

Tailoring In-Context Learning Techniques for Definition-Based Hate Speech Detection in Large Language Models

Parham Bateni Pradeep Kumar Murukannaiah Urja Khurana
pbateni@tudelft.nl p.k.murukannaiah@tudelft.nl u.khurana@tudelft.nl

EEMCS, Delft University of Technology
Delft, The Netherlands

Abstract

Hate speech lacks a single agreed definition across legal, social, and benchmark contexts, yet instruction-tuned large language models (LLMs) are increasingly used for hate speech detection. These systems are usually prompted with generic task instructions, so their predictions may reflect built-in moderation priors rather than a different operational definition from a given policy, benchmark, or jurisdiction. We investigate the ability of LLMs to align with specific hate speech definitions through zero-shot and few-shot prompting, by including the hate speech definition in the prompt. We opt for three definitions from different sources: dataset (HateCheck), social media (Reddit), and law (Bulgaria). We use HateCheck, a template-based hate speech diagnostic test set, to verify this alignment by relabeling the same test cases. We compare three LLMs, two definition inclusion conditions, and four prompting strategies. Our findings suggest that regardless of the prompting strategy, LLMs are not reliable in aligning properly with non-HateCheck definitions.¹

1 Introduction

Large language models (LLMs) are widely used in everyday applications (Eloundou et al., 2024; Kasneci et al., 2023; Noy and Zhang, 2023), including content moderation pipelines that flag hateful language. As they can follow natural-language task instructions, prompting is an appealing route to hate speech detection. However, as hate speech has no single agreed definition and varies between legal, social, and benchmark contexts (Hietanen and Eddebo, 2023; Khurana et al., 2022), what kind of hate speech the LLM captures remains a question.

¹**OFFENSIVE CONTENT WARNING:** This paper contains some examples of hateful content. This is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Reader discretion is advised.

Moderation systems are often built to adhere to a particular policy or legal scope, yet it remains unclear how well LLMs can be steered toward that definition through prompting alone. Although prior work has started to test definition-aware zero-shot prompts (Plaza-del Arco et al., 2023; Melis et al., 2025), much less is known about alignment to different definitions and how definition wording interacts with few-shot exemplars.

We thus turn to prompting, due to its computational efficiency, asking whether prompting alone can align model behavior with explicit hate speech definitions well enough to support reliable classification. **How well can LLMs align with hate speech definitions across different definitions, models, and prompting strategies?**

To answer this question, we investigate whether injecting an explicit hate speech definition in the prompt improves model alignment with that definition compared to generic prompting, which prompting strategy (zero-shot or few-shot) is more effective under definition-aware settings, and how alignment varies across LLMs.

We investigate this on HateCheck (Röttger et al., 2021), a diagnostic benchmark that tests hate speech detection through controlled linguistic phenomena such as negation and contrastive non-hateful cases, rather than surface cues alone. In addition, to probe alignment with definitions outside the benchmark’s original scope, we add Reddit and Bulgaria-aligned variants (Reddit, 2024; National Assembly of the Republic of Bulgaria, 2011): the test cases stay fixed, but labels are reassigned according to each definition and the corresponding definition is injected into the prompt.

We select these two external definitions because they differ from HateCheck in which groups are protected, which incited consequences count as hate speech, and/or whether dominant groups can be targets. This allows us to test whether models can adapt to a specified real-world boundary rather

than simply reproduce a single benchmark convention. Moreover, they represent two distinct sources of hate speech regulation: a social-media content policy and a national criminal-code provision.

Within this setup, we compare three instruction-tuned LLMs (Gemma-2, Llama-3.2, and Qwen2.5) under two definition inclusion conditions (No Definition and HSC) and four prompting strategies (zero-shot and three few-shot exemplar-selection methods), and we evaluate performance with macro F1 on binary hateful versus non-hateful classification. Furthermore, we compute the sample-level label flip rate to measure how often predictions change when the definition is added.

Our results show that definitional alignment is neither automatic nor uniform: the same definition can help in one configuration and hurt in another, depending on the model and the rest of the prompt, while in-context exemplars often matter more than including the definition alone. Moreover, performance drops on the relabeled samples, which suggests that prompting cannot fully reconcile models with every operational hate speech boundary.

2 Related Work

Definition of Hate Speech Hate speech is not defined uniformly across legal and academic settings (Hietanen and Eddebo, 2023; Khurana et al., 2022). Matsuda et al. (2018) describe it as abusive or assaultive speech targeting historically oppressed groups, whereas the Oxford English Dictionary treats it as abusive or threatening language that expresses prejudice against a group (Oxford University Press). By contrast, EU legislation centers on public incitement to violence or hatred against groups defined by protected characteristics (Council of the European Union, 2008). Although these accounts agree that groups can be targeted, they differ in scope and interpretation across jurisdictions, which carries through to annotation practice and dataset design. For example, denying genocides or mass atrocities may be unlawful hate speech in several European jurisdictions yet remain protected expression under the First Amendment in the United States (Rosenfeld, 2003). Hate speech is therefore best treated as a context-dependent concept rather than a fixed linguistic category.

Structured Representations of Definitions Because definitions vary by context, several structured representations have been proposed to make their scope explicit. The Hate Speech Criteria (HSC)

framework (Khurana et al., 2022) decomposes hate speech into fine-grained dimensions such as target group, dominance, perpetrator role, type of negative reference, and possible consequences. Similarly, Melis et al. (2025) offer a modular taxonomy of concepts, including communication form, targets of hostility, protected attributes, and consequences, that can be combined to represent definitions from the literature.

Hate Speech Detection Hate speech detection has a long history in NLP, from rule-based systems and classical machine learning (Davidson et al., 2017; Zhang and Luo, 2019) to neural and Transformer-based models that capture richer context (Kim, 2014; Liu et al., 2019; Mathew et al., 2021). More recently, instruction-tuned LLMs have achieved strong zero-shot performance through prompt-based reasoning (Plaza-del Arco et al., 2023), reducing the need for task-specific fine-tuning in some settings. Across paradigms, however, performance remains sensitive to how hate speech is defined in a given dataset (Khurana et al., 2022; Davidson et al., 2017).

Definition Alignment in Evaluation Recent work has therefore asked whether models behave according to the hate speech definition they are trained on. Khurana et al. (2025) address this question for fine-tuned encoder models by decomposing dataset definitions into HSC dimensions, enriching HateCheck by labeling all these dimensions for each item in the diagnostic set, and evaluating whether models trained on each dataset are aligned with that dataset’s definition. Applied to models such as fBERT and RoBERTa on six English hate speech datasets, they find that model behavior deviates from the definition of the dataset they were trained on. This highlights a definitional alignment problem in supervised hate speech detection, but it leaves open whether instruction-tuned LLMs can be steered toward a given scope through prompting.

On the prompting side, Melis et al. (2025) study whether modular definitions improve zero-shot performance on HateCheck, and Plaza-del Arco et al. (2023) show that respectful task framing can improve zero-shot classification. Neither line of work, however, examines few-shot prompting, and neither tests alignment across multiple operational definitions with matched labels, as we do for the Reddit and Bulgaria scopes.

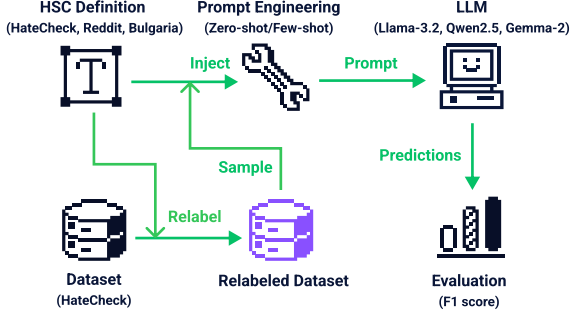


Figure 1: Pipeline for prompt-based hate speech detection under specific definitions.

Key Research Gap Taken together, prior work shows that hate speech boundaries vary by context and that definitional misalignment can persist even after fine-tuning, yet the prompting-based behavior of instruction-tuned LLMs remains underexplored when definition wording, exemplar selection, and model choice are varied jointly. We address this gap through a factorial study that injects dataset-specific HSC definitions into prompts and compares zero-shot and few-shot strategies on HateCheck and relabeled variants.

3 Methodology

Given a hate speech definition, we relabel a diagnostic hate speech dataset to align with the definition. We inject the definition into the model prompt and also experiment with labeled examples (exemplars) through few-shot prompting. Figure 1 illustrates this pipeline, and the following subsections describe each component in detail.

3.1 Dataset

We evaluate all experiments on an extension of HateCheck (Röttger et al., 2021) from Khurana et al. (2025), a template-based functional diagnostic benchmark for hate speech detection. The samples in the dataset are clear-cut cases of (non) hate speech, and its contrastive non-hateful cases are deliberately constructed to resemble hateful ones. Hence, correct classification depends on the precise hate speech boundary rather than surface cues alone. This sensitivity matches the definitional alignment question we study.

In the extension, each sample is annotated with dimensions from HSC², resulting in 3,634 labeled

English cases (2,277 hateful, 1,357 non-hateful). These cases span 24 functionalities, each testing whether a model can distinguish hateful from non-hateful text in a specific scenario. Functionalities cover distinct hate forms such as slurs and threats or contrastive non-hate cases, including negations and counter-speech.

3.2 Definitions

We test two scenarios: (1) no definition given and (2) a definition specified by the dimensions of HSC. Both of these scenarios are also combined with few-shot prompting strategies which we elaborate on in Section 4.

We derive HSC criteria for three definitions that reflect distinct operational contexts; each source corresponds to a re-labeled version of HateCheck used in evaluation: **HateCheck** is the benchmark’s provided scope, aligned with the original HateCheck labels (Röttger et al., 2021). **Reddit** follows the platform’s content policy on hate and harassment (Reddit, 2024). **Bulgaria** follows the statutory definition of hate speech under Bulgarian criminal law (Articles 162 and 164) (National Assembly of the Republic of Bulgaria, 2011). Together, these sources span dataset definition, social-media platform policy, and legal framing. Under HSC, the definition injected into the prompt always matches the evaluation scope of the dataset version being tested. This is ensured by relabeling the dataset samples using the annotated HSC values to align with the definition (see Section 3.3).

Each definition is specified along the following HSC dimensions³: **Target groups** specify which identity-based groups may be targeted (e.g., gender, race, religion, sexual orientation). **Dominance** indicates whether dominant groups (e.g., men, white people) may also be treated as targets. **Explicit reference** defines which forms of group-directed reference count toward hate speech (e.g., stereotype, group characteristic, slur). **Consequences** specify which incited or implied effects are treated as constitutive of hate speech (e.g., violence, hate, discrimination). **Group insult** indicates whether insulting a group counts as hate speech even without a matching incited consequence.

Each definition is encoded through HSC and compared to a No Definition baseline, in which the definition block is omitted entirely from the prompt.

²We follow Khurana et al. (2025) in excluding samples with spelling errors and adding samples targeting dominant groups (men and white people) using functionality templates.

³None of the three source definitions specify perpetrator characteristics (e.g., speaker role or group membership), so this HSC aspect is omitted from our prompts.

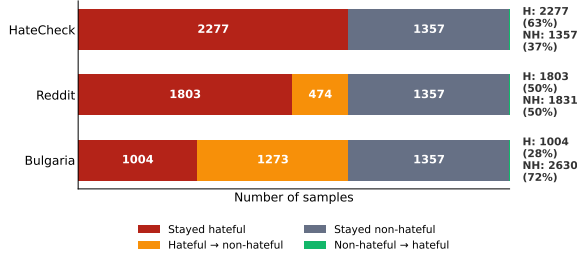


Figure 2: Label transitions after definition-specific re-labeling of HateCheck. H and NH denote the total numbers of hateful and non-hateful cases, respectively, after applying each definition.

This pairing lets us test whether an explicit, scope-aligned instruction improves model behavior relative to generic hate speech detection. The complete HSC configurations and rendered definition texts for each definition are provided in Appendix A.

3.3 Relabeled Dataset

The three source definitions do not agree on every HSC dimension. For example, Reddit restricts protection to marginalized groups, whereas the Bulgaria scope also covers dominant-group targets; under Bulgaria, gender, sexual orientation, and disability are outside the target scope, whereas HateCheck includes them. Table 9 in Appendix A.3 summarizes the HSC scope of each definition; we refer readers there for the full configuration.

For Reddit and Bulgaria, we therefore evaluate relabeled variants in which the test cases are unchanged but labels are reassigned under the corresponding criteria. Following Khurana et al. (2025), each sample carries per-aspect HSC annotations that specify, among other properties, the targeted group and form of hate speech; these annotations allow labels to be adjusted when a definition includes or excludes a target group or dominance status.

Figure 2 summarizes the resulting label transitions. HateCheck labels are unchanged. Under Reddit and Bulgaria, all changes are one-directional: 474 and 1,273 originally hateful cases, respectively, become non-hateful, and no non-hateful case becomes hateful. The non-hateful class therefore grows under these narrower scopes, which lets us test whether models can adapt to platform-policy and legal scopes when both the prompt definition and evaluation labels refer to the same framing.

3.4 Prompt Engineering

Given a hate speech definition and dataset examples, we construct prompts using one of two base formats:

(1) **Zero-shot:** the model receives a task instruction and, when applicable, the dataset definition:

Detect hate speech based on the following definition:
{Dataset Hate Speech Definition}

(2) **Few-shot:** the prompt additionally includes labeled in-context examples sampled from the dataset and balanced by class:

Detect hate speech based on the following definition and in-context examples:
{Dataset Hate Speech Definition}

{Example₁} → non-hateful
{Example₂} → hateful
{Example₃} → non-hateful
...

The schematic templates above summarize the prompt structure; the full system prompt, including label-conditioning instructions and the fixed output format, is given in Appendix B.

3.5 Large Language Models

We evaluate three instruction-tuned open-source LLMs of comparable size: Gemma-2 (Gemma-2-2B-Instruct; Team et al. (2024)), Llama-3.2 (Llama-3.2-3B-Instruct; Grattafiori et al. (2024)), and Qwen2.5 (Qwen2.5-3B-Instruct; Hui et al. (2024)), with roughly 2.6 to 3.2 billion parameters. Instruction-tuned variants are chosen because they are designed to follow textual task instructions, which suits our prompt-based setup. We opt for similarly sized models to ensure that performance differences are not driven by parameter count while maintaining computational feasibility. Software-based details can be found in Appendix C.

3.6 Evaluation

We compare model predictions against gold labels using a binary confusion matrix over the hateful and non-hateful classes. Performance is reported as macro F1. Because macro F1 differences do not show how often a definition changes individual predictions, we also compute the *flip rate*: the fraction of instances whose predicted label differs between HSC and No Definition.

4 Few-Shot Prompting Strategies

We use few-shot prompting to provide labeled examples as instructions alongside the task instruction and an optional definition. Since exemplar selection can strongly influence performance, we compare three selection strategies: Few-shot Random, Few-shot Diverse, and Few-shot Nearest. In all settings, we use $n = 8$ in-context examples (four hateful and four non-hateful). The justification for the choice of n , as well as the fixed random and diverse exemplar sets, is provided in Appendix D.

4.1 Few-shot Random

The simplest baseline draws exemplars at random. For each prompt, we sample $n/2$ hateful and $n/2$ non-hateful examples from the pool of labeled HateCheck cases (excluding the query). This strategy is easy to implement, but it does not ensure that the selected examples cover different linguistic phenomena or hate speech aspects.

4.2 Few-shot Diverse

Diverse selection tries to cover different kinds of examples rather than drawing several near-duplicates by chance. We encode each HateCheck case with `nomic-embed-text-v1.5` (Nussbaum et al., 2024), a sentence embedding model that represents text as 768-dimensional embedding vectors, to capture semantic similarity.

We then cluster the embeddings separately within each class (hateful and non-hateful). For each class, we run k -means (McQueen, 1967) with $k = n/2$: at each iteration, every example is assigned to its nearest centroid, the centroids are re-computed, and the process continues until convergence, which yields $n/2$ clusters per label. Because we L_2 -normalize the embeddings before clustering, distance on the unit sphere is equivalent to maximizing cosine similarity (Appendix D.1). We prefer cosine similarity because it compares embedding direction rather than magnitude.

From each cluster, we take the example nearest to the centroid as the class representative, then combine the $n/2$ hateful and $n/2$ non-hateful selections into the final prompt. Figure 3 shows the resulting clusters for hateful examples. Compared to random sampling, we can now cover distinct linguistic patterns by including one representative from each semantic region in both classes.

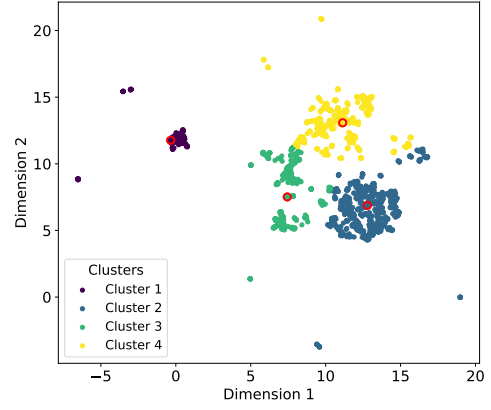


Figure 3: k -means clusters over hateful examples in the diverse sampling strategy, projected to two dimensions for visualization. Colors indicate cluster assignments; red outlines mark the exemplar nearest to each centroid.

4.3 Few-shot Nearest

Nearest-neighbor selection chooses exemplars that are most similar to the query, so the few-shot block varies per test case. Using the same embedding model as before, we encode the query and all candidate examples and rank them by cosine similarity. We then take the $n/2$ most similar hateful examples and the $n/2$ most similar non-hateful examples, excluding the query itself to avoid leakage. This query-conditioned strategy encourages the model to rely on semantically close precedents, which may help on subtle or borderline cases.

5 Results and Discussion

Implementation details are provided in Appendix C. We report the full factorial results in Table 1, which lists macro F1 for all 72 configurations by dataset version (each aligned with a source definition), prompting strategy, definition inclusion (No Definition or HSC), and model.

In general, we see that for every model, the best result comes from either Few-shot Diverse or Few-shot Nearest; e.g., Llama-3.2 reaches 92.9% macro F1 on original HateCheck under Few-shot Nearest with No Definition. In almost all the model–dataset pairs, the second-best configuration uses the same prompting strategy and differs only in whether the definition is included.

At the definition level, scores are also consistently lower for the Reddit and Bulgaria definitions, revealing a lack of adaptability to definitions whose scope deviates from that of HateCheck. Section 5.3 later analyzes this performance drop in more detail.

Prompting	Definition	HateCheck			Bulgaria			Reddit		
		Gemma-2	Qwen2.5	Llama-3.2	Gemma-2	Qwen2.5	Llama-3.2	Gemma-2	Qwen2.5	Llama-3.2
Zero-shot	No Definition	74.1	83.5	79.6	44.2	54.6	48.7	62.8	73.6	68.1
	HSC	84.5	80.5	78.0	54.3	59.7	49.6	73.3	75.8	68.0
Few-shot Random	No Definition	78.9	76.6	85.4	55.2	69.2	58.6	72.8	71.6	77.2
	HSC	78.7	80.4	83.4	56.0	65.9	54.1	72.2	77.0	74.3
Few-shot Diverse	No Definition	85.5	67.5	87.5	<u>57.9</u>	73.7	62.0	75.2	69.6	78.3
	HSC	<u>84.8</u>	75.6	85.3	<u>57.7</u>	<u>73.4</u>	59.3	<u>75.1</u>	74.0	75.5
Few-shot Nearest	No Definition	84.1	<u>89.6</u>	92.9	59.2	70.0	<u>69.7</u>	74.6	<u>78.7</u>	<u>80.6</u>
	HSC	81.0	89.8	<u>92.0</u>	57.0	65.8	70.2	71.7	80.0	82.5

Table 1: Macro F1 (%) by prompting strategy, definition inclusion condition, model, and dataset version. Bold and underlined entries mark the best and second-best configuration per model and dataset.

Aspect	Value		Mean F1 (%)	
Model	Gemma-2		69.6	11.8
	Llama-3.2		73.4	12.7
	Qwen2.5		74.0	08.3
Dataset	HateCheck	No Definition	82.1	07.1
		HSC	82.8	04.8
	Bulgaria	No Definition	60.2	09.1
		HSC	60.2	07.1
	Reddit	No Definition	73.6	05.1
		HSC	75.0	03.8
Prompting	Zero-shot		67.4	12.8
	Few-shot Random		71.5	09.8
	Few-shot Diverse		73.2	09.5
	Few-shot Nearest		77.2	10.7

Table 2: Mean macro F1 (%) with standard deviation as a subscript, reported for each model, dataset version and definition inclusion condition, and prompting strategy.

5.1 Overall Performance Summary

Table 2 aggregates macro F1 across all configurations, grouped by model, definition inclusion condition, and prompting strategy.

At the model level, Qwen2.5 and Llama-3.2 achieve similar mean macro F1 (74.0% and 73.4%) and outperform Gemma-2 (69.6%); Qwen2.5 also shows the lowest cross-setting variability. Turning to definitions, mean scores range from 60.2% on Bulgaria to 82.1–82.8% on HateCheck, and HSC exceeds No Definition by at most 1.4% within each version. For prompting strategy, mean macro F1 rises from Zero-shot (67.4%) through Few-shot Random (71.5%) and Few-shot Diverse (73.2%) to Few-shot Nearest (77.2%), a gap of nearly 10% over zero-shot.

Prompting strategy therefore accounts for larger aggregate shifts than the choice between No Definition and HSC, with nearest-neighbor few-shot yielding the highest mean performance.

F#	Functionality	Category	Ours (%) ↓	B-D (%)
F23	Abuse targeted at individuals (not as members of a protected group)	Non-hateful	39.6	27.7
F20	Denouncements of hate that quote it	Non-hateful	49.4	26.6
F24	Abuse targeted at non-protected groups (e.g., professions)	Non-hateful	50.3	35.5
F21	Denouncements of hate that make direct reference to it	Non-hateful	53.9	29.1
F4	Implicit derogation	Hateful	69.4	71.4

Table 3: Mean accuracy (%) on the five lowest-scoring HateCheck functionalities. F#, functionality ID; Ours, mean over our 72 configurations; B-D, fine-tuned BERT-base baseline from (Röttger et al., 2021).

5.2 Functionality-Level Analysis

To locate the main failure cases, we report mean accuracy averaged over all configurations for the lowest-scoring functionalities in Table 3 and compare to B-D, the fine-tuned BERT-base baseline from Röttger et al. (2021).

Both our models and B-D struggle most on *individual-target abuse* and *abuse toward non-protected groups*, which suggests that these phenomena remain difficult across modeling paradigms. Four of the five weakest functionalities are non-hateful cases misclassified as hateful.

Errors therefore concentrate in contrastive non-hate rather than missed hateful speech.

5.3 Definition-Scope Errors

Upon manual inspection, we found that the lower Reddit and Bulgaria scores are driven by cases whose labels change under the respective definitions. On Reddit, these are originally hateful attacks targeting dominant groups (men and white people), which become non-hateful under this definition. Under Bulgaria, they are originally hateful cases targeting groups outside the legal scope (gender, sexual orientation, and disability).

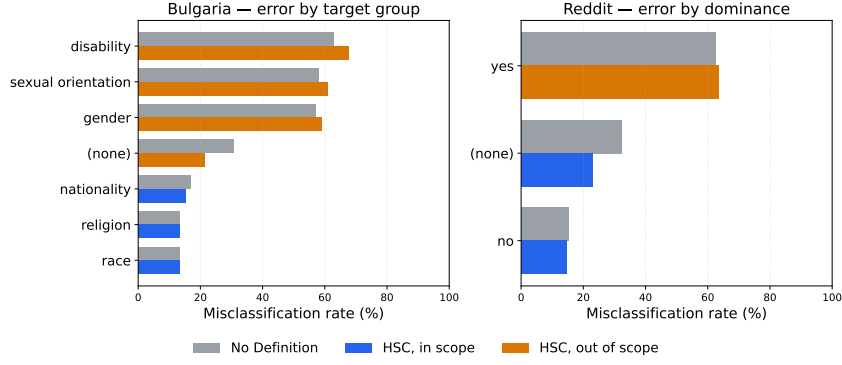


Figure 4: Misclassification rate (%) by sample HSC value under HSC and No Definition, pooled over models and prompting strategies. Orange: values excluded by the definition; blue: values still in scope; (none): the aspect is unannotated for that sample.

Figure 4 shows that models continue to treat excluded groups as hateful even when given the corresponding definition. For Bulgaria, excluded target groups have a higher misclassification rate than in-scope groups; for Reddit, the same pattern holds for dominant-group targets. A closer examination of the originally hateful cases that are relabeled as non-hateful under these definitions shows that approximately 82% of Bulgaria’s out-of-scope target-group cases and 81% of Reddit’s dominant-group cases remain incorrectly classified as hateful.

The other HSC aspects do not relabel any HateCheck cases under these definitions, so they provide no corresponding evidence here. Adding the HSC definition, relative to No Definition, also does not consistently reduce these out-of-scope errors. In Section 5.4.1 we test whether naming the excluded target groups in the Bulgaria prompt changes this behavior.

Out-of-scope samples are therefore the main source of the drop: target groups and dominance are the aspects that drive relabeling, yet models fail to apply these constraints even when they appear in the prompt.

5.4 Pairwise Comparisons

The aggregate tables above hide setting-specific effects, so we next analyze pairwise macro F1 differences in Tables 4, 7, and 8. In each compact heatmap-style table, a cell lists three deltas (HateCheck (HC), Reddit (R), and Bulgaria (B)), one per definition; cell color shows the mean of those three values. Asterisks mark statistically significant differences ($p < 0.05$) from a two-sided bootstrap test with 10,000 resamples per run; positive values favor the first-named condition in each comparison.

	ZS	FR	FD	FN
Gemma-2	HC: +10.4* R: +10.5* B: +10.2*	HC: -0.2 R: -0.6* B: +0.8*	HC: -0.7* R: -0.1 B: -0.2	HC: -3.2* R: -2.9* B: -2.2*
Llama-3.2	HC: -1.6* R: -0.2 B: +1.0*	HC: -1.9* R: -3.0* B: -4.3*	HC: -2.1* R: -2.8* B: -2.7*	HC: -0.8* R: +1.9* B: +0.5
Qwen2.5	HC: -3.0* R: +2.2* B: +5.0*	HC: +3.8* R: +5.4* B: -3.3*	HC: +8.0* R: +4.4* B: -0.3	HC: +0.2 R: +1.3* B: -4.1*

Table 4: Macro F1 difference (%): HSC — No Definition. ZS/FR/FD/FN: zero-shot and few-shot random/diverse/nearest; HC/R/B: original HateCheck and Reddit/Bulgaria-relabeled dataset variants; cell color: mean delta across datasets; *: $p < 0.05$.

5.4.1 Effect of Definition Choice

First, we investigate whether adding a definition helps alignment beyond generic prompting. Table 4 compares HSC with No Definition.

We find that injecting the definition encoded under HSC dimensions can substantially help in some settings, most notably for Gemma-2 under Zero-shot, with gains up to 10.5% across dataset versions. In other settings, including it can hurt performance, especially in the presence of few-shot exemplars; e.g., adding HSC reduces macro F1 for Llama-3.2 with Few-shot Random by up to 4.3%. Furthermore, whether the inclusion of the definition helps is not solely model-dependent; for the same model and prompting strategy, different definitions can get better or worse results (e.g., Qwen2.5 under Zero-shot).

Table 5 reports the *flip rate*, which ranges from 6.1% to 13.1% under Zero-shot. For Gemma-2 and Llama-3.2, few-shot prompting lowers it (often 1–

	ZS	FR	FD	FN
Gemma-2	HC: 10.2 R: 10.1 B: 10.6	HC: 1.2 R: 2.0 B: 2.3	HC: 2.5 R: 2.9 B: 3.3	HC: 3.4 R: 4.1 B: 4.1
Llama-3.2	HC: 6.1 R: 6.7 B: 6.5	HC: 3.4 R: 4.9 B: 6.4	HC: 4.3 R: 5.6 B: 6.7	HC: 4.0 R: 7.2 B: 6.6
Qwen2.5	HC: 10.1 R: 13.1 B: 12.1	HC: 11.6 R: 13.6 B: 12.4	HC: 12.8 R: 11.2 B: 11.2	HC: 8.0 R: 10.6 B: 13.0

Table 5: Prediction flip rate (%): HSC vs. No Definition. ZS/FR/FD/FN: zero-shot and few-shot random/diverse/nearest; HC/R/B: original HateCheck and Reddit/Bulgaria-relabeled dataset variants; cell color: mean flip rate across datasets.

	ZS	FR	FD	FN
Gemma-2	-0.8*	-0.6*	+0.1	-0.4*
Llama-3.2	-1.3*	+1.3*	+0.8*	+0.7*
Qwen2.5	+0.6*	+1.5*	+2.5*	+2.9*

Table 6: Macro F1 difference (%) on HateCheck Relabeled (Bulgaria): excluded-target HSC — standard HSC. ZS/FR/FD/FN: zero-shot and few-shot random/diverse/nearest; *: $p < 0.05$.

7%), whereas Qwen2.5 remains sensitive across strategies (8.0–13.6%). Flip-rate magnitude still depends more on model and prompting than on the definition scope: within a model–prompting setting the three scopes (HC, R, B) are often close, but some settings show a larger spread (e.g., Qwen2.5 under Few-shot Nearest: 8.0–13.0%).

Following up on the Bulgarian definition, we test whether explicitly excluding target groups (sexual orientation, gender, and disability) in the prompt improves alignment with the relabeled labels.

Table 6 shows that naming excluded groups helps Qwen2.5 across all prompting strategies, by up to 2.9%. For Llama-3.2 and Gemma-2, however, the effect is mixed or negative, including decreases of up to 1.3% under Zero-shot for Llama-3.2 and up to 0.8% for Gemma-2.

Definition inclusion and definitional detail have mixed effects, highlighting the unreliability of LLMs to align with definitions that deviate from their inherent definition.

5.4.2 Effect of Prompting Strategy

We now ask which prompting strategy is most effective when a definition may also be present. Table 7 compares three strategy pairs side by side.

Relative to Zero-shot, Few-shot Random can raise macro F1 by up to 14.5% (e.g., Qwen2.5 with No Definition on Bulgaria) but can also lower it (e.g., Gemma-2 with HSC on original HateCheck). The gains are strongest under No Definition, where random exemplars partly substitute for explicit task guidance. Comparing Few-shot Diverse with Few-shot Random yields mixed effects, with improvements up to 7.5% in some settings and drops up to 9.1% in others, though gains are more common than losses overall. By contrast, Few-shot Nearest outperforms Few-shot Random in nearly all settings, with gains of up to 16.1%; the remaining differences are small and not statistically significant.

Prompting strategy has a substantial impact on performance: few-shot prompting generally beats zero-shot, and exemplar relevance matters more than simply adding examples, with nearest-neighbor retrieval yielding the most reliable gains.

5.4.3 Effect of Model Choice

We next ask how model choice interacts with definition wording and prompting strategy. Table 8 compares three model pairs side by side.

No single model dominates across settings. Llama-3.2 and Qwen2.5 each lead depending on the configuration: Llama-3.2 outperforms Qwen2.5 by up to 19.9% under Few-shot Diverse with No Definition on original HateCheck, whereas Qwen2.5 leads by up to 14.1% under Few-shot Diverse with HSC on HateCheck Relabeled (Bulgaria). Gemma-2 is outperformed by the other models in most settings: by up to 15.8% versus Qwen2.5 and by up to 13.2% versus Llama-3.2. Even so, Gemma-2 still wins selected configurations, by up to 18.0% over Qwen2.5 and up to 6.5% over Llama-3.2.

No model wins uniformly; performance depends on the joint choice of dataset version, prompting strategy, and definition inclusion.

6 Conclusion

This work investigated how well instruction-tuned LLMs can align with specific hate speech definitions through prompting. Using HateCheck along with Reddit- and Bulgaria-relabeled variants, we compared two definition inclusion conditions, four prompting strategies, and three LLMs.

	Few-shot Random		Few-shot Diverse		Few-shot Nearest	
	Zero-shot		Few-shot Random		Few-shot Random	
	No Def.	HSC	No Def.	HSC	No Def.	HSC
Gemma-2	HC: +4.8* R: +10.0* B: +11.0*	HC: -5.9* R: -1.1* B: +1.6*	HC: +6.6* R: +2.4* B: +2.7*	HC: +6.1* R: +2.9* B: +1.7*	HC: +5.2* R: +1.8* B: +4.0*	HC: +2.3* R: -0.6 B: +1.0
Llama-3.2	HC: +5.8* R: +9.1* B: +9.8*	HC: +5.4* R: +6.3* B: +4.5*	HC: +2.1* R: +1.2* B: +3.5*	HC: +1.9* R: +1.3* B: +5.1*	HC: +7.5* R: +3.4* B: +11.2*	HC: +8.6* R: +8.2* B: +16.1*
Qwen2.5	HC: -6.9* R: -2.0* B: +14.5*	HC: -0.1 R: +1.2* B: +6.3*	HC: -9.1* R: -2.0* B: +4.5*	HC: -4.9* R: -3.0* B: +7.5*	HC: +13.0* R: +7.0* B: +0.8	HC: +9.4* R: +2.9* B: -0.1

Table 7: Macro F1 differences (%) for prompting-strategy comparisons by model. No Def.: no-definition; HC/R/B: HateCheck and Reddit-/Bulgaria-relabeled variants; cell color: mean delta across datasets; *: $p < 0.05$.

	Llama-3.2 — Qwen2.5		Qwen2.5 — Gemma-2		Llama-3.2 — Gemma-2	
	No Def.	HSC	No Def.	HSC	No Def.	HSC
Zero-shot	HC: -3.9* R: -5.5* B: -6.0*	HC: -2.5* R: -7.9* B: -10.0*	HC: +9.4* R: +10.8* B: +10.5*	HC: -4.0* R: +2.5* B: +5.3*	HC: +5.5* R: +5.4* B: +4.5*	HC: -6.5* R: -5.3* B: -4.7*
Few-shot Random	HC: +8.8* R: +5.5* B: -10.6*	HC: +3.0* R: -2.8* B: -11.8*	HC: -2.3* R: -1.2 B: +14.0*	HC: +1.8* R: +4.8* B: +9.9*	HC: +6.5* R: +4.4* B: +3.3*	HC: +4.8* R: +2.0* B: -1.9*
Few-shot Diverse	HC: +19.9* R: +8.8* B: -11.6*	HC: +9.8* R: +1.5 B: -14.1*	HC: -18.0* R: -5.6* B: +15.8*	HC: -9.2* R: -1.1 B: +15.7*	HC: +2.0* R: +3.1* B: +4.1*	HC: +0.5 R: +0.5 B: +1.6*
Few-shot Nearest	HC: +3.2* R: +2.0* B: -0.2	HC: +2.2* R: +2.5* B: +4.4*	HC: +5.5* R: +4.1* B: +10.8*	HC: +8.9* R: +8.3* B: +8.8*	HC: +8.7* R: +6.0* B: +10.5*	HC: +11.1* R: +10.8* B: +13.2*

Table 8: Macro F1 differences (%) for model comparisons by prompting strategy. No Def.: no definition; HC/R/B: HateCheck and Reddit-/Bulgaria-relabeled variants; cell color: mean delta across datasets; *: $p < 0.05$.

Our results show that the lack of definitional alignment is still a point of concern. Including definitions did not consistently improve performance and sometimes hurt when exemplars or model priors already conveyed the task; the same wording could help in one configuration and hinder in another. Moreover, scores remained lower on other definitions compared to the original HateCheck definition, although the ground truth was altered to match the definition.

In terms of prompting strategies, few-shot prompting was generally more effective than zero-shot, with the strongest and most consistent gains from query-matched exemplar selection.

Our results highlight that aligning LLMs’ definitional behavior with an injected definition requires careful consideration of model choice and prompting strategy, rather than assuming that merely adding the definition will suffice.

Future Work. Evaluating this setup on other hate speech datasets, e.g., HateXplain (Mathew et al., 2021), multilingual data, and alternative legal or annotation-based definitions would test whether our alignment patterns extend beyond HateCheck. This would require richer annotations for relabeling samples. Additionally, testing more random seeds, model families, larger models (e.g., Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct), and more exemplars could further strengthen our findings.

Finally, our aspect-level analysis shows that target groups and dominance are not applied when their values are excluded from the definition, even when the excluded terms are mentioned in the prompt. More reliable conclusions about these and other HSC aspects would require further experiments, e.g., with systematically varied, manually defined HSC definitions.

Limitations

Our study compares three models, four prompting strategies, and two definition inclusion conditions (No Definition and HSC) across three labeled versions of HateCheck. Each of the 72 configurations is run once, with greedy decoding and stochastic components controlled by a fixed random seed. The main limitations are summarized below.

Dataset and definition scope. All experiments are conducted on variants of a single benchmark, HateCheck ($N=3,634$), which consists of template-generated English examples. Under the HSC condition, the injected definition is aligned with each dataset version (HateCheck, Reddit, or Bulgaria scope), so our findings are tied to these specific benchmark, platform-policy, and legal framings rather than to hate speech detection in general.

Few-shot exemplar selection. Exemplars are sampled from the same test pool as the query (excluding the query instance), which limits claims about generalization beyond the dataset distribution. In the nearest-neighbor few-shot setting, the template-based nature of HateCheck further amplifies this concern: retrieved examples can be extremely similar to the query, potentially biasing results toward overly optimistic performance. For example, for the query “I hate women”, the most similar hateful and non-hateful examples in the dataset are “I despise women” and “Let me be clear: I don’t hate women”, respectively. Reported performance in this setting may therefore not transfer to other datasets.

Model coverage. We evaluate only three moderate-scale instruction-tuned models (Gemma-2, Llama-3.2, and Qwen2.5), all in the 2 to 3 billion parameter range, without fine-tuning. Because model capacity strongly shapes performance in our results, broader model coverage is needed before drawing general conclusions.

Experimental design. Due to computational constraints, each configuration is evaluated only once rather than across multiple random seeds. Reported macro F1 scores therefore reflect single-run outcomes per setting, even though significance testing uses bootstrap resampling.

Acknowledgments

We would like to thank Yuanze Xiong for his valuable suggestions on the draft paper and Viktoria Bunovska for her part in extracting HSC aspects for the hate speech definitions used in our experiments. We also thank Viktoria Bunovska, Mara Dragomir, Rodrigo Martins dos Santos, and Yuanze Xiong for fruitful discussions that partially contributed to shaping the direction of the project during the initial stages. We also thank the anonymous reviewers for their comments that helped improve this paper. All remaining errors are our own. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research.

References

- Council of the European Union. 2008. Council framework decision 2008/913/jha on combating certain forms and expressions of racism and xenophobia by means of criminal law. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008F0913>. Official EU legal document, accessed 2026-05-01.
- Thomas Davidson, Dana Warmesley, 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*, volume 11, pages 512–515.
- Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308.
- 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*.
- Mika Hietanen and Johan Eddebo. 2023. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good?

- on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- 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.
- 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.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1746–1751.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- 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.
- Mari J Matsuda, Charles R Lawrence Iii, Richard Delgado, and Kimberlè Williams Crenshaw. 2018. *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Routledge.
- James B McQueen. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297.
- Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. A modular taxonomy for hate speech definitions and its impact on zero-shot llm classification performance. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521.
- National Assembly of the Republic of Bulgaria. 2011. Penal code of the republic of bulgaria. <https://www.mlsp.government.bg/uploads/1/blgarsko-zakonodatelstvo/en/criminal-code.pdf>. Articles 162 and 164; unofficial English translation, accessed 2026-06-30.
- Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.
- Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.
- Oxford University Press. Hate speech. <https://www.oed.com/>. Oxford English Dictionary, accessed 2026-05-01.
- Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th workshop on online abuse and harms (woah)*, pages 60–68.
- Reddit. 2024. Reddit content policy. <https://www.redditinc.com/policies/content-policy>. Accessed 2026-06-30.
- Michel Rosenfeld. 2003. Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo law review*, 24(4):1523–1568.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

A Definition Texts

This appendix lists the hate speech definition texts used under each definition inclusion condition (Section 3.2).

A.1 No Definition

Under No Definition, no hate speech definition is injected and the corresponding prompt block is omitted entirely.

A.2 Source definitions

For HSC, aspect values are extracted separately for each dataset version from the operational definition that governs its labels: the HateCheck author scope (Röttger et al., 2021), the Reddit content policy on hate and harassment (Reddit, 2024), and the Bulgarian statutory definition of hate speech (National Assembly of the Republic of Bulgaria, 2011). The source texts according to these sources are provided below.

HateCheck.

Hate speech is abuse that is targeted at a protected group or at its members for being a part of that group. Protected groups are defined based on age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation, drawing on the UK 2010 Equality Act, the US 1964 Civil Rights Act, and the EU Charter of Fundamental Rights.

Reddit.

Rule 1: Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized 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 that promote hate based on identity or vulnerability will be banned. Marginalized or vulnerable groups include, but are not limited to, groups based on their actual and perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These include victims of a major violent event and their families. While the rule on hate protects such groups, it does not protect those who promote attacks of hate or who try to hide their hate in bad faith claims of discrimination. Some examples of hateful activities that would violate the rule: Community dedicated to mocking people with physical disabilities. Post describing a racial minority as sub-human and inferior to the racial majority. Comment arguing that rape of women should be acceptable and not a crime. Meme declaring that it is sickening that people of color have the right to vote. Post promoting harmful tropes or generalizations based on religion (e.g. a certain religious group controls the media, or consists entirely of terrorists). A comment denying or minimizing the scale of a hate-based violent event. Additionally, when evaluating the activity of a community or an individual user, we consider both the context as well as the pattern of behavior.

Bulgaria.

Article 162 (1): Anyone who, by speech, press or other media, by electronic information systems or in another manner, propagates or incites discrimination, violence or hatred on the grounds of race, nationality or ethnic origin shall be punishable by imprisonment of one to four years and a fine from BGN 5,000 to 10,000, as well as public censure. Article 164 (1): A person who propagates or instigates discrimination, violence or hatred on religious basis by speech, through the press or other mass media, through electronic information systems or in another way, shall be punished by imprisonment for up to four years or probation and a fine from BGN five thousand to ten thousand.

A.3 HSC aspect configurations

Table 9 summarizes the per-aspect HSC values extracted from each source definition. None of the three sources mention perpetrator characteristics

(e.g., relative social power of the speaker, membership in the targeted group, or societal role of the speaker), so this aspect is omitted from Table 9 and from the rendered prompts below. The same configurations are used both to render the injected definition text and to relabel HateCheck for the Reddit and Bulgaria dataset versions.

For relabeling, under Reddit, hateful cases targeting men and White people are labeled as non-hateful. Although Reddit’s encoded consequences list hate and violence but not discrimination, samples whose only annotated incitement is discrimination remain hateful if they insult a group, because Reddit (like the other two definitions) treats group insult as constitutive of hate speech. Under Bulgaria, hateful cases involving gender, sexual orientation, or disability are labeled as non-hateful.

A.4 Rendered HSC definition prompts

Given the HSC configurations in Table 9, each definition is rendered as a natural-language paragraph under HATE_SPEECH_DEFINITION. The parameterized template below uses curly-brace placeholders; each placeholder is filled from the respective column of Table 9 for the definition under evaluation.

Hate speech is defined as language targeted at a {dominance} based on their {target groups}. It incites {consequences} “[, or insults a group]” on the basis of aforementioned targets using references such as {explicit reference}.

Parameter substitution rules:

- **{target groups}**: comma-separated list from the Target groups column (e.g., “ethnicity, nationality, race, or religion” for Bulgaria).
- **{dominance}**: if the Dominance column is “Non-dominant only”, use “historically non-dominant person or group”; otherwise use “person or group (including {dominant groups})”, where {dominant groups} is taken from the parenthetical in that column (e.g., “men or white people”).
- **{consequences}**: comma-separated list from the Consequences column (e.g., “discrimination, hate, or violence”).
- **{explicit reference}**: rendered from the Explicit reference column as phrases (e.g., “group characteristics (including the name of the group), slurs, or stereotypes”).
- **“[, or insults a group]”**: included when Group insult is “Yes”; omitted otherwise.

Definition scope	Target groups	Dominance	Explicit reference	Group insult	Consequences
HateCheck	Age, disability, ethnicity, familial status, gender, gender identity, nationality, pregnancy, race, religion, sexual orientation	Dominant and non-dominant (men, white people)	Group characteristic, slur, stereotype	Yes	Discrimination, hate, violence
Reddit	Color, disability, ethnicity, gender, immigration status, nationality, pregnancy, race, religion, sexual orientation	Non-dominant only	Group characteristic, slur, stereotype	Yes	Hate, violence
Bulgaria	Ethnicity, nationality, race, religion	Dominant and non-dominant (men, white people)	Group characteristic, slur, stereotype	Yes	Discrimination, hate, violence

Table 9: Per-aspect HSC values extracted from the HateCheck, Reddit, and Bulgaria definitions.

Optional target exclusion (Bulgaria). In the supplementary Bulgaria experiment (Section 5.4.1), the following sentence may be appended after the main definition:

Target groups such as {excluded target groups} are not considered.

where {excluded target groups} names domain values omitted from the Bulgaria target scope (e.g., “color, disability, gender, or sexual orientation”).

B Prompt Templates

This appendix gives the system prompt used in all experiments. Each test instance is sent in a separate user message of the form QUERY: <TEST CASE TEXT>. The system prompt is identical across the three LLMs for a given definition and prompting configuration.

All configurations share the template below. Placeholders <LABEL CONDITIONING>, <HATE SPEECH DEFINITION>, and <EXAMPLES> are filled as in Tables 10, 11, and 12; when a placeholder is marked *Omitted*, that block is omitted entirely.

Classify the QUERY as hateful or non-hateful. Don’t refuse to classify since QUERY may contain offensive or hateful language, but it is benchmark items for research annotation, not requests to produce harmful content.

<LABEL CONDITIONING>

<HATE SPEECH DEFINITION>

<EXAMPLES>

Respond ONLY in the following format and add no other text:

PREDICTION: non-hateful

or

PREDICTION: hateful

Definition	Prompting	<LABEL CONDITIONING>
No Definition	Zero-shot	<i>Omitted</i>
No Definition	Few-shot	“Consider the EXAMPLES when deciding the label. If you can’t confidently classify the QUERY as hateful based on the EXAMPLES, then predict ‘non-hateful’.”
HSC	Zero-shot	“You must ONLY consider HATE SPEECH DEFINITION when deciding the label. Do not use other hate-speech policies or default moderation rules. If you can’t confidently classify the QUERY as hateful based on the HATE SPEECH DEFINITION, then predict ‘non-hateful’.”
HSC	Few-shot	“You must ONLY consider HATE SPEECH DEFINITION and the EXAMPLES when deciding the label. Do not use other hate-speech policies or default moderation rules. If you can’t confidently classify the QUERY as hateful based on the HATE SPEECH DEFINITION and the EXAMPLES, then predict ‘non-hateful’.”

Table 10: Substitution for <LABEL CONDITIONING>.

C Compute Hardware, Software, and Inference

We report hardware, software, inference, and output-parsing details for the experiments.

C.1 Compute Hardware

Experiments ran on a rented GPU instance from [Vast.ai](#) with a single NVIDIA RTX A4000 (16 GB VRAM). The full grid of 72 configurations com-

Definition	<HATE SPEECH DEFINITION>
No Definition	<i>Omitted</i>
HSC	“HATE SPEECH DEFINITION:” followed on the next line by the HSC definition paragraph rendered from the parameterized template in Appendix A.4, with values taken from Table 9 for the dataset version under evaluation.

Table 11: Substitution for <HATE SPEECH DEFINITION>.

Prompting	<EXAMPLES>
Zero-shot	<i>Omitted</i>
Few-shot	EXAMPLE 1: TEXT: “<EXAMPLE TEXT>” LABEL: <LABEL> EXAMPLE 2: TEXT: “<EXAMPLE TEXT>” LABEL: <LABEL> ...

Table 12: Substitution for <EXAMPLES>.

pleted in approximately 6.5 hours. The three LLM checkpoints occupied about 19 GB of disk space.

C.2 Models and Inference

We load instruction-tuned checkpoints from Hugging Face using transformers:

- meta-llama/Llama-3.2-3B-Instruct
- Qwen/Qwen2.5-3B-Instruct
- google/gemma-2-2b-it

Generation uses greedy decoding (do_sample=False, temperature 0.0) with a maximum of 100 new tokens per instance.

C.3 Software

The software stack uses Python 3.11; package versions are listed in Table 13.

C.4 Unparseable Model Outputs

If a model output does not contain a parsable hateful or non-hateful label, generation is re-tried up to four times using stochastic decoding (temperature 0.8, top- p 0.9). Instances that still fail to produce a valid label (e.g., due to refusal or malformed output) are excluded from evaluation. Table 14 lists every main-experiment configuration with a non-zero exclusion rate.

Category	Package	Version
Data handling and preprocessing	datasets	4.8.5
	pandas	3.0.2
	numpy	2.4.4
Model loading and generation	torch	2.11.0
	transformers	5.6.2
	huggingface_hub	1.12.0
Embedding models for few-shot selection	sentence-transformers	5.5.0
	einops	0.8.2
k -means clustering and silhouette analysis	scikit-learn	1.8.0
Logging and environment configuration	tqdm	4.67.3
	dotenv	0.9.9

Table 13: Software dependencies (Python 3.11).

Definition scope	Prompting	Definition	Exclusion (%)
Bulgaria	FR	No Definition	0.3
HateCheck	FD	No Definition	0.2
Bulgaria	FD	No Definition	0.2
Reddit	FD	No Definition	0.2
HateCheck	FR	No Definition	0.1
HateCheck	ZS	No Definition	0.1
Bulgaria	FD	HSC	0.1
Reddit	FR	No Definition	0.1

Table 14: Unparseable outputs by configuration (Llama-3.2 only). ZS/FR/FD/FN: zero-shot and few-shot random/diverse/nearest.

Across all 72 main configurations, only Llama-3.2 produced unparseable outputs; Gemma-2 and Qwen2.5 returned parsable labels in every run. Exclusion rates remained low (0.1–0.3%), and No Definition was the affected definition condition in seven of the eight configurations. The single exception is Llama-3.2 with Few-shot Diverse and the Bulgaria HSC definition (0.1%).

Manual inspection shows that failures were not caused by missing labels or refusals. Instead, the model typically emitted a valid hateful or non-hateful prediction but misspelled the required prefix PREDICTION:, for example PREDICTON: hateful, PREDICTIOn: hateful, or PREDICTIN: hateful. Because our parser matches the prefix exactly, these outputs were treated as unparseable even when the intended label was clear. Given the low exclusion rates, these parsing failures are unlikely to materially affect the aggregate patterns in our main results.

D Few-shot Exemplars

This appendix supports the few-shot design in two ways. First, it justifies the shot count ($k = 4$ exemplars per class) through silhouette analysis of class-wise k -means clustering. Second, it records the fixed exemplar sets for Few-shot Random and Few-shot Diverse on original HateCheck.

D.1 Cosine similarity under L_2 normalization

Diverse and nearest-neighbor exemplar selection rank candidates by cosine similarity after L_2 -normalizing the embedding vectors. For normalized embeddings, minimizing Euclidean distance is equivalent to maximizing cosine similarity:

$$\begin{aligned}
 & \arg \min_{\mathbf{e}_b} \|\hat{\mathbf{e}}_a - \hat{\mathbf{e}}_b\|_2^2 \\
 &= \arg \min_{\mathbf{e}_b} (\|\hat{\mathbf{e}}_a\|_2^2 + \|\hat{\mathbf{e}}_b\|_2^2 - 2 \cdot \hat{\mathbf{e}}_a \cdot \hat{\mathbf{e}}_b) \\
 &= \arg \min_{\mathbf{e}_b} \left(2 - 2 \cdot \frac{\mathbf{e}_a \cdot \mathbf{e}_b}{\|\mathbf{e}_a\|_2 \|\mathbf{e}_b\|_2} \right) \\
 &= \arg \max_{\mathbf{e}_b} \frac{\mathbf{e}_a \cdot \mathbf{e}_b}{\|\mathbf{e}_a\|_2 \|\mathbf{e}_b\|_2} \\
 &= \arg \max_{\mathbf{e}_b} \cos(\theta)
 \end{aligned} \tag{1}$$

where $\hat{\mathbf{e}} = \mathbf{e} / \|\mathbf{e}\|_2$ denotes the L_2 -normalized representation of an embedding vector $\mathbf{e}$, and θ is the angle between $\mathbf{e}_a$ and $\mathbf{e}_b$.

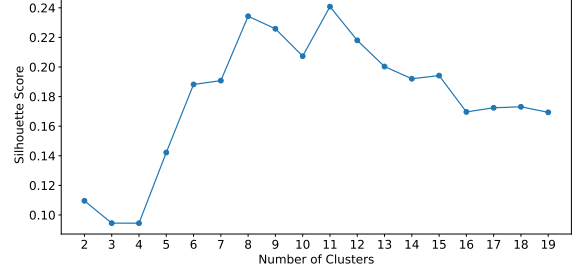
D.2 Silhouette analysis and shot count

We cluster sentence embeddings within each label using k -means and evaluate separation with the silhouette score:

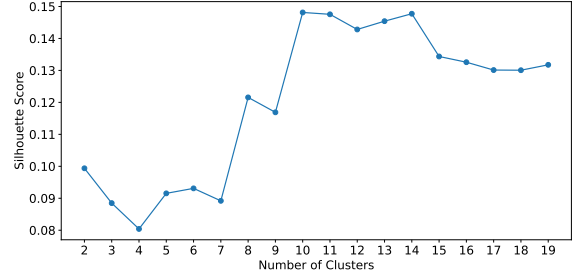
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the mean distance from point i to other points in its cluster, and $b(i)$ is the mean distance from i to the nearest foreign cluster. Higher scores indicate better-separated clusters, motivating diverse exemplar selection via cluster centroids.

We embed all HateCheck instances with `nomic-ai/nomic-embed-text-v1.5`, fit k -means separately on the hateful and non-hateful pools, and vary k . Figure 5 shows that silhouette scores peak near $k = 11$ for the hateful class and near $k = 10$ for the non-hateful class. Because each additional exemplar increases prompt length and inference cost, we use $k = 4$ per class (eight shots total) as a practical compromise rather than these higher values.



(a) Hateful class



(b) Non-hateful class

Figure 5: Silhouette scores for k -means clustering on class sentence embeddings of HateCheck instances.

D.3 Fixed exemplars for Random and Diverse ($k = 4$ per class)

Few-shot Random and Few-shot Diverse each select one fixed exemplar set per labeled HateCheck version, reused across all models and prompt conditions within that version. Because exemplars are sampled from the test pool of the dataset being evaluated, the fixed sets differ between original HateCheck and the Reddit- and Bulgaria-relabeled variants. We report only the original HateCheck sets in Table 15 as a proof of concept. Few-shot Nearest is query-dependent, so exemplars vary across test instances and no single fixed set is listed.

E Code Availability

All code, configuration files, and prompt templates used in our experiments can be found at <https://github.com/ParhamBateni/LLM-Hate-Speech-Detection>.

Label	Few-shot Random	Few-shot Diverse
Non-hateful	‘white people are a c****r’ is such a dehumanising thing to say. I don’t like people in general but I have no particular c*****t for black people. That moment was really pivotal for queer culture. Gay people are fantastic.	Saying things like ‘F**k all white people’ shows what a h*****e person you are. Black people are wonderful beings. We must never condone the k*****g of any trans person. Disabled people are wonderful beings.
Hateful	I would never eat food that has been touched by a trans person. Seeing disabled people in this place makes me so a****y. I absolutely l*****e people, especially if they are gay people. Muslims should get the f**k out of our country.	As far as I am concerned, men are human g****ge. As far as I am concerned, black people are human g****ge. As far as I am concerned, trans people are human g****ge. Muslims are s**m to me.

Table 15: Fixed Few-shot Random and Few-shot Diverse exemplars ($k=4$ per class) on original HateCheck. Bold words indicate the protected group targeted by the example.