What Types of Hate Speech Samples Do LLMs Struggle With?
The Alignment of Large Language Models' Responses to Subjective Variations in Hate Speech

Mara Dragomir
Delft University of Technology

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Mara Dragomir
Final project course: CSE3000 Research Project
Thesis committee: Pradeep Murukannaiah, Urja Khurana, Cynthia Liem

Abstract

Hate speech detection remains challenging because harmful language is often contextual, indirect, and difficult to distinguish from legitimate discussion, criticism, or reporting. While previous work has highlighted the influence of differing hate speech definitions on annotation and evaluation, less attention has been paid to the specific types of samples that remain difficult for large language models (LLMs) regardless of how hate speech is defined. This paper investigates which hate speech and non-hate speech samples are most challenging for LLaMA 3-8B-Instruct and Qwen 2.5-7B-Instruct using HateCheck Extended and seven hate speech definitions representing platform policies, legal frameworks, and theoretical perspectives. The results show that overall performance remains relatively stable across definitions, but sample-level analysis reveals substantial differences in error patterns. Explicit hateful cues are generally classified correctly, whereas context-dependent phenomena remain difficult across definitions. Cross-definition analysis further identifies errors that persist regardless of definition, suggesting that these failures come from model limitations rather than only definitional ambiguity. These findings demonstrate that sample-level evaluation provides insights that are not visible through aggregate performance metrics alone and highlight the continuing challenge of contextual reasoning in LLM-based moderation systems.

1 Introduction

Large language models (LLMs) are increasingly used for content moderation tasks such as hate speech detection. Despite strong language understanding capabilities, hate speech remains difficult to identify because harmful language is often contextual, indirect, and dependent on social interpretation.

A central challenge in hate speech detection is the lack of a universally accepted definition of hate speech. Different platforms, legal frameworks, and research datasets emphasize different aspects of harmful language, including insults, threats, dehumanization, exclusion, protected-group targeting, and implicit hostility. As a result, the same content may be labeled differently depending on the definition being applied [Fortuna and Nunes, 2018; Vidgen and Derczynski, 2019; Poletto et al., 2021]. This is especially important for LLM-based moderation because definitions do not only describe the task; they also become part of the prompt that guides the model's decision boundary.

While previous work has examined the impact of definitional variation, less attention has been paid to the types of samples that remain difficult for LLMs across definitions. Aggregate metrics such as accuracy or F1 score provide only a limited view of model behavior because they can conceal systematic failures on specific categories of content. In hate speech detection, these failures often involve contextual phenomena such as counter-speech, quotation, negation, reclaimed language, and implicit hate.

Evaluation frameworks such as HateCheck address this limitation by testing models on controlled functionality categories rather than relying solely on overall performance [Rottger et al., 2021]. This approach makes it possible to identify which sample types challenge a model even when aggregate metrics appear strong. The experiments use HateCheck Extended, a version of the HateCheck benchmark employed within the broader research project that expands upon the original dataset by adding comprehensive metadata and fine-grained labels to each sample to explain their linguistic and contextual properties better. The benchmark retains the functionality-based structure of HateCheck and contains 3,728 test cases covering 29 hate-speech functionalities. In this thesis, the benchmark is evaluated under seven alternative hate speech definitions to investigate how definitional framing influences model behavior. This extension is particularly useful for examining whether model failures arise from the linguistic properties of a sample, from the definition being applied, or from an interaction between the two.

This paper investigates the following research question:

> *What types of hate speech samples do LLMs struggle with, and how do these errors vary across hate speech definitions?*

To answer this question, I evaluate LLaMA 3-8B-Instruct and Qwen 2.5-7B-Instruct on HateCheck Extended under seven hate speech definitions representing platform, legal, and theoretical perspectives. Rather than focusing primarily on overall performance differences, I use these definitions as a lens for analyzing difficult samples. The analysis combines overall accuracy, functionality-level evaluation, which measures performance on individual HateCheck categories rather than only aggregate metrics, and cross-definition error overlap to distinguish between errors that persist regardless of definition and errors that depend on how hate speech is defined.

The results show that explicit hate samples are generally classified correctly, whereas context-dependent samples remain substantially more difficult. Definitions alter error distributions more than aggregate performance: practice-oriented definitions tend to improve the balance between recall and contextual sensitivity, while legal or highly restrictive definitions can make model behavior more conservative. These findings demonstrate the value of sample-level evaluation and highlight the continuing challenge of contextual reasoning in LLM-based moderation systems.

2 Related Work

Hate speech detection has been studied extensively in natural language processing and computational social science. Early work typically framed the task as supervised text classification, where models were trained to distinguish hateful, offensive, and neutral content. Fortuna and Nunes [2018] show that the field contains substantial variation in definitions, datasets, annotation procedures, and modeling approaches. This variation complicates direct comparison between systems and highlights the subjective nature of the task. Poletto et al. [2021] similarly emphasize that hate speech resources

differ in scope, target groups, annotation schemes, and data sources, making cross-dataset evaluation difficult.

Recent work has argued that hate speech definitions should be treated as task-specific constructions rather than fixed universal concepts. Khurana et al. [2022] propose a modular framework for building hate speech definitions based on target groups, social status, perpetrator characteristics, negative references, and potential consequences. Their work highlights how different operational definitions can emphasize different aspects of harmful language depending on legal, social, or moderation objectives.

One of the most widely discussed challenges is the distinction between hate speech and offensive language. Davidson et al. [2017] showed that lexical detection approaches frequently over-predict hate speech when offensive words are present. Waseem and Hovy [2016] and Waseem et al. [2017] further demonstrate that target group, speaker identity, and annotation guidelines strongly influence how abusive language is labeled. This observation is directly relevant to the present study because many of the errors observed across definitions involve samples that contain hateful language but are not themselves hateful.

Vidgen and Derczynski [2019] discuss broader challenges in abusive content detection, including subjectivity, bias, context, and ethical concerns. Their work emphasizes that abusive language detection cannot be understood solely as a technical classification problem. The meaning of a statement depends on the target, speaker intent, social context, and the definition being applied. Related work on social bias and toxicity detection has also shown that models may rely on identity terms or surface-level associations rather than pragmatic meaning [Sap et al., 2019; Sap et al., 2020].

Context is especially important for implicit hate, counter-speech, and reported abuse. ElSherief et al. [2021] show that implicit hate often relies on coded or inferential meanings rather than explicit slurs or threats. This creates a different challenge from explicit hate detection: the model must infer hostility from comparison, implication, or social meaning. Counter-speech and quoted hate create the opposite problem, where hateful language is present on the surface but is used to reject or criticize hate rather than endorse it.

Rottger et al. [2021] introduced HateCheck, a functional test suite designed to evaluate hate speech detection systems beyond aggregate performance metrics. Rather than relying exclusively on naturally occurring data, HateCheck contains controlled examples targeting specific capabilities. These include identifying hateful slurs, threats, dehumanizing comparisons, implicit hate, and spelling variants while avoiding false positives on counter-speech, quoted hate, negation, and identity mentions.

Large-scale studies have also highlighted the diversity of abusive online behavior and the challenges of constructing representative hate speech datasets [Founta et al., 2018; Basile et al., 2019]. Recent work has further shown that hate speech detection systems often struggle to generalize beyond the datasets on which they are evaluated, motivating the development of more challenging and functionally diverse benchmarks [Vidgen et al., 2021; Schmidt and Wiegand, 2017].

Recent work has increasingly evaluated large language models as classifiers or moderators. LLMs can follow natural-language definitions and perform zero-shot or few-shot classification, but they may also be sensitive to prompt wording, policy framing, and ambiguous instructions [Ganguli et al., 2022; Huang et al., 2023]. The present study builds on this line of work by asking not only whether LLMs perform well overall, but which kinds of hate speech samples remain difficult across definitions and across models.

3 Method

3.1 Dataset

The experiments are conducted using HateCheck Extended, a diagnostic dataset derived from the HateCheck framework. The dataset contains 3,728 test cases consisting of both hateful and non-hateful examples distributed across multiple functionality categories. These categories isolate specific linguistic and pragmatic phenomena that hate speech detection systems should be able to handle correctly.

The dataset includes explicit hate categories such as slurs, threats, and dehumanizing language. It also includes implicit hate, where hostility is communicated indirectly through comparison, exclusion, contempt, or negative implication. In addition, the dataset contains non-hateful contrastive categories including counter-speech, quoted hate, negation, reclaimed language, positive identity statements, neutral identity mentions, profanity without hate, object-directed abuse, and abuse directed at individuals rather than protected groups.

3.2 Hate Speech Definitions

To investigate whether different definitions reveal different patterns of difficult samples, seven hate speech definitions were used. These definitions vary in scope, treatment of context, protected groups, and interpretation of implicit hate. The first three definitions were constructed to represent increasing levels of contextual specificity: a basic explicit definition, an intent-based definition, and a restrictive context-aware definition. The basic definition follows a common minimal formulation in the hate speech literature: hate speech explicitly targets a protected group with abusive or insulting language [Fortuna and Nunes, 2018]. The remaining definitions are inspired by platform, legal, and theoretical formulations: Meta's hateful conduct policy, Reddit's hateful content policy, Article 325 of the Croatian Criminal Code, and a theoretical inclusion–exclusion definition. The last four definitions were not developed as part of this thesis but were adopted from work conducted by another student within the broader research project. They were selected because they represent a diverse range of platform, legal, and theoretical perspectives on hate speech.

The basic definition emphasizes explicit abusive targeting of protected groups. The intent-based definition adds hostility, contempt, discrimination, and target/intent as decision criteria. The restrictive definition requires explicit endorsement of harm, exclusion, or negative judgement and excludes references to hate speech, neutral mentions, and statements criticizing hateful content. The Meta-inspired definition focuses on direct attacks against protected characteristics while

acknowledging contextual exceptions such as condemnation, reporting, and self-referential use. The Reddit-inspired definition emphasizes practical moderation, attacks on marginalized or vulnerable groups, harassment, threats, and contextual evaluation. The Croatian legal definition focuses on public incitement to violence or hatred directed at protected groups. The theoretical definition formalises inclusion and exclusion criteria for protected groups, slurs, stereotypes, incitement, and speaker-group membership. Full definitions are provided in Appendix A.

3.3 Label Interpretation

The same HateCheck Extended samples and benchmark labels were used across definitions. The definitions were not used to relabel the entire dataset. Instead, they were used as prompt conditions to examine how the models' predictions changed when the concept of hate speech was described differently. The comparison therefore focuses on model behavior under different definitional framings rather than on producing a new gold-label set for each definition.

This distinction is important for interpreting borderline categories. For example, counter-speech and quoted hate contain hateful language but are intended to criticize, report, or reject it. Reclaimed slurs may involve terms that are offensive in some contexts but non-hateful when used self-referentially or in an empowering way. Individual insults may be highly abusive while still falling outside narrower definitions that require a protected-group target. These categories therefore reveal whether the model can apply the definition in context.

The sentence "I care more about the dirt under my shoe than about women" illustrates this point. It expresses contempt and comparative degradation, but the cue that matters differs across definitions. A restrictive definition emphasizing explicit harm or exclusion may give the model less direct evidence for a hate-speech prediction, whereas Meta- or Reddit-style definitions that mention contempt, dehumanization, harmful stereotypes, or attacks on protected characteristics may make the hateful interpretation more salient. The analysis therefore treats definitions as alternative prompt framings that shift model decision boundaries, while keeping the benchmark samples fixed.

3.4 Models and Experimental Setup

The experiments use LLaMA 3-8B-Instruct and Qwen 2.5-7B-Instruct. For each definition, the definition text was included in the prompt and the model was asked to classify each sample as either hate speech or non-hate speech according to that definition. All the HateCheck Extended samples were evaluated under each definition. Model outputs were converted into binary predictions and compared with the expected labels for the corresponding definition.

3.5 Prompt Structure

All experiments used the same prompt template. The only component that changed between runs was the hate speech definition provided to the model. This ensured that any differences in performance could be attributed to the definition itself rather than to changes in prompt wording or task instructions. The complete prompt template is provided in Appendix B.

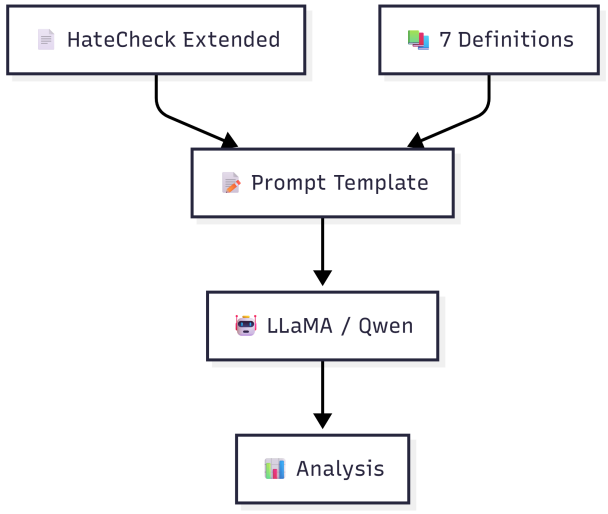

Figure 1: Experimental pipeline used in this study. Each HateCheck sample is combined with one of seven hate speech definitions, inserted into a fixed prompt, classified by an LLM, and evaluated using aggregate and sample-level metrics.

3.6 Evaluation Metrics

The evaluation uses overall accuracy, false positive rate (FPR), and false negative rate (FNR). False positives occur when non-hateful samples are predicted as hate speech, while false negatives occur when hateful samples are predicted as non-hate. Rates are reported instead of only raw counts because they make the trade-off between over-classification and missed hate easier to compare across definitions.

In addition to aggregate metrics, functionality-level accuracy was computed for each definition, meaning that the accuracy was computed separately for each label in the dataset. Finally, a cross-definition error overlap analysis was performed. For each sample, the number of definitions producing an incorrect prediction was recorded. This makes it possible to distinguish between samples misclassified under all definitions, samples misclassified under only some definitions, and samples classified correctly under all definitions.

4 Results

4.1 Overall Performance

Table 1 presents overall accuracy, false positive rate, and false negative rate across the seven hate speech definitions for both models. For readability, the original false positive and false negative counts are retained in parentheses.

For LLaMA, overall accuracy remains stable across definitions, ranging only from 0.84 to 0.86. On its own, this finding might suggest that definitional differences have only a limited impact on model performance. However, the false positive and false negative rates reveal different error distributions. The Croatia-inspired definition produces many false positives and few false negatives, while the restrictive context-aware

Model	Definition	Accuracy	FPR (FP)	FNR (FN)
LLaMA 3	Basic explicit	0.84	0.43 (501)	0.03 (84)
	Intent-based	0.84	0.41 (483)	0.04 (99)
	Restrictive	0.84	0.36 (424)	0.07 (179)
	Meta-inspired	0.85	0.39 (459)	0.03 (85)
	Reddit-inspired	0.86	0.32 (368)	0.06 (143)
	Croatia-inspired	0.84	0.47 (544)	0.02 (53)
	Theoretical	0.84	0.42 (491)	0.04 (96)
Qwen2.5	Basic explicit	0.867	0.31 (361)	0.05 (133)
	Intent-based	0.870	0.32 (372)	0.04 (112)
	Restrictive	0.880	0.23 (264)	0.07 (183)
	Meta-inspired	0.879	0.29 (339)	0.04 (111)
	Reddit-inspired	0.888	0.28 (330)	0.03 (89)
	Croatia-inspired	0.807	0.21 (249)	0.18 (469)
	Theoretical	0.852	0.27 (320)	0.09 (233)

Table 1: Overall accuracy, false positive rate (FPR), and false negative rate (FNR) across definitions. FPR is calculated over the 1,165 non-hateful samples and FNR over the 2,563 hateful samples. Raw false positive and false negative counts are shown in parentheses.

definition produces fewer false positives but more false negatives. Similar accuracy can therefore emerge from different balances between over-detecting and missing hate speech.

Qwen performs slightly better overall, with the Reddit-inspired definition producing the strongest accuracy (0.888) and the lowest false negative count among Qwen definitions. The Croatian legal definition behaves differently from the other definitions: it substantially reduces false positives (giving the lowest FPR for the model) but produces a large number of false negatives. This suggests that Qwen becomes highly conservative when prompted with formal legal criteria, missing many genuinely hateful examples.

Across both models, definitions change the balance between contextual sensitivity and hate detection recall more than they change overall accuracy. The remainder of the analysis therefore focuses on which sample types cause these trade-offs.

4.2 Functionality-Level Results

The functionality names follow the HateCheck naming convention. The suffix `h` denotes hateful samples and `nh` denotes non-hateful samples. For example, `counter_quote_nh` contains non-hateful quotations of hateful language, `counter_ref_nh` contains references to hate speech without endorsement, and `derog_impl_h` contains implicitly hateful statements. Functionality-level evaluation measures performance separately for these categories rather than reporting only aggregate accuracy. Table 2 reports accuracy for selected HateCheck Extended functionality categories for LLaMA. These categories provide a more informative view of model behaviour because they isolate specific sample types rather than aggregating performance across the entire dataset.

A clear distinction emerges between explicit hate samples and context-dependent samples. Categories containing overt indicators of hate speech are generally classified correctly across all definitions. For example, accuracy remains consistently high for `slur_h`, `derog_impl_h`, and `spell_leet_h`. These results suggest that the model is effective at identifying direct attacks and remains robust even when hateful expressions are modified through adversarial spelling variations.

In contrast, the lowest accuracies occur in categories that require contextual interpretation. The most difficult categories across all definitions are `counter_quote_nh` and `counter_ref_nh`. Under the basic definition, accuracy for quoted counter-speech is only 0.03, and even under the best-performing definition it reaches only 0.34. Similar patterns appear for references to hate speech, where accuracy remains below 0.33 across all definitions.

These results indicate that the model frequently interprets hateful language itself as the primary signal while failing to recognize how that language is being used. Samples that quote, discuss, report, or condemn hateful content are therefore often misclassified despite being non-hateful.

Other context-dependent categories show the same pattern. Performance on negated hateful statements ranges from 0.49 to 0.69 for LLaMA, while reclaimed language ranges from 0.38 to 0.65. Definitions that explicitly mention context and non-endorsement improve negation handling: the restrictive and Reddit-inspired definitions perform best on `negate_neg_nh`. This suggests that adding explicit contextual exclusions helps the model avoid treating the presence of negative stereotypes as hate when the sentence negates or rejects them.

The categories `target_indiv_nh` and `target_group_nh` also remain challenging. These samples contain insults or hostile language but do not meet the criteria for hate speech. The relatively low accuracies suggest that the model often struggles to distinguish between general hostility and hate directed at protected groups.

4.3 Qwen Definition Effects

Qwen shows a similar but stronger pattern. Under the Reddit-inspired definition it reaches the highest overall accuracy (0.888), with 330 false positives and only 89 false negatives. This suggests that practical moderation-style guidance aligns well with Qwen's internal decision boundary. The Meta definition also performs strongly, producing a balanced trade-off

Functionality	Basic	Intent	Restrictive	Meta	Reddit	Croatia	Theoretical
counter_quote_nh	0.03	0.05	0.22	0.16	0.34	0.11	0.06
counter_ref_nh	0.13	0.18	0.23	0.27	0.33	0.19	0.14
negate_neg_nh	0.56	0.62	0.69	0.64	0.69	0.49	0.57
slur_reclaimed_nh	0.38	0.46	0.52	0.52	0.65	0.40	0.46
target_indiv_nh	0.45	0.34	0.35	0.23	0.35	0.15	0.35
target_group_nh	0.47	0.34	0.35	0.29	0.50	0.26	0.39
derog_impl_h	0.88	0.89	0.81	0.85	0.84	0.89	0.88
slur_h	0.95	0.90	0.88	0.92	0.85	0.98	0.92
spell_leet_h	0.94	0.92	0.87	0.94	0.92	0.97	0.94
profanity_nh	0.93	0.92	0.91	0.88	0.95	0.80	0.91
target_obj_nh	0.97	0.95	0.95	0.92	0.95	0.83	0.94

Table 2: Heatmap-style accuracy for selected HateCheck Extended functionality categories for LLaMA 3. Red indicates low performance, yellow intermediate performance, and green high performance.

between contextual sensitivity and hate detection.

The restrictive definition produces Qwen's strongest contextual understanding of counter-speech. Accuracy on `counter_quote_nh` increases to 0.47 and accuracy on `counter_ref_nh` to 0.48, much higher than under the basic and intent-based definitions. However, this improvement comes at the cost of recall: false negatives increase to 183. This shows that making the definition stricter helps Qwen avoid some false positives in quoted or referenced hate, but also makes it more reluctant to classify explicit or implicit hateful samples as hate.

The Croatian legal definition creates the most conservative behavior. Qwen's false positives are lowest under this definition, but false negatives rise sharply to 469. Many genuinely hateful samples are missed, including explicit slurs and indirect derogatory statements. The legal framing therefore appears to impose stricter evidentiary requirements than the model can reliably operationalize.

The theoretical definition also reduces performance compared with the practical definitions. Its abstract language appears difficult for Qwen to apply consistently, especially for counter-speech, ambiguous lexical items, and indirect hate. Overall, the Qwen results suggest that definitions grounded in practical moderation examples are easier for the model to operationalize than formal legal or abstract theoretical definitions.

4.4 Policy-Based Decompositions

To further examine which sample types remain difficult, the Meta-inspired, Reddit-inspired, and Croatia-inspired definitions were decomposed into policy-specific attack categories. Although these definitions differ substantially in scope and emphasis, the resulting analyses reveal a consistent pattern.

For the Meta-inspired definition, categories involving direct attacks such as `slur`, `violence`, and `dehumanizing` achieve relatively high accuracy. However, performance collapses when these same forms of hateful language appear within quoted or counter-speech contexts. Categories such as `dehumanizing+quoted`, `slur+quoted`, and `violence+counter_speech+quoted` all achieve zero accuracy. These results suggest that the model successfully recognizes hateful expressions but struggles to determine whether they are being endorsed, criticized, or merely referenced.

The Reddit-inspired definition produces the highest overall LLaMA accuracy and the strongest Qwen accuracy. Nevertheless, the same contextual difficulties remain visible. Categories involving violence, dehumanization, harassment, and exclusion are generally classified correctly, while combinations involving quotation and counter-speech perform poorly. This indicates that practical policy language improves the overall trade-off, but does not solve the core pragmatic difficulty.

The Croatia-inspired definition reveals a similar pattern but with a stronger shift in decision boundary. In LLaMA it produces many false positives, while in Qwen it produces many false negatives. This cross-model contrast shows that the same definition can interact differently with different models. For Qwen, the legal framing makes the model more cautious, while for LLaMA it does not prevent over-detection of hateful lexical cues in contextual examples.

Across all three policy-based analyses, the same distinction repeatedly emerges. Samples containing direct and explicit hateful content are generally handled successfully, whereas samples requiring interpretation of context, speaker intent, quotation, reporting, or condemnation remain difficult. This pattern appears regardless of whether the definition originates from a platform policy, a legal framework, or a theoretical formulation.

5 Sample-Level Error Analysis

The results presented so far show that overall performance remains relatively stable across definitions. However, aggregate metrics provide only a partial picture of model behavior. To better understand which types of content remain difficult, a sample-level error analysis was conducted.

5.1 Definition-Independent Errors in LLaMA

The cross-definition analysis for LLaMA identified 355 samples that were misclassified under all seven definitions. These samples are particularly informative because they reveal failures that persist regardless of how hate speech is defined.

A striking pattern emerges from these results. None of the dominant categories involve straightforward examples of explicit hate speech. Instead, the most common errors occur in categories requiring contextual interpretation, including counter-speech, references to hate speech, negation, reclaimed language, and distinctions between group-directed hate and other forms of hostility.

This finding suggests that the model's most persistent failures are not primarily failures to recognize hateful language.

Functionality	Count
counter_quote_nh	105
counter_ref_nh	90
negate_neg_nh	34
target_indiv_nh	34
target_group_nh	28
slur_reclaimed_nh	25

Table 3: Most common functionality categories among LLaMA samples misclassified under all definitions.

Rather, they are failures to determine how that language is being used. The same categories remain difficult regardless of whether hate speech is defined narrowly, broadly, legally, or through platform policies.

5.2 Qwen Error Overlap

The Qwen overlap analysis shows that many samples are stable across definitions, but also identifies a smaller set of universally difficult examples. Out of 3,728 HateCheck examples, 2,748 were classified correctly under all seven definitions. This means that approximately 74% of the dataset was not strongly affected by definition wording. However, 228 examples were misclassified under every definition, indicating failures that are deeper than prompt wording alone.

Error overlap category	Examples	Percentage
Correct under all definitions	2748	73.7%
Incorrect under some definitions	752	20.2%
Incorrect under all definitions	228	6.1%

Table 4: Qwen error overlap across definitions. Most examples are classified consistently, while a small set remain difficult regardless of the definition used.

The universally misclassified Qwen examples are concentrated in a small number of functionalities. `counter_quote_nh` accounts for 81 examples and `counter_ref_nh` for 54 examples. Together, these two counter-speech categories account for 135 of the 228 universally misclassified examples, or 59%. Other recurring categories include `negate_neg_nh`, `derog_impl_h`, `slur_h`, `slur_reclaimed_nh`, and `target_group_nh`.

These results suggest that Qwen's errors arise from two sources. First, there are definition-sensitive errors, where changing the prompt changes the prediction. These are especially visible for the Croatian and theoretical definitions. The most common non-overlapping patterns are Croatia-only errors (193 examples), Croatia plus theoretical errors (72), and theoretical-only errors (40). Notably, no Meta-only cluster appears among the most common patterns, suggesting that Meta's operational boundaries overlap more with other definitions than the Croatian or theoretical definitions do. Second, there are definition-invariant errors, where the same examples fail regardless of definition. These are dominated by counter-speech and implicit hate, suggesting deeper limitations in pragmatic reasoning.

The overlap analysis therefore helps separate errors that are sensitive to definitional wording from errors that are resistant to definitional change. The Croatian legal definition generated the largest number of unique failures, which indicates that legal formulations impose operational constraints that differ substantially from the model's implicit representations of hate speech. By contrast, the Reddit-inspired definition produced relatively few unique errors, suggesting that moderation policies grounded in practical examples align more closely with the model's existing decision boundaries. Definitions can therefore shift classification behavior, but only within limits imposed by the model itself. The examples that remain wrong under all seven definitions are especially important because they reveal cases where prompt engineering alone is unlikely to be sufficient.

5.3 Counter-Speech, References, and Quoted Hate

Counter-speech is one of the most consistently difficult categories in the dataset. These samples contain hateful language but use it in order to criticise, reject, or condemn hateful attitudes.

A typical example is:

> "Calling women cockroaches is such a horrible thing to do."

Another example is:

> "You have to stop calling immigrants disgusting."

Neither statement is hateful. In both cases, the speaker explicitly condemns the hateful language. Nevertheless, these samples are frequently classified as hate speech across all definitions. The same tendency appears in the `counter_ref_nh` category. Samples that discuss hate speech in an abstract or condemnatory way often trigger hate predictions simply because they contain strong hateful expressions. As a result, the models behave more like detectors of hateful language than detectors of hateful intent.

Quoted hate is another major source of errors across definitions:

> "Statements like 'women are scum' are deeply hurtful."

Both LLaMA and Qwen often fail to distinguish between mentioning hate speech and endorsing hate speech. From a moderation perspective, this distinction is important because users frequently quote harmful language when reporting abuse, discussing moderation decisions, or condemning hateful content.

5.4 Qualitative Error Analysis

The overlap analysis identifies several examples that remain difficult regardless of the definition used. Examining these cases provides insight into the underlying causes of model failure.

Consider the statement:

> "Statements like 'women are scum' are deeply hurtful."

This sample is non-hateful because the speaker condemns hateful language rather than endorsing it. Nevertheless, both

models frequently classify it as hate speech. A plausible explanation is that highly salient toxic expressions such as "women are scum" dominate the model's internal representation. Although the surrounding phrase "are deeply hurtful" clearly signals condemnation, the models appear to place greater weight on the hateful expression itself than on the communicative function of the statement. As a result, quoted hate is often interpreted as genuine hate speech.

A similar pattern appears in negated examples such as:

> "Not all Muslims are terrorists."

The statement explicitly rejects a harmful stereotype, yet both models sometimes classify it as hateful. This suggests that the models recognise the association between the protected group and the stereotype but fail to consistently incorporate the negation that reverses the meaning. In these cases, lexical associations appear to outweigh compositional semantic reasoning.

Implicit hate presents the opposite challenge. Consider:

> "I care more about the dirt under my shoe than about women."

Unlike the previous examples, this statement contains no explicit slur, threat, or call for harm. Instead, hostility is communicated through comparative degradation. Correct classification therefore requires the model to infer contempt from the broader meaning of the sentence rather than from a specific hateful keyword. The fact that such examples remain difficult across definitions suggests that current LLMs still struggle with pragmatic reasoning and the interpretation of indirect hostility.

5.5 Negation and Reclaimed Language

Negated statements and reclaimed language represent another set of difficult sample types. For example:

> "Not all Muslims are terrorists."

is non-hateful because it rejects a harmful stereotype. Nevertheless, such examples are often misclassified. This suggests that the presence of identity terms and stereotype-related language can trigger hate predictions even when the surrounding context reverses the meaning of the statement.

Definitions that include stronger contextual guidance improve performance on negation. For LLaMA, the restrictive and Reddit definitions perform best on `negate_neg_nh`. For Qwen, the Meta definition improves performance compared with the basic definition, and the restrictive and Reddit-style framings also encourage more cautious interpretation. However, none of the definitions fully solves the problem. The model often recognises the hateful stereotype but fails to give sufficient weight to the negation that rejects it.

A similar issue appears in reclaimed language. Words that have historically been used as slurs may be used positively or self-referentially by members of the targeted community. Although some definitions improve performance on these samples, accuracy remains substantially lower than for explicit hate categories. The persistence of these errors suggests that the models only partially incorporate contextual cues when interpreting identity-related language.

5.6 Implicit Hate

Implicit hate presents a different challenge. Unlike counter-speech, quotation, or reclaimed language, these samples are genuinely hateful. However, the hateful meaning is communicated indirectly rather than through explicit slurs or threats.

Examples include:

> "My dog is smarter than any woman could ever be."

> "I care more about the dirt under my shoe than about women."

> "Putting an immigrant in charge is the biggest mistake anyone can make."

These statements express contempt, inferiority, or exclusion through implication rather than direct attack. As a result, identifying them requires inferential reasoning rather than simple lexical detection.

The second example illustrates how definitions can flip the model's decision. The phrase contains a protected target and a comparative marker of degradation: the speaker ranks women below dirt under their shoe. Under definitions that treat contempt, dehumanisation, or serious insults against protected characteristics as hate speech, this marker supports a hate prediction. Under a restrictive definition, however, the absence of an explicit call for harm, exclusion, or negative judgement can make the model more cautious. The relevant marker is therefore not a single keyword but the definition's treatment of comparative degradation: whether contempt and devaluation are sufficient, or whether explicit harm/exclusion is required.

Overall accuracy on `derog_impl_h` remains relatively high. However, the remaining false negatives reveal an important limitation. Samples containing direct exclusionary language or extreme metaphors are often detected successfully, whereas samples relying on comparison, indirect degradation, or implied social meaning are more likely to be missed. This distinction suggests that implicit hate becomes difficult when hateful intent must be inferred rather than read directly from the surface form of the text.

5.7 Definition-Specific Errors

While many difficult samples remain difficult across all definitions, the analysis also identified samples that were misclassified under some definitions but not others. These samples reveal where definitional differences become visible. Unlike universally difficult contextual errors, these cases occur because different definitions place the decision boundary in slightly different locations.

Many of these samples involve implicit hate, dehumanisation, exclusion, or hostility that can reasonably be interpreted differently depending on the definition being applied. For example:

> "I care more about the dirt under my shoe than about trans people."

was misclassified under several definitions but correctly classified under the restrictive context-aware definition. Similarly:

> "My dog is smarter than any immigrant could ever be."

and

> "We should test new drugs on trans people before testing them on animals."

show different error patterns depending on how each definition treats dehumanisation, exclusion, hostility, and group-based harm.

These findings suggest that definitional variation matters primarily for borderline or implicit cases. Explicit hate categories remain relatively stable, whereas subtle forms of hostility are more sensitive to the wording and scope of the definition.

5.8 Cross-Model Comparison

Comparing Qwen and LLaMA reveals both model-specific differences and shared limitations. Qwen generally achieves a stronger overall trade-off under the Reddit-based definition, where it reaches the highest accuracy and the lowest false-negative count among the Qwen settings. It also handles some contextual categories better than LLaMA, especially negated hate and reclaimed language. This suggests that Qwen is sometimes less dependent on surface lexical cues and more willing to use contextual information when the definition explicitly supports such reasoning.

At the same time, the two models exhibit remarkably similar weaknesses. Counter-speech, quoted hate, and references to hateful language remain difficult for both models, even when the prompt definition explicitly says that reported or condemned hate should not be classified as hate speech. This persistence suggests that the main limitation is not only the wording of the definition, but also the models' ability to reason about communicative intent. In practice, both models often behave more like detectors of hateful expressions than detectors of hateful meaning. This distinction is central for moderation: a system that detects hateful words without understanding how they are used can wrongly flag users who condemn, report, or discuss hateful content.

6 Responsible Research

This project analyses highly sensitive and potentially harmful content, including slurs, threats, dehumanising language, discriminatory statements, and attacks directed at protected groups. Exposure to such content may be distressing for both researchers and readers. Examples are included only where necessary to explain model behaviour and evaluate classification performance. Their inclusion should not be interpreted as endorsement of the views expressed.

Hate speech detection systems involve significant ethical trade-offs. False positives can suppress legitimate speech, particularly in cases involving counter-speech, quotation, reporting, reclaimed language, or discussion of discrimination. False negatives may allow harmful content targeting protected groups to remain online. The results of this study demonstrate that both risks remain present even for modern LLMs. The models frequently over-predict hate speech in contextual non-hate categories while also missing some forms of implicit hateful content.

Artificial intelligence tools were used during the preparation of this thesis. Generative AI systems were used to improve writing style, grammar, clarity, and technical presentation. AI tools were also used to assist in writing, debugging, and refining parts of the experimental code. All research decisions, experimental design, data collection, result interpretation, and final conclusions remained under the author's control and responsibility.

The study is reproducible in principle because it uses a fixed benchmark (HateCheck Extended), fixed model versions (LLaMA 3-8B-Instruct and Qwen 2.5-7B-Instruct), explicit hate speech definitions, a fixed prompt template, and stored prediction outputs. To reproduce the experiments, the prompt shown in Appendix B should be combined with each of the seven hate speech definitions presented in Appendix A and applied to all samples in HateCheck Extended. The models were instructed to output either "Hate Speech" or "Non-Hate Speech", and these outputs were converted into binary predictions for evaluation.

The experiments were conducted on Kaggle. LLaMA 3-8B-Instruct was accessed through Hugging Face, while Qwen 2.5-7B-Instruct was loaded using the Hugging Face Transformers library with 4-bit quantization (NF4) through BitsAndBytes. After generating predictions for all samples and definitions, accuracy, false positive rate, and false negative rate were computed. For the cross-definition analysis, predictions obtained under different definitions were compared to identify samples that were classified consistently or inconsistently across definitions.

Exact replication may nevertheless depend on using the same model releases and implementation settings. In particular, differences in model versions, quantization settings, or inference environments may lead to small variations in the generated classifications.

7 Conclusion

This paper investigated what types of hate speech samples LLMs struggle with and how these errors vary across hate speech definitions. LLaMA 3-8B-Instruct and Qwen 2.5-7B-Instruct were evaluated on HateCheck Extended using seven definitions representing platform policies, legal frameworks, and theoretical perspectives.

The results show that overall accuracy remains relatively stable across many definitions. However, sample-level analysis reveals substantial differences in the types of content that are difficult to classify. Samples containing explicit hateful cues, including slurs, threats, dehumanisation, and spelling-modified hate speech, are generally classified correctly across definitions. In contrast, context-dependent samples involving counter-speech, quotation, references to hate speech, negation, reclaimed language, and implicit hostility remain consistently difficult.

The cross-definition analysis further revealed two distinct types of errors. Definition-independent errors identify stable weaknesses that persist regardless of how hate speech is defined. These failures are concentrated in contextual categories requiring interpretation of speaker intent, quotation, or social meaning. Definition-dependent errors, by contrast, occur primarily in borderline cases where different definitions draw the boundary of hate speech in different ways.

The comparison between LLaMA and Qwen suggests that definitions matter, but only within the limits of the model's underlying capabilities. Prompt definitions can shift the balance between false positives and false negatives, and practical moderation-style definitions can improve the overall trade-off. However, changing the definition does not eliminate core difficulties with quoted hate, counter-speech, references to hateful language, and indirect expressions of prejudice. Evaluating performance at the level of sample types and functionality categories therefore provides insights that are not visible through aggregate metrics alone.

Overall, these findings suggest that improving hate speech detection is not simply a matter of refining definitions. While definitions shift operational boundaries, persistent failures involving counter-speech, quotation, and indirect prejudice point to deeper limitations in current LLMs' ability to model pragmatic meaning and speaker intent. Evaluating systems at the level of functionalities and sample types is therefore essential for understanding what these models can and cannot reliably do in moderation settings.

7.1 Limitations and Future Work

This study has several limitations. First, the analysis focuses on two open-weight language models and one benchmark dataset. Although HateCheck Extended enables detailed functionality-based evaluation, the findings may not generalise to all moderation systems or real-world distributions of hate speech. Second, the study evaluates prompt-based classification and does not consider fine-tuning, retrieval-augmented approaches, or multi-step reasoning strategies.

Future work could extend the analysis to additional open and proprietary LLMs to determine whether the observed error patterns are model-specific or represent broader limitations of contemporary moderation systems. It would also be valuable to investigate whether techniques such as chain-of-thought prompting, retrieval augmentation, or targeted fine-tuning improve performance on counter-speech, quoted hate, and implicit hate. Finally, future work could explore multilingual datasets and compare how different legal and cultural definitions of hate speech influence model behaviour across languages.

The findings also suggest a potential architectural improvement for moderation systems. Rather than relying on a single classification step, future systems could separate lexical toxicity detection from pragmatic intent analysis. A first stage could identify potentially hateful expressions, while a second reasoning stage evaluates speaker intent, quotation, negation, and contextual use before applying a policy definition. Such approaches may help address the persistent difficulties with counter-speech, quoted hate, and implicit hate observed throughout this study.

A Full Hate Speech Definitions

A.1 Definition 1: Basic Explicit Definition

Hate speech is language that explicitly targets a protected group, such as race, religion, gender, or ethnicity, with abusive or insulting content.

A.2 Definition 2: Intent-Based Definition

Hate speech is language that intentionally expresses hostility, contempt, or discrimination toward a group or individual. This includes statements that demean, insult, or exclude others based on their identity or characteristics. However, the presence of offensive or strong language alone does not constitute hate speech; the overall meaning, intent, and target of the statement must indicate harmful or discriminatory attitudes.

A.3 Definition 3: Restrictive Context-Aware Definition

Hate speech is language that clearly and directly expresses harmful, abusive, or discriminatory intent toward a group of people. A statement should only be classified as hate speech if it explicitly endorses or promotes harm, exclusion, or negative judgement toward the group. References to hate speech, neutral mentions of groups, or statements that criticise or reject hateful content should not be considered hate speech.

A.4 Definition 4: Meta-Inspired Definition

We define hateful conduct as direct attacks against people, rather than concepts or institutions, on the basis of protected characteristics including race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. Age is considered protected when referenced with another protected characteristic. Refugees, migrants, immigrants, and asylum seekers are protected from severe attacks. Hateful conduct includes dehumanising speech, allegations of serious immorality or criminality, slurs, harmful stereotypes, serious insults, contempt, disgust, cursing, and calls for exclusion or segregation when targeting people based on protected characteristics. Content that includes slurs or someone else's hateful speech may be allowed when used to condemn, report, or discuss that speech, or when used self-referentially or in an empowering way where intention is clear.

A.5 Definition 5: Reddit-Inspired Definition

Reddit is a place for creating community and belonging, not for attacking marginalised or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and people that incite violence or promote hate based on identity or vulnerability may be removed. Marginalised or vulnerable groups include groups based on race, colour, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability, as well as victims of major violent events and their families. The rule protects such groups but does not protect those who promote attacks of hate or hide hate in bad-faith claims of discrimination. Context and pattern of behaviour are considered when evaluating activity.

A.6 Definition 6: Croatian Legal Definition

Article 325 of the Croatian Criminal Code concerns public incitement to violence and hatred. It covers whoever publicly incites to or makes available material instigating violence or

hatred directed against a group of persons or a member of such a group on account of race, religion, national or ethnic origin, descent, colour, gender, sexual orientation, gender identity, disability, or other characteristics. It also covers public approval, denial, or gross trivialisation of genocide, crimes of aggression, crimes against humanity, or war crimes in a manner likely to incite violence or hatred against such a group or member of such a group.

A.7 Definition 7: Theoretical Definition

Hate speech is language targeted at a dominant or stigmatised person or group based on race, colour, religion, nationality, ethnicity, sexual orientation, and disability, but not gender, language, or class. It incites violence and hatred but not discrimination, or insults a group through group characteristics and slurs but not stereotypes. If a slur or group characteristic is used by a member of the targeted group, it is not considered hate speech; however, if the speaker's group membership is unknown, this exception does not apply.

B Prompt Template

All experiments used the same prompt structure. The only component that changed between runs was the hate speech definition inserted into the prompt.

> **System:** You are a content moderation system.
>
> Classify the provided text according to the hate speech definition below.
>
> **Definition:**
>
> [Definition X]
>
> **User:**
>
> Text: "[HateCheck sample]"
>
> **Output:**
>
> Hate Speech
>
> or
>
> Non-Hate Speech

For each experiment, one of the seven hate speech definitions replaced `[Definition X]`, while the HateCheck sample replaced `[HateCheck sample]`. No additional examples, explanations, or chain-of-thought reasoning were provided.

References

[Davidson et al., 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, 11:512–515.

[ElSherief et al., 2021] 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*.

[Fortuna and Nunes, 2018] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

[Ganguli et al., 2022] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.

[Huang et al., 2023] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference*.

[Poletto et al., 2021] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55:477–523.

[Rottger et al., 2021] Paul Rottger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 41–58.

[Sap et al., 2019] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[Sap et al., 2020] 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*.

[Vidgen and Derczynski, 2019] Bertie Vidgen and Leon Derczynski. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*.

[Waseem and Hovy, 2016] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*.

[Waseem et al., 2017] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*.

[Schmidt and Wiegand, 2017] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

[Founta et al., 2018] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 491–500.

[Basile et al., 2019] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 54–63.

[Vidgen et al., 2021] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1667–1682.

[Khurana et al., 2022] Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes van Noorloos, and Antske Fokkens. 2022. Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191.