What Types of Hate Speech Samples Do LLMs Struggle With?

The Alignment of LLM Responses to Subjective Variations in Hate Speech

Mara Dragomir

1. MOTIVATION

- Hate speech is subjective: platforms, laws, and researchers draw boundaries differently.
- Overall accuracy hides whether errors are false positives on legitimate speech or false negatives on harmful speech.
- The goal is to identify sample types that remain difficult across **seven definitions** and **two LLMs**.

Research question

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

2. DEFINITIONS COMPARED

Definition	Main focus	Context handling
Basic explicit	Explicit abusive or insulting attacks on protected groups	Minimal context
Intent-based	Hostility, contempt, discrimination, target, and intent	Meaning and target matter
Restrictive	Explicit endorsement of harm, exclusion, or negative judgement	Strong exclusions for quotation, rejection, and neutral mentions
Meta-inspired	Direct attacks, slurs, stereotypes, contempt, disgust, exclusion	Allows reporting, condemnation, and self-reference
Reddit-inspired	Attacks on marginalized/vulnerable groups, harassment, bullying, threats	Practical moderation context and behavior patterns
Croatia-inspired	Public incitement to violence or hatred toward protected groups	Formal legal threshold
Theoretical	Inclusion/exclusion criteria for groups, slurs, stereotypes, speaker membership	Abstract boundary rules

Definition role

The definitions were used as prompt framings, not as seven new gold-label datasets. The analysis therefore shows how model behavior shifts under different operational descriptions of hate speech.

Label interpretation

Borderline categories reveal whether the model can apply the definition in context. Counter-speech and quoted hate contain hateful words but should be non-hate; individual abuse can be hostile without targeting a protected group.

Decision boundary effect

Broader platform-style definitions tend to keep recall high, while restrictive or legal definitions can reduce false positives by making the model more cautious.

3. DATASET: HATECHECK EXTENDED

HateCheck Extended contains **3,728 controlled test cases** covering **29 functionality categories**.

H — Explicit hate

Q — Quoted / contextual

I — Implicit hate

N — Negation / variation

- Hateful categories: slurs, threats, dehumanization, spelling variants, implicit derogation.
- Non-hateful contrasts: counter-speech, quotation, reporting, negation, reclaimed language.
- Useful because each category isolates a specific linguistic or pragmatic capability.

Why this benchmark helps

Instead of only testing random examples, HateCheck Extended separates **what** the model must understand: direct attacks, target groups, quotation, negation, speaker intent, and indirect hostility.

Most informative cases

The most revealing samples are not obvious slurs. They are cases where hateful words appear in non-hateful use, or where hateful meaning appears without obvious hate words.

Metric denominators

False positive rate is computed over **1,165 non-hateful samples**; false negative rate is computed over **2,563 hateful samples**. This makes over-flagging and missed-hate trade-offs comparable across definitions.

4. EXPERIMENTAL SETUP

HateCheck samples → **Insert one definition** → **Classify with LLaMA + Qwen** → **Collect predictions** → **Metrics + error overlap**

- Same prompt template for all runs; only the inserted definition changed.
- Models: LLaMA 3-8B-Instruct and Qwen 2.5-7B-Instruct.
- Metrics: accuracy, FPR, FNR, functionality accuracy, and cross-definition error overlap.

Prompt condition

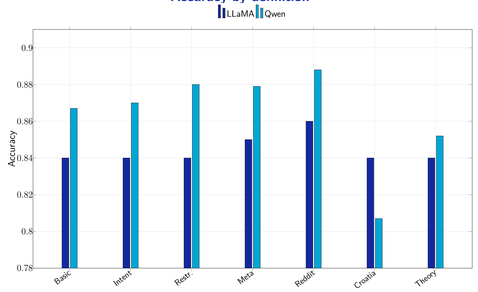
Each sample was classified as **Hate Speech** or **Non-Hate Speech** according to the active definition. This isolates definition wording as the experimental variable.

Overlap analysis

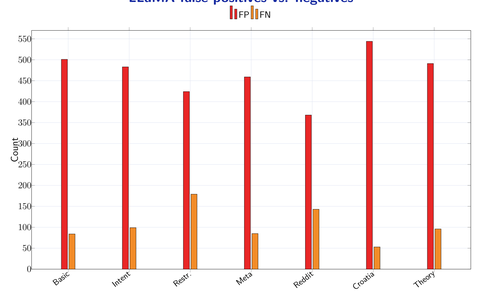
For each sample, errors were counted across all seven definitions to separate definition-sensitive mistakes from persistent model limitations.

5. OVERALL RESULTS

Accuracy by definition



LLaMA false positives vs. negatives



Main pattern

LLaMA accuracy stays between 0.84–0.86, but the false positive/false negative balance changes strongly. Qwen performs best overall with the Reddit-inspired definition (0.888), while the Croatian legal framing lowers Qwen false positives but sharply increases false negatives (469).

Model	Best trade-off	Lower false positives	Main risk
LLaMA	Reddit-inspired: 0.86 accuracy, 368 FP, 143 FN	Restrictive reduces FP to 424 compared with 501 under Basic	Still over-flags counter-speech and quoted hate
Qwen	Reddit-inspired: 0.888 accuracy, 330 FP, 89 FN	Croatian gives lowest FP (249)	Croatian produces many FN (469), missing harmful samples

Accuracy is not enough

Two definitions can have similar accuracy while making opposite moderation errors: suppressing legitimate speech or missing hateful content.

Best-performing style

Practical moderation language, especially Reddit-inspired framing, aligns better with model decision boundaries than legal or abstract definitions.

Croatian legal contrast

The same legal framing affects the two models differently: LLaMA over-flags contextual examples, while Qwen becomes highly conservative and misses many hateful samples.

6. CATEGORY-LEVEL AND MODEL FINDINGS

LLaMA category	Basic	Intent	Restr.	Meta	Reddit	Croatia	Theory
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
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

Qwen effects

Qwen handles some contextual categories better than LLaMA. Restrictive guidance improves quoted/reference counter-speech, but raises false negatives.

Policy decompositions

Meta, Reddit, and Croatia analyses all show the same split: direct attacks are easier; quoted/counter-speech variants collapse.

Finding	LLaMA	Qwen
Best framing	Reddit-inspired accuracy reaches 0.86	Reddit-inspired accuracy reaches 0.888
Conservative shift	Restrictive lowers false positives but raises false negatives	Croatian legal definition creates many missed hateful samples
Stable strengths	Explicit slurs, spelling variants, violent threats	Similar explicit-hate strengths, often stronger trade-off
Stable weaknesses	Quoted hate, references, negation, reclaimed language	Same core weaknesses, fewer universal errors

Key interpretation

The hard categories cluster around one distinction: models can often recognize hateful expressions, but do not reliably infer whether the speaker is endorsing, rejecting, quoting, or indirectly implying hate.

LLaMA universal errors

Most common all-definition failures: quoted counter-speech (105), references to hate speech (90), negation (34), individual abuse (34), group-target ambiguity (28), reclaimed slurs (25).

Qwen overlap summary

Qwen is more stable overall: 73.7% of examples are correct under all definitions. The remaining universal errors still concentrate in counter-speech and implicit/contextual categories.

7. ERROR ANALYSIS: WHAT STAYS HARD?

Counter-speech / references

Non-hateful condemnation is often flagged because hateful language appears on the surface.

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

"You have to stop calling immigrants disgusting."

Negation / reclaimed language

The stereotype or slur is recognized, but the context that rejects or reclaims it is underweighted.

"Not all Muslims are terrorists."

"I'm proud to be gay."

Quoted hate

Models struggle to separate mentioning hateful language from endorsing it.

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

Implicit hate

Indirect contempt requires inference rather than keyword matching.

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

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

Cross-definition overlap

LLaMA: 355 samples were wrong under all seven definitions, dominated by counter_quote_nh (105), counter_ref_nh (90), negate_neg_nh (34), target_indiv_nh (34), target_group_nh (28), and slur_reclaimed_nh (25).

Qwen: 2,748 samples were correct under all definitions; 752 were wrong under some definitions; 228 were wrong under all definitions.

Counter-speech categories account for 135 of the universal Qwen errors.

False-positive mechanism

A toxic phrase becomes too salient, so the model flags reporting, criticism, or quotation as hate speech. This risks suppressing users who call out abuse.

False-negative mechanism

Indirect hate can avoid obvious toxic words. The model must infer contempt, exclusion, or devaluation from full sentence meaning.

Definition-sensitive errors

Some samples flip depending on definition wording, especially implicit degradation and dehumanization. These cases show where the boundary of hate speech is genuinely definition-dependent rather than simply a model failure.

Shared model limitation

Even when definitions explicitly mention reporting, condemnation, or non-endorsement, both models still often classify quoted or referenced hate as hate speech. This points to a deeper weakness in reasoning about communicative function.

8. KEY TAKEAWAYS

- Explicit hateful cues are usually detected correctly, including slurs and spelling-modified hate.
- Contextual non-hate is the central false-positive problem: reporting, quoting, condemning, negating, and reclaiming hateful language.
- Definitions shift false positives and false negatives more than they shift aggregate accuracy.
- Practical moderation-style definitions align best with Qwen; legal or abstract definitions can create conservative or inconsistent behavior.
- Persistent cross-definition errors suggest deeper limits in pragmatic reasoning and speaker-intent modeling.

Moderation implication

A model that detects hateful words without understanding communicative intent may wrongly flag counter-speech and reporting, while still missing subtle prejudice. Evaluation should inspect functionality categories and sample-level overlap, not only aggregate scores.

False positives matter

Over-flagging can silence people who report, quote, or condemn hate speech.

False negatives matter

Under-flagging lets indirect abuse remain visible, especially when hostility is subtle or coded.

9. FUTURE WORK

- Evaluate additional open and proprietary LLMs and real-world moderation datasets.
- Test chain-of-thought prompting, retrieval augmentation, and targeted fine-tuning for contextual categories.
- Extend to multilingual datasets and culturally specific legal definitions.