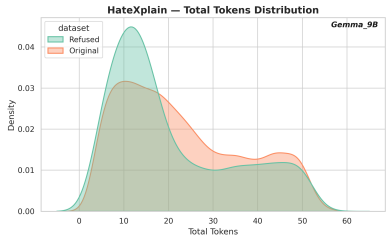
Figure 15: Token count distributions across datasets for Qwen2.5 7B and Qwen3 30B.



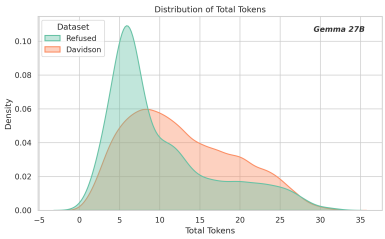
(a) Davidson



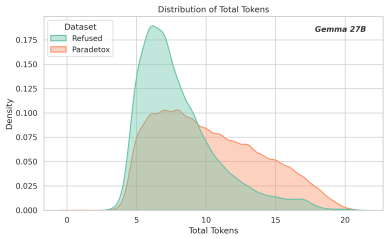
(b) Paradetox



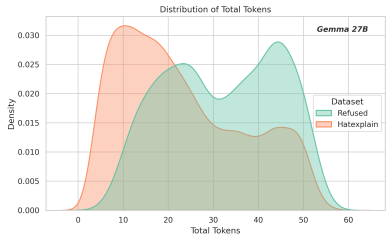
(c) HateXplain



(d) Davidson

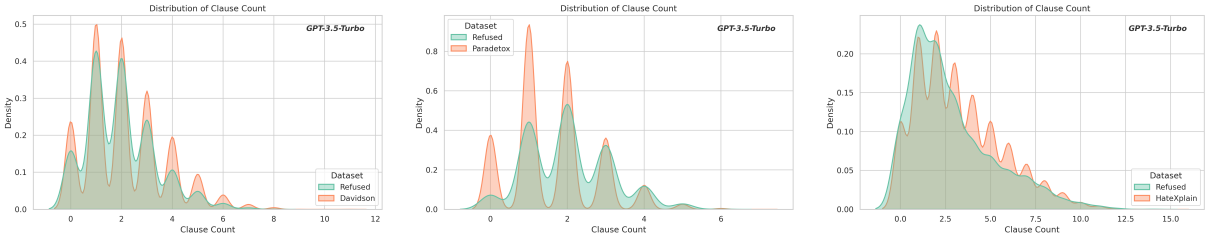


(e) Paradetox

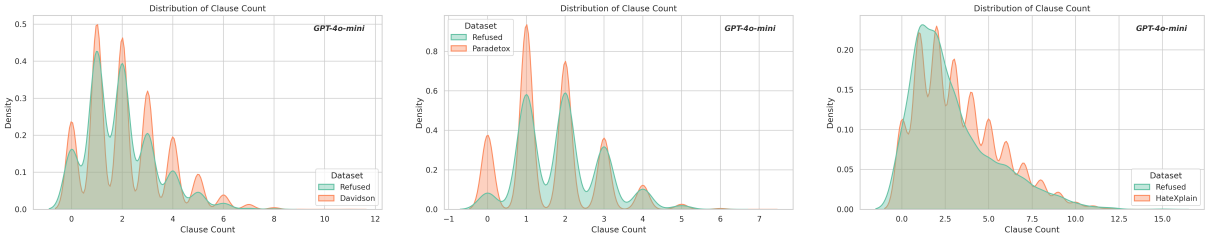


(f) HateXplain

Figure 16: Token count distributions across datasets for Gemma2 9B and Gemma3 27B.

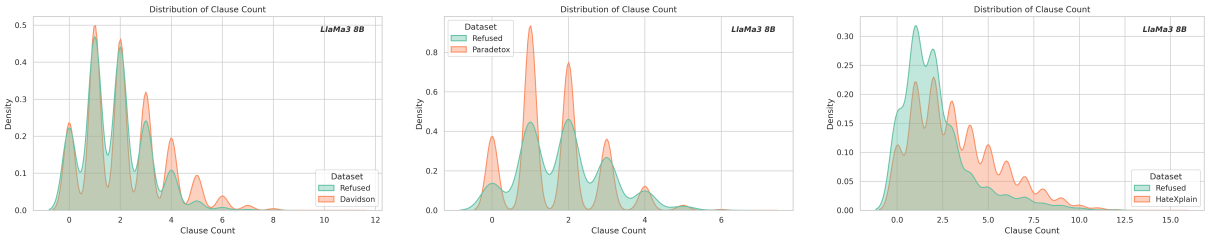


(a) Davidson (b) Paradetox (c) HateXplain



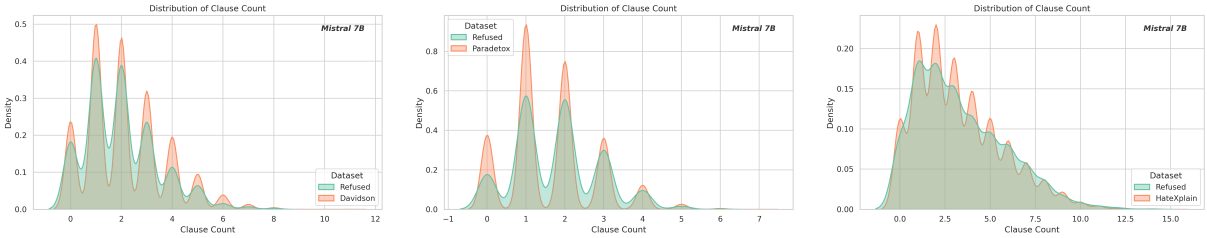
(d) Davidson (e) Paradetox (f) HateXplain

Figure 17: Clause count distributions across datasets for GPT-3.5-Turbo and GPT-4o-mini.

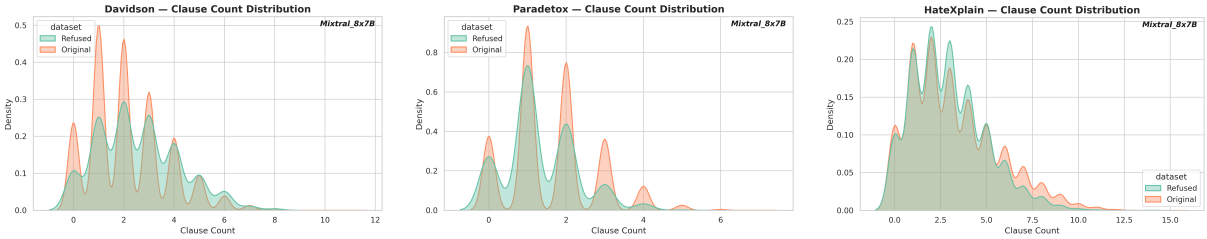


(a) Davidson (b) Paradetox (c) HateXplain

Figure 18: Clause count distributions across datasets for Llama3 8B.

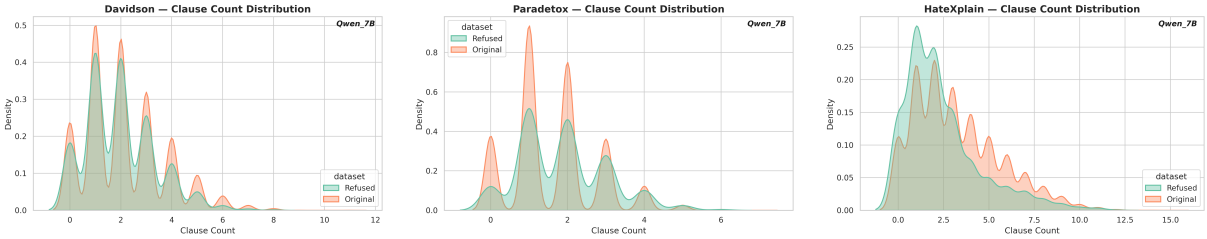


(a) Davidson (b) Paradetox (c) HateXplain

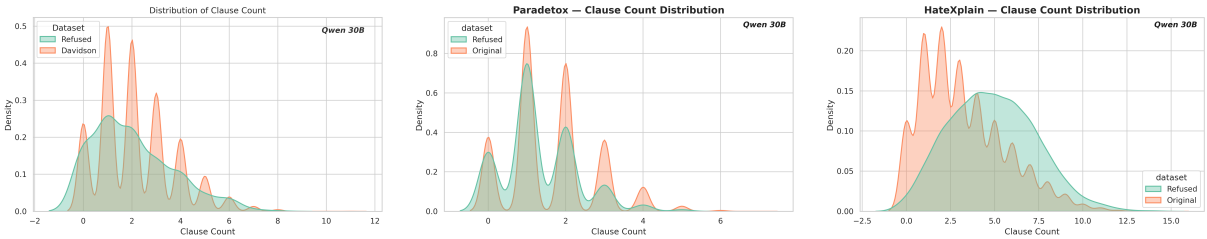


(d) Davidson (e) Parardetox (f) HateXplain

Figure 19: Clause count distributions across datasets for Mistral 7B and Mixtral 8×7B.

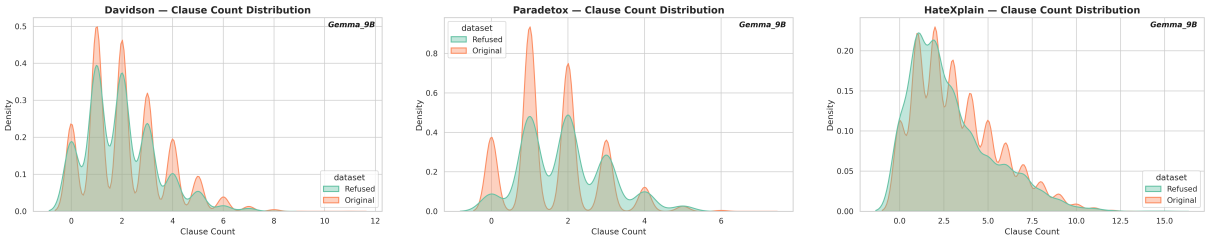


(a) Davidson (b) Paradetox (c) HateXplain

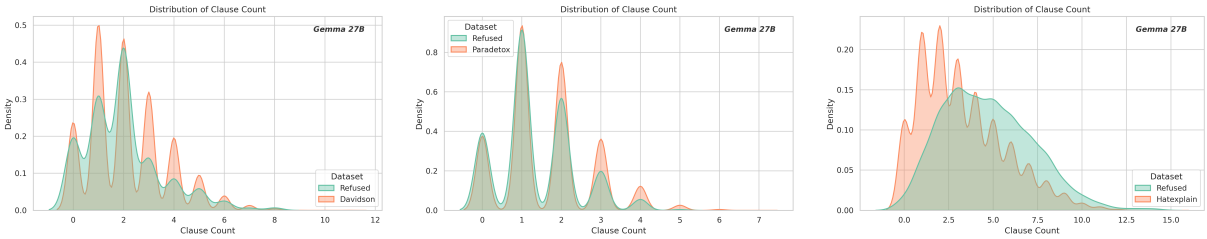


(d) Davidson (e) Paradetox (f) HateXplain

Figure 20: Clause count distributions across datasets for Qwen2.5 7B and Qwen3 30B.

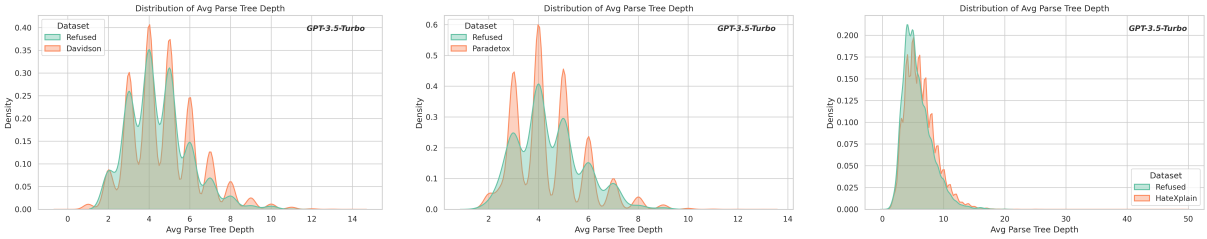


(a) Davidson (b) Paradetox (c) HateXplain

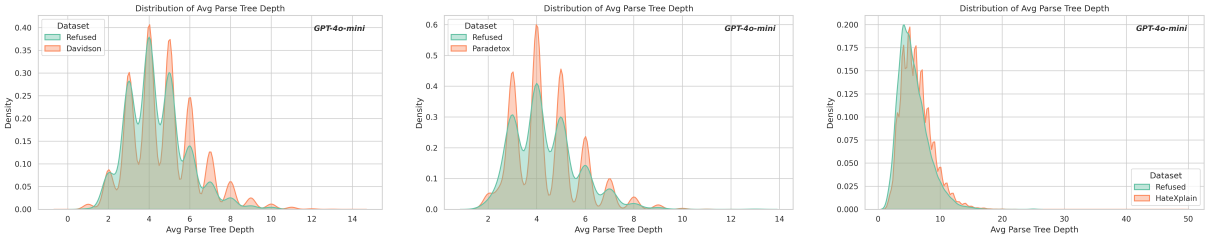


(d) Davidson (e) Paradetox (f) HateXplain

Figure 21: Clause count distributions across datasets for Gemma2 9B and Gemma3 27B.

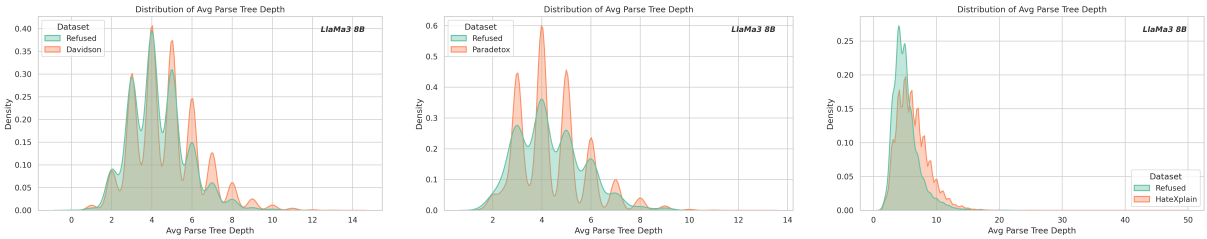


(a) Davidson (b) Paradetox (c) HateXplain



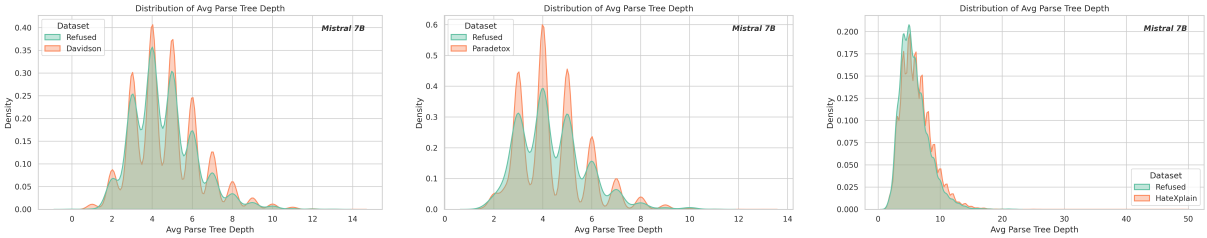
(d) Davidson (e) Paradetox (f) HateXplain

Figure 22: Parse tree depth distributions across datasets for GPT-3.5-Turbo and GPT-4o-mini.

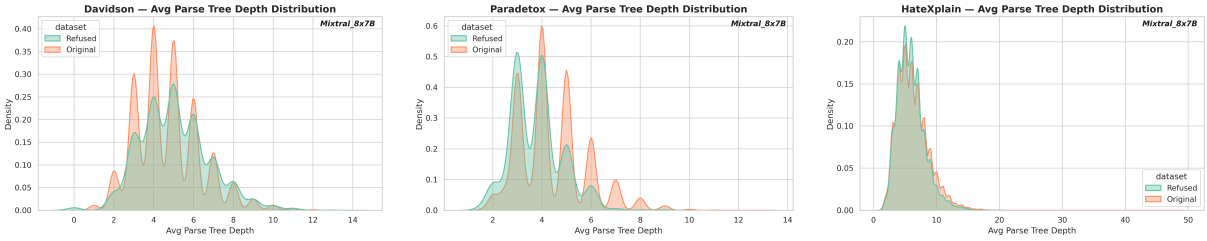


(a) Davidson (b) Paradetox (c) HateXplain

Figure 23: Parse tree depth distributions across datasets for Llama3 8B.

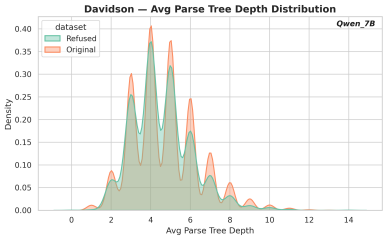


(a) Davidson (b) Paradetox (c) HateXplain

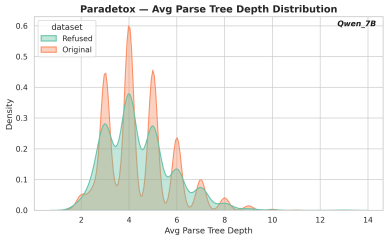


(d) Davidson (e) Paradetox (f) HateXplain

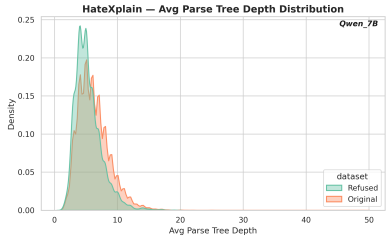
Figure 24: Parse tree depth distributions across datasets for Mistral 7B and Mixtral 8×7B.



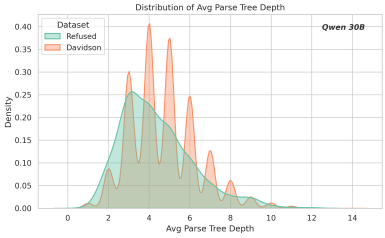
(a) Davidson



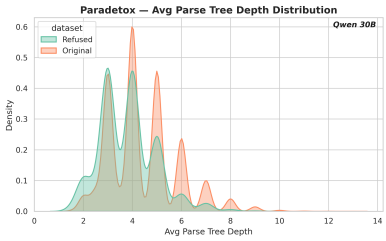
(b) Paradetox



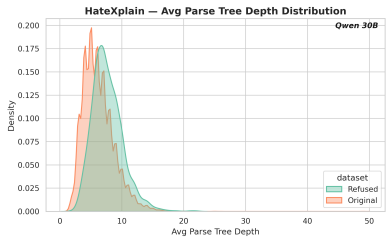
(c) HateXplain



(d) Davidson

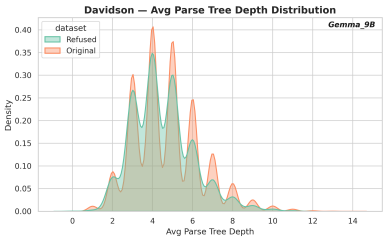


(e) Paradetox

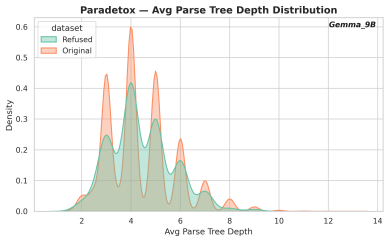


(f) HateXplain

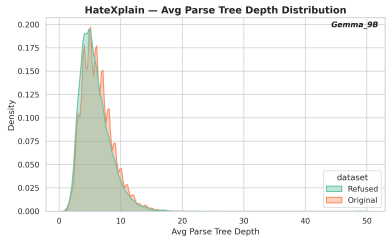
Figure 25: Parse tree depth distributions across datasets for Qwen2.5 7B and Qwen3 30B.



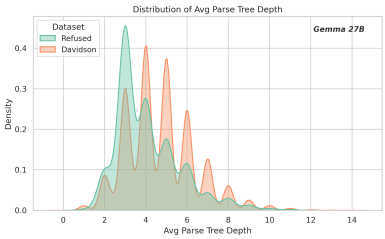
(a) Davidson



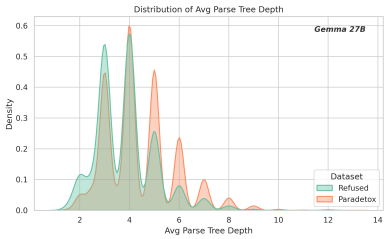
(b) Paradetox



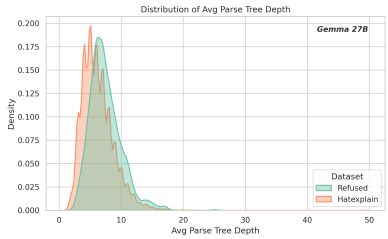
(c) HateXplain



(d) Davidson



(e) Paradetox



(f) HateXplain

Figure 26: Parse tree depth distributions across datasets for Gemma2 9B and Gemma3 27B.

This suggests that Spanish false refusals are largely driven by oversensitivity to direct lexical cues. In contrast, French shows substantially lower swear-word ratios (3.68–7.74%), indicating that false refusals are more influenced by semantic misinterpretation than explicit profanity. German, Korean, and Chinese exhibit consistently low ratios (2–7%), suggesting that false refusals in these languages are not primarily triggered by profanity but by broader semantic or structural mismatches with English-centric safety alignment. Overall, these results highlight the need for multilingual detoxification mechanisms that jointly calibrate lexical and semantic sensitivity.

J.3 Sentence Length

Figure 29 presents the token-length distributions of false refusals and original samples for Spanish and Chinese across all evaluated models. The corresponding distributions for French, German, and Korean are shown in Figures 30, 31, and 32, respectively.

Across all languages, the token-length distributions of falsely refused samples largely overlap with those of the original datasets, indicating that sentence length and surface-level verbosity do not systematically trigger refusal behavior. In Spanish, despite substantially higher false-refusal ratios, only minor shifts toward shorter inputs are observed for certain models, such as Gemma2 9B and Mistral 7B. This suggests that over-refusals are more likely driven by semantic or cultural factors, including informal registers or idiomatic expressions, rather than lexical complexity.

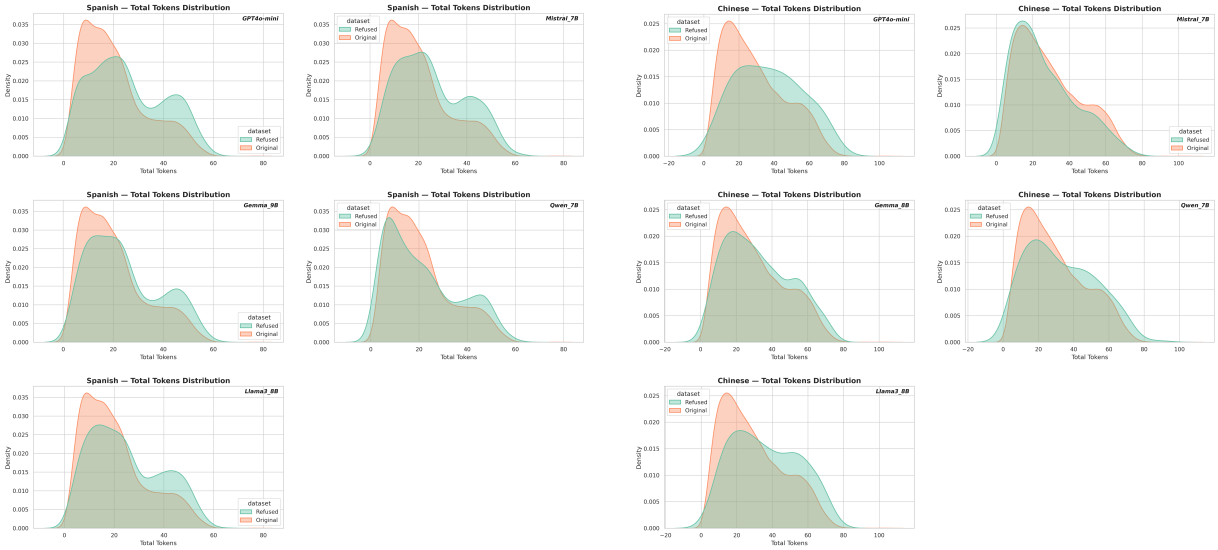
By contrast, Chinese exhibits similarly overlapping length distributions alongside consistently low refusal rates, implying under-sensitivity to toxic content despite comparable input lengths. Taken together, these results indicate that multilingual refusal behavior is shaped primarily by semantic and alignment-related factors, rather than by token-level properties.

J.4 Sentence Complexity

Figure 33 presents the clause-count distributions for Spanish and Chinese across all evaluated models. Corresponding distributions for French, German, and Korean are shown in Figures 34, 35, and 36, respectively. Parse tree depth distributions are shown in Figures 37, 38, and 39.

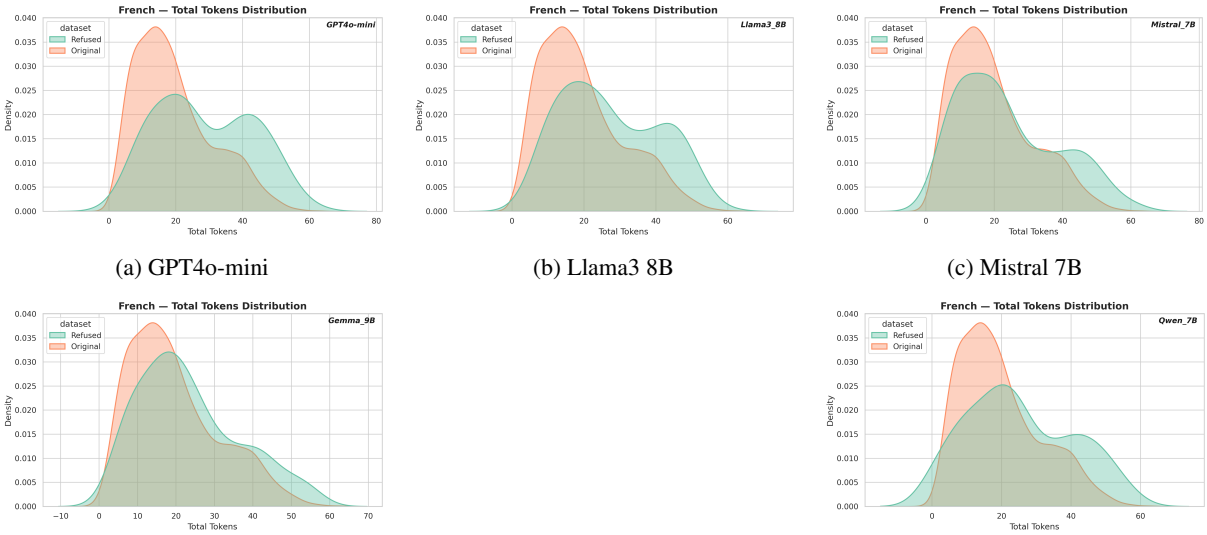
Similar to the token-length patterns, the distributions of falsely refused samples largely overlap with those of the original datasets across all languages. Only minor shifts toward simpler structures are observed in a few models. This indicates that clause-level and parse-tree syntactic complexity do not meaningfully contribute to false refusal behavior in multilingual hate speech detoxification.

Despite substantial differences in overall refusal rates between Spanish and Chinese, the comparable distributions of clause counts and parse-tree depths suggest that refusals are primarily driven by semantic and contextual factors, rather than by structural complexity.



(a) **Spanish** (b) **Chinese**

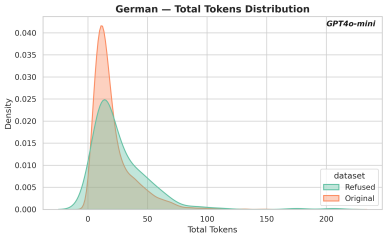
Figure 29: Token count distribution for false refusals versus original samples across representative models in Spanish and Chinese dataset.



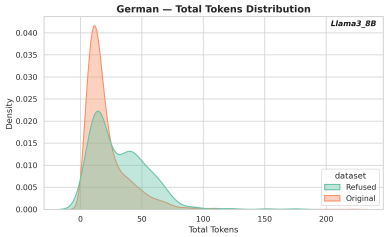
(a) GPT4o-mini (b) Llama3 8B (c) Mistral 7B

(d) Gemma2 9B (e) Qwen2.5 7B

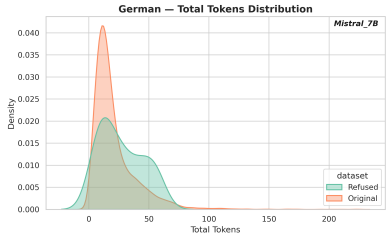
Figure 30: Token count distributions for French dataset.



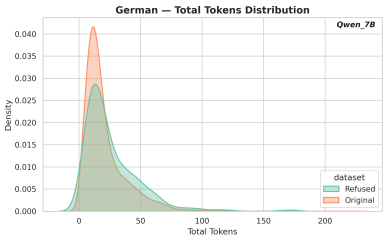
(a) GPT4o-mini



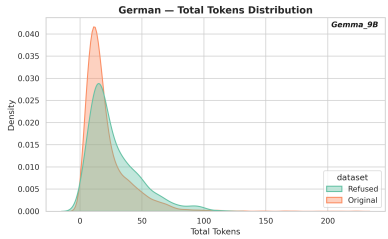
(b) Llama3 8B



(c) Mistral 7B

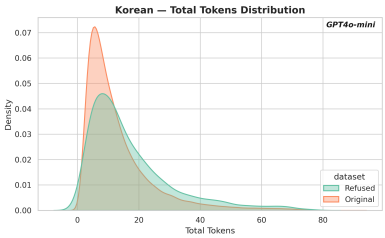


(d) Qwen2.5 7B

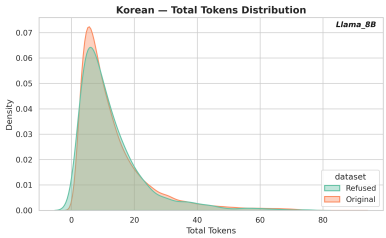


(e) Gemma2 9B

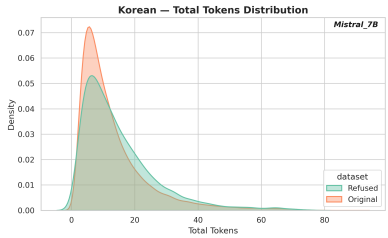
Figure 31: Token count distributions for German dataset.



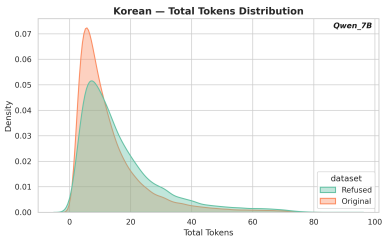
(a) GPT4o-mini



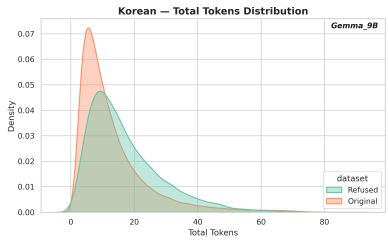
(b) Llama3 8B



(c) Mistral 7B

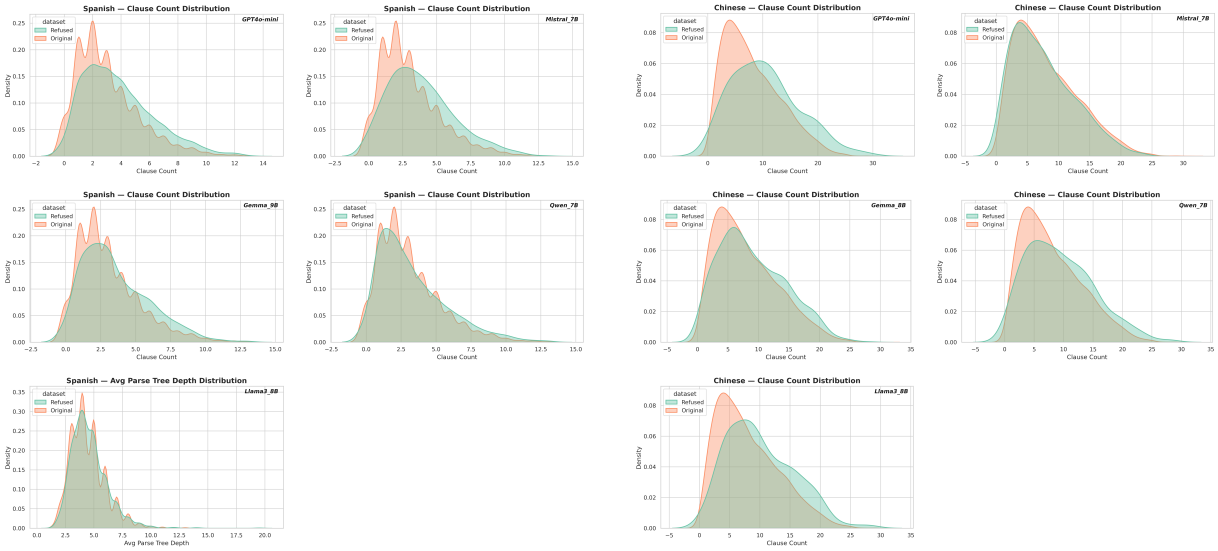


(d) Qwen2.5 7B



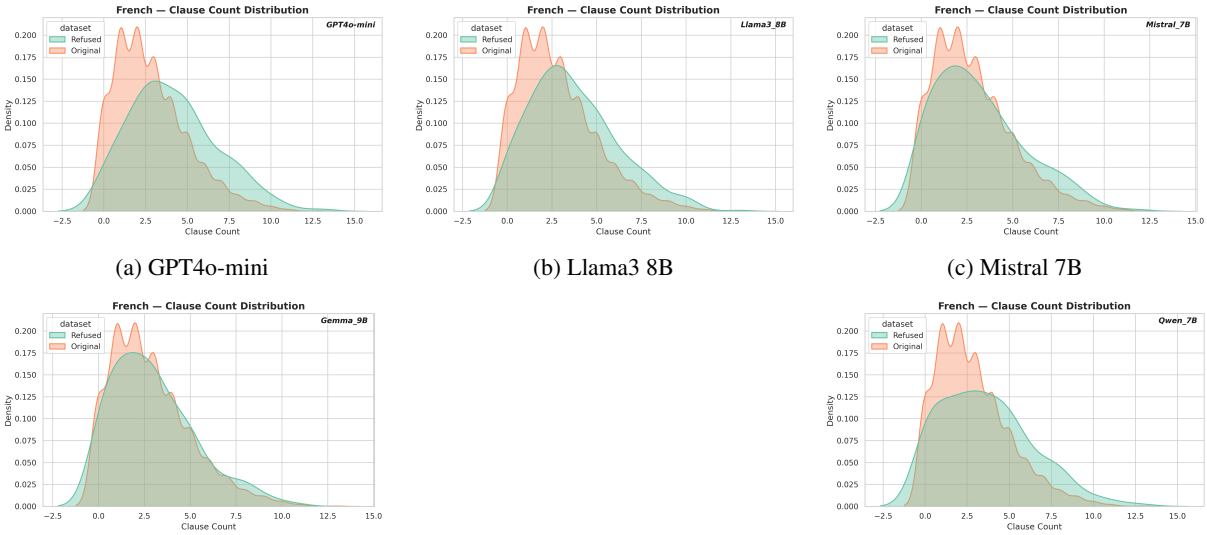
(e) Gemma2 9B

Figure 32: Token count distribution of Korean dataset.



(a) **Spanish** (b) **Chinese**

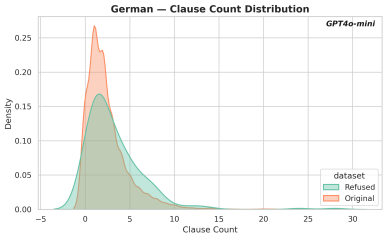
Figure 33: Clause count distribution for false refusals versus original samples across representative models in Spanish and Chinese datasets.



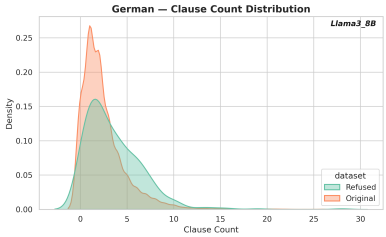
(a) GPT4o-mini (b) Llama3 8B (c) Mistral 7B

(d) Gemma2 9B (e) Qwen2.5 7B

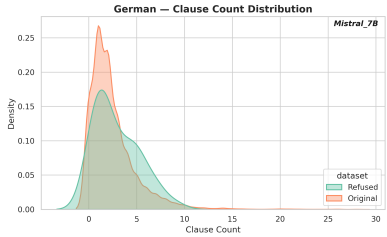
Figure 34: Clause count distributions for French dataset.



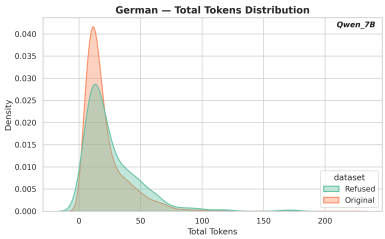
(a) GPT4o-mini



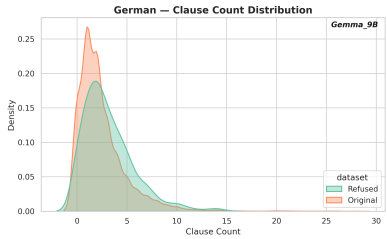
(b) Llama3 8B



(c) Mistral 7B

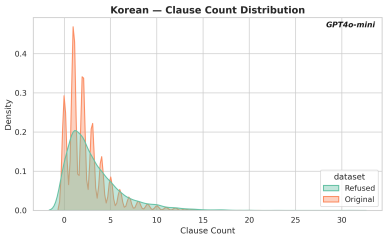


(d) Qwen2.5 7B

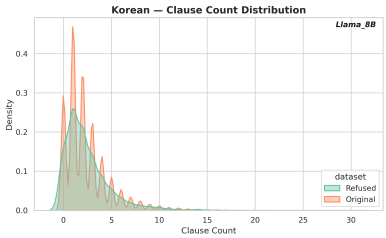


(e) Gemma2 9B

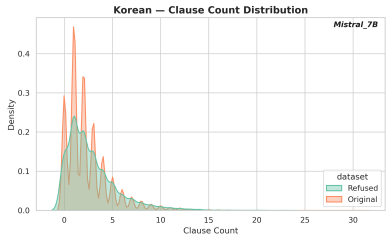
Figure 35: Clause count distributions for German dataset.



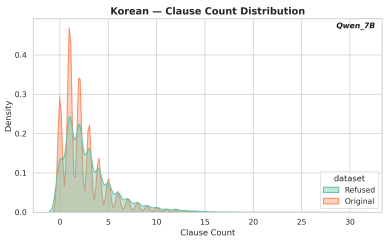
(a) GPT4o-mini



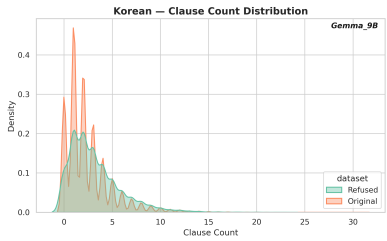
(b) Llama3 8B



(c) Mistral 7B

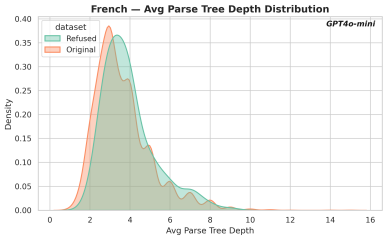


(d) Qwen2.5 7B

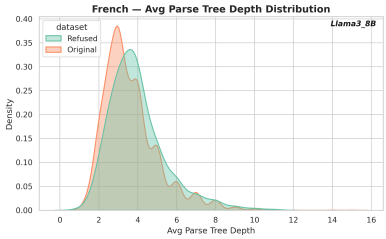


(e) Gemma2 9B

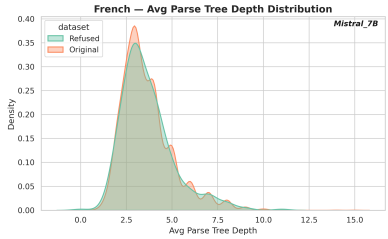
Figure 36: Clause count distribution of the Korean dataset.



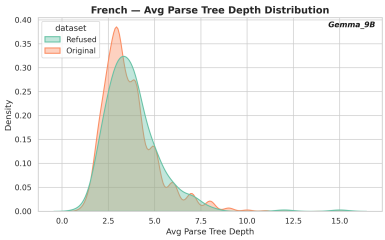
(a) GPT4o-mini



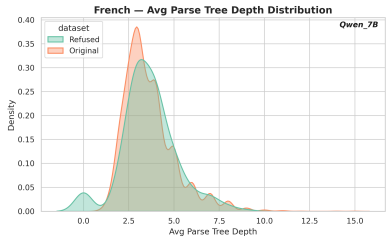
(b) Llama3 8B



(c) Mistral 7B

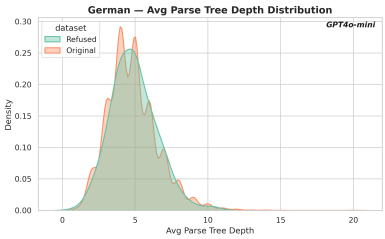


(d) Gemma2 9B

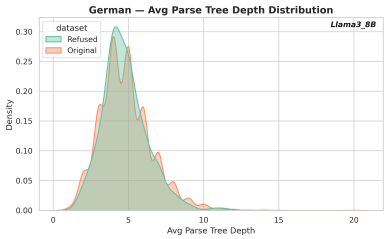


(e) Qwen2.5 7B

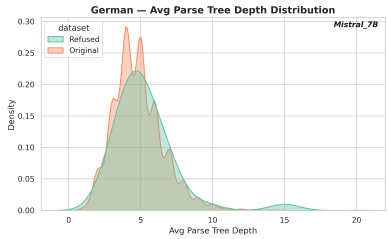
Figure 37: Parse tree depth distributions for French dataset.



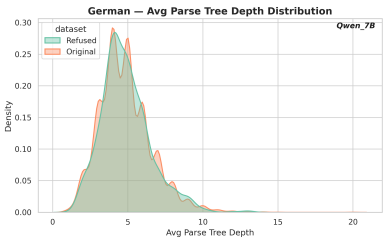
(a) GPT4o-mini



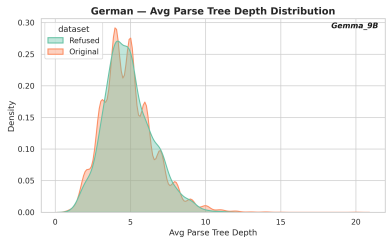
(b) Llama3 8B



(c) Mistral 7B

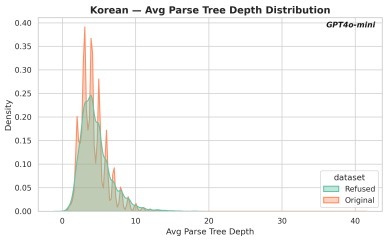


(d) Qwen2.5 7B

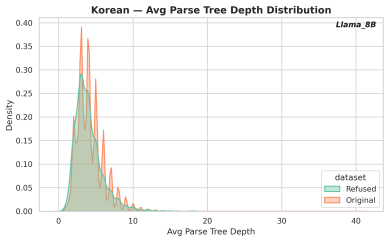


(e) Gemma2 9B

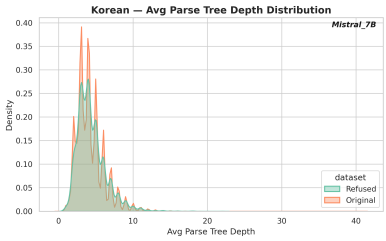
Figure 38: Parse tree depth distributions for German dataset.



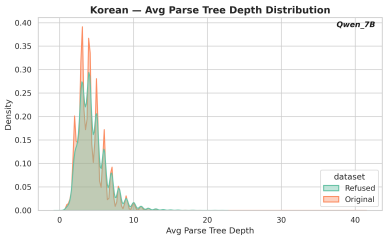
(a) GPT4o-mini



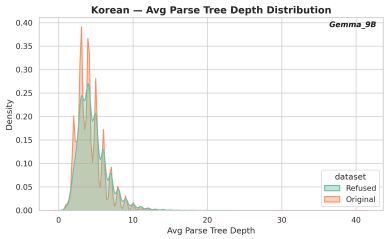
(b) Llama3 8B



(c) Mistral 7B



(d) Qwen2.5 7B



(e) Gemma2 9B

Figure 39: Parse tree depth distribution of the Korean dataset.