The Alignment of Large Language Models' Responses to Subjective Variations in Hate Speech

Comparing Alignment to Real-Life-Inspired Definitions in Zero-Shot Hate Speech Classification

Viktoria Bunovska[1]

Supervisors: Pradeep Murukannaiah [1]**, Urja Khurana**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

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 20, 2026

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

Abstract

Detecting hateful content on social media has become an active area of research, with recent approaches focusing on the use of Large Language Models (LLMs). Rather than using datasets to train classifiers, researchers are exploring methods that embed hate speech definitions directly in the model's prompt. However, hate speech is a subjective concept, and its definition varies across contexts. As a result, LLMs must align their classifications with the specific definition provided in the prompt. To make the creation process more systematic, frameworks for constructing context-specific definitions of hate speech have been proposed. Yet, no work has compared how framework-based formulations influence LLM alignment relative to the definitions used in real-life regulation, such as laws and social media policies. This study, therefore, compares definitions from the Hate Speech Criteria (HSC) framework, legal texts, and platform policies by evaluating how precisely two LLMs align with each type under a zero-shot prompting setup. Our results indicate that while the level of alignment is model-dependent, legal and policy definitions generally guide LLM behavior more effectively than framework-based formulations. Nevertheless, definitions created with the framework still steer models in the intended direction, suggesting that further refinement of these frameworks could improve their effectiveness in prompt-based hate speech detection.

Warning: This paper discusses hate speech and contains examples of potentially offensive language.

1 Introduction

The rapid growth of social media has made hate speech increasingly prevalent (Reichelmann et al., 2021). Research shows that exposure to hateful content can negatively affect mental health, as it has been linked to PTSD symptom severity (Shmulewitz et al., 2025). As a result, accurately identifying hate speech has become an important task for social media platforms.

Automated hate speech detection has become necessary due to the scale of online content moderation. Earlier approaches relied on machine learning models trained on annotated datasets. However, these datasets are limited by sampling and annotation bias (Khurana et al., 2025), poor generalization to new data (Fortuna et al., 2020), and internal inconsistencies (Fortuna et al., 2020; Awal et al., 2020). To reduce the need for task-specific training data, researchers have started exploring the use of large language models (LLMs) for hate speech detection. Rather than learning the task from a dedicated dataset, LLM-based approaches can classify content by incorporating hate speech definitions directly into prompts. This can be effectively achieved through zero-shot prompting, where the model evaluates text using only the definition without relying on labeled examples (Plaza-del arco et al., 2023; Melis et al., 2025). But what exactly is the definition of hate speech?

Hate speech is a subjective concept, and its interpretation varies across individuals and contexts (Fortuna et al., 2020). Therefore, no universally accepted definition of hate speech exists, and thus models must align their responses with the one provided in the prompt. This definitional variation has important implications for NLP research. Specifically, commonly used definitions in research have been shown to diverge from the platform guidelines and legal texts applied in practice (Rizwan et al., 2025), raising concerns about whether findings from research settings generalize to real-world moderation contexts. This study, therefore, focuses on practical applications and incorporates definitions drawn from three real-world-inspired sources. Particularly, we look at national laws, social media platform policies, and a recently proposed framework-based approach grounded in law and social science.

Each of the three sources we explore reflects different goals and characteristics. First, hate speech laws are defined by jurisdictional terms and tone, which can cause inconsistent interpretation by annotators (Korre et al., 2024). In contrast, social media platform policies need to be understandable and unambiguous due to their broad audience[1]. The third definition type is created manually using theoretical frameworks that standardize commonly encountered components in real-life definitions. These frameworks address the subjectivity of hate speech by enabling definitions to be constructed from key components (e.g., targets, actors, or expressions) suited to the task and context (Khurana et al., 2022). Definitions derived from these frameworks are referred to as *theoretical*, *theoretically-created* or *framework-based* definitions.

Previous studies have evaluated LLM hate speech classification using framework-based definitions (Khurana et al., 2025; Melis et al., 2025). However, no work has directly compared framework-based definitions with the legal and platform-policy formulations used in practice to evaluate their effects on LLM alignment.

This gap raises the main question of this study: How well do LLMs align their hate speech classifications with different types of real-world-inspired definitions in a zero-shot setting?

We divide it into three sub-questions:

- RQ1: Do framework-based definitions lead to similar levels of LLM alignment as legal and social media policy definitions?
- RQ2: Do broader or more concrete real-world-inspired definitions better support LLM alignment?
- RQ3: Do samples associated with the hate speech components present in a real-world-inspired definition benefit from including that definition in the prompt?

In this study, we investigate the classification behavior of two LLMs, `Flan-T5-XL` and `Qwen2.5-3B-Instruct`, when provided with six different hate speech definitions in a zero-shot prompting setting. We find that alignment varies across both models and definition types. Legal and platform-policy definitions generally support stronger alignment than theoretically-

[1]Article 14, Terms and conditions - the Digital Services Act (DSA), https://www.eu-digital-services-act.com/Digital_Services_Act_Article_14.html

created ones. In contrast, differences between broad and concrete definitions are limited, while the effects of individual hate speech components vary across models. Nevertheless, the results suggest that definitions constructed from frameworks still hold potential. Thus, further refinement of framework guidelines may improve their effectiveness in guiding LLM-based classification.

2 Related Work

LLM-based hate speech detection Recent research has begun investigating LLMs for hate speech detection by providing definitions in prompts, thereby reducing the dependence on annotated datasets (Roy et al., 2023). Plaza-del arco et al. (2023) demonstrate that both model choice and prompt formulation significantly affect performance; however, they do not provide definitions to their zero-shot prompts. Roy et al. (2023) find that adding context and target group specifications can improve performance, but the effect is inconsistent across models and datasets. Closest to our work, Melis et al. (2025) show that definition content and specificity affect zero-shot classification differently depending on the model architecture. However, their study does not mention relabeling the dataset according to each definition to catch potential alignment differences. Furthermore, it focuses exclusively on theoretically constructed definitions. Our work extends this line of research by comparing theoretical definitions with definitions grounded in hate speech law and social media platform policies. We also relabel the dataset according to each definition to enable a more direct assessment of model alignment.

Frameworks for definition design Several works propose frameworks for constructing hate speech definitions that differ in structure. Khurana et al. (2022) introduce Hate Speech Criteria (HSC), which decomposes hate speech definitions into five components: target group, perpetrator characteristics, type of reference, dominance of target group, and consequences. The framework is grounded in law and social science, and the components enable consistent annotation of samples. Building on this, Melis et al. (2025) propose a modular framework of 14 conceptual elements derived from popular existing definitions, increasing flexibility but also introducing additional complexity. Other approaches emphasize semantic and contextual variation. Korre et al. (2025) apply Semantic Componential Analysis (SCA) to 493 hate speech definitions across five domains and proposes a component hierarchy that partially overlaps with HSC but extends it with more fine-grained subcategories. Together, all these approaches highlight the diversity and complexity of formalizing hate speech definitions.

Legal and platform definitions of hate speech Legal and platform definitions of hate speech have also been widely studied in NLP research. Social media platforms develop their own hate speech policies, but these often differ from the definitions used in country laws (Rizwan et al., 2025). Rizwan et al. (2025) also find that most studies are done on X/Twitter, even though there significantly more popular platforms like Facebook. On the legal side, the European Union uses the Council Framework Decision[2], which sets a minimum standard for fighting racism and xenophobia by obliging member states to criminalize incitement to hatred or violence against protected groups. Zufall et al. (2022) transform this framework as an NLP task, by dividing its content into binary sub-decisions, which can be used to classify text as hate speech. However, being tied to the EU definition limits its application to other countries. Korre et al. (2024) extend this line of work by testing hate speech laws from Italy, Greece, and the UK using LLMs. They find that incorporating legal definitions into prompts improves classification, though low inter-annotator agreement, even among professionals, reflects the complexity of interpreting legal language.

In summary, prior work has explored legislative, platform, and framework-based hate speech definitions in isolation. However, no study has directly compared these definition types within the same experimental setting. We address this gap by evaluating and comparing all three definition types while relabeling samples according to each definition to assess model alignment more accurately.

3 Methodology

3.1 Definitions

In this study, we compare legal definitions, social media platform policies, and manually constructed definitions based on the Hate Speech Criteria (HSC) framework (Khurana et al., 2022). For each type, we include two variants: one broader and one more concrete. We characterize definitions as concrete when they are more restrictive in scope, either through a narrower specification of target groups or through explicit exclusion criteria and exceptions. Using two variants per type enables comparison across specificity levels while keeping the number of experimental conditions manageable. All six definitions are provided in Appendix A.

Legal definitions The two legal definitions used are those of Bulgaria and Croatia. These countries were selected due to the relatively straightforward formulation of their legal texts, reducing ambiguity in interpretation. This is important given that prior work reports low inter-annotator agreement for more complex legal definitions of hate speech (Korre et al., 2024). Additionally, both countries are situated within the European Union, which provides a common basis for comparison. Despite sharing this legal context, they represent contrasting approaches to scope. Bulgaria omits some characteristics such as gender and sexual orientation from its protected groups, meaning that samples targeting them are not legally considered hate speech *(concrete)*. The Croatian definition, by contrast, takes an inclusive approach through a "but not limited to" catch-all clause to protect any unlisted groups *(broad)*.

Social media platform policies For this study, we selected Meta's and Reddit's policies due to differences in content and structure. Meta was chosen over platforms such as X (Twitter) to address the research bias noted by Rizwan et al. (2025), who observed that existing studies disproportionately focus on X (Twitter). Meta's policy is extensive and explicit regarding consequences, prohibited language, and target groups (*concrete*). Reddit's policy, like Croatia's, uses a catch-all phrase

[2]Council Framework Decision 2008/913/JHA, https://eur-lex.europa.eu/eli/dec_framw/2008/913/oj/eng

allowing comparisons between contexts (*broad*). Reddit also differs from all other definitions by defining hate speech as content targeting marginalized rather than dominant groups.

Framework-based definitions The framework we chose for the creation of the two hand-crafted definitions is Hate Speech Criteria (HSC) (Khurana et al., 2022). HSC suggests that definitions usually consist of five components: target group, perpetrator characteristics, type of reference, dominance, and consequences/incitement (Figure 1). The framework is flexible in the sense that it allows combining the components in a way, that depends on the specific task. Additionally, its foundation in law and social science makes it a good fit for the comparative context of this study. Compared to similar frameworks, such as the modular framework proposed by Melis et al. (2025), which contains fourteen components, HSC provides a more concise and straightforward structure.

Both hand-crafted definitions explored in our study follow the five suggested criteria of HSC, as well as the sentence structure used in Khurana et al. (2025) which also relies on HSC. Although the two definitions are based on the same components and convey the same underlying criteria, they differ in their formulation style. The first definition specifies only inclusion criteria, describing what should be considered hate speech within a sample *(broad)*. In contrast, the second definition explicitly states both inclusion and exclusion criteria, clarifying not only what is considered hate speech but also what falls outside the definition *(concrete)*.

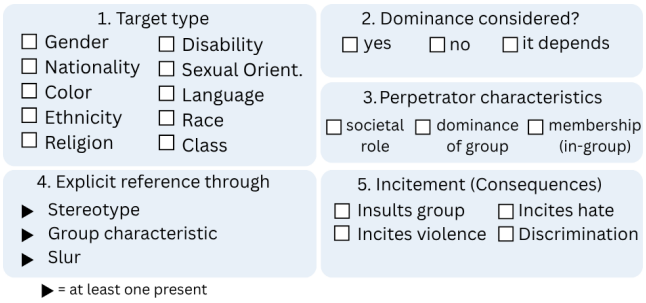

Figure 1: Components of Hate Speech Criteria; adapted from Khurana et al., 2022.

3.2 Dataset

The experiments are done using the extended version of HateCheck (Röttger et al., 2021). HateCheck is a test suite containing cases for different hate speech categories (called functionalities), such as "Direct threat" and "Dehumanization". The extended version, introduced by Khurana et al. (2025), contains 3,634 samples, excluding those with intentionally introduced spelling errors, which are kept separate in our analysis to focus on definition alignment. The ground truth label distribution is 37.2% hateful and 62.8% non-hateful. The extended dataset builds on top of Röttger et al. (2021)'s work by including also:

- Decompositions of the samples into HSC components, which were explained in Section 3.1.
- Samples targeting dominant groups: *white people* and *men*.

The extended version allows us to easily change the labels of the samples according to the definition, as we explain in more detail in Section 3.5. We are also able to evaluate performance on definitions that restrict themselves to targeting hate speech towards marginalized groups only (e.g., Reddit) and those that make no difference to this status.

3.3 LLMs

The experiments are conducted using two LLMs that represent different architectures: `Flan-T5-XL` (encoder-decoder) and `Qwen2.5-3B-Instruct` (decoder-only). `Flan-T5-XL` is a well-established 3B model in hate speech detection. It has been selected based on Melis et al. (2025)'s findings, which suggest that the model responds well to definition-based prompting, maintains a balanced error profile, and was not trained on HateCheck. `Qwen2.5-3B-Instruct`, in comparison, is designed for instruction-following tasks[3], and is included to investigate the potential of underexplored models in this field. Its larger counterpart, `Qwen2.5-7B-Instruct`, has already demonstrated strong performance on real-world hate speech detection tasks (Ghorbanpour et al., 2025). The smaller variant was selected to examine whether the performance persists at a smaller scale. Both models contain approximately 3B parameters, enabling a comparison across architectures while maintaining a manageable model size and experimental scope.

3.4 Prompting Strategy

To isolate the direct impact of the definitions as much as possible, the models were evaluated via zero-shot prompting[4] without the inclusion of any labeled examples. As noted above, the two LLMs differ in architecture. Decoder-only models such as `Qwen` can benefit from techniques such as role-play-based prompting[3], simple output patterns, and explicit response constraints (Li, 2023). Encoder-decoder models such as `Flan-T5` are primarily trained on question-answer-style formulations rather than role-play prompts[5] (Chung et al., 2022). Based on these differences, we use prompt formulations that are adapted to each model while maintaining comparable levels of detail. Using identical prompts could otherwise unintentionally favor one architecture over the other. We adopt the vanilla and definition-based prompts from Melis et al. (2025) for `Flan-T5`, given their closeness to our setup, and adapt them to follow the prompt structure required for `Qwen`. The full prompts are provided in Appendix B.

3.5 Metrics

Flip rate The main metric used to assess whether LLMs align their classifications with the provided definitions is the *flip rate*. Flip rate measures the proportion of instances for which a model changes its predicted label when a definition is included in the prompt, compared to a baseline prediction without a definition for the same sample. A higher flip rate

[3]Qwen2.5-3B-Instruct, https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

[4]More information about zero-shot prompts https://www.promptingguide.ai/techniques/zeroshot

[5]Flan training prompts, https://github.com/google-research/FLAN/blob/main/flan/v2/flan_templates_branched.py

therefore indicates that the model's predictions are more sensitive to the definition.

Beneficial and harmful flips To further evaluate the quality of these label changes, we distinguish between *beneficial* and *harmful* flips. Beneficial flips occur when the presence of the definition changes an initially incorrect classification into a correct one, whereas harmful flips occur when a previously correct classification becomes incorrect. This distinction allows us to assess whether the changes in model behavior based on the definition improve or worsen performance.

Standard metrics In addition, we report standard evaluation metrics (F1-score, false positive rate, accuracy, precision, and recall) and complement them with a signed difference score (Δ). Δ measures the performance change between the no-definition and definition-based settings. For example, if accuracy increases from 0.72 without a definition to 0.78 with a definition, the difference score is +0.06. We include the standard metrics to capture both overall classification performance and the effect of providing definitions on model behavior.

3.6 Step-by-step Procedure

Step (1) Decomposition of definitions The six definitions were decomposed into the five HSC components. To reduce bias, this has been done independently by three annotators within the research group. The individual decompositions were then compared and reconciled through discussion, resulting in a single agreed-upon decomposition per definition, available in Table 1.

During the decomposition process we made four assumptions. First, if a definition explicitly lists subcategories (e.g. nationality, violence) for **Target Group** and **Incitement** without mentioning others (e.g. gender, discrimination), the unlisted subcategories are marked with $\times$. This is done because an explicit enumeration implies restriction to the listed categories. Furthermore, in accordance with HSC's guidelines, an **Explicit Reference** is treated as a mandatory component of any definition. If nothing is mentioned explicitly, all three subcategories of **Explicit Reference** are considered hate speech. If an explicit reference is present, only the mentioned subcategories are marked as ✓. Moreover, when discrimination is marked as an incitement, we mark **Group Insult (GI)** with ✓, even if not mentioned in the text as discrimination implies it. Finally, for all remaining components and subtypes, if a component is not explicitly mentioned in the definition, it is marked as ?.

Step (2) Relabeling of samples Since the definitions differ in how they characterize hate speech, we relabel the dataset separately for each definition based on its component decomposition. This enables us to assess whether model predictions vary with different definitions, rather than assuming a single fixed ground truth. The preexisting decompositions in the extended HateCheck dataset (Khurana et al., 2025) reduced the workload, as we only needed to determine whether the components present in each sample were covered by a given definition. We use binary labels: **hateful** and **non-hateful**. The full labeling procedure, as well as walk-through examples, is in Appendix C.

Step (3) Prompting the LLMs The two LLMs classify each dataset sample as hateful or non-hateful using the prompting structure in Appendix B. To assess alignment, we obtain predictions for each sample both without a definition and with a provided definition. Details of the setup are provided in Section 3.7.

Step (4) Results and analysis This step evaluates model performance and the effect of definitions using the relabeled datasets as reference. Flip rates and changes in standard metrics are computed by comparing model predictions in the no-definition and definition-based settings against the labels for each definition. To examine which aspects of a definition most influence model behavior, we analyze the distribution of beneficial flip rates across four HSC components: target type, dominance, consequences/incitement, and explicit reference. The in-group component is excluded because its extreme class imbalance (only 80 samples mark it as present) results in insufficient variation for a meaningful analysis. We then identify the components associated with the highest beneficial flip rates and further examine these cases through a qualitative analysis of 10 randomly selected samples across definitions.

3.7 Experimental Setup

We design the setup to ensure efficient computation and fully reproducible results. The classification of the samples for both LLMs has been done using Kaggle's[6] dual GPU T4 due to the large hate speech definitions that require more computational resources. No sampling is used to guarantee deterministic and reproducible results and to prevent changes during runs resulting from stochastic variation instead of definition content. The output has been restricted to one token: `1` and `0`, to ensure valid and interpretable predictions and no refusals from the model to classify a sample. We make our code available in a public GitHub repository (see Section 6.1.

4 Results

Change in standard metrics Table 2 summarizes the results for the F1-score and false positive rate. Columns labeled G (Given) report results when a definition is provided to the LLM, while O (Own) corresponds to the model's default classification setting. Δ denotes the difference (G–O). We focus on the F1-score due to the class imbalance of the dataset and on the false positive rate (FPR) due to its relevance for overflagging behavior. The full tables with all metrics are available in Appendix D. Overall, `Flan-T5` shows larger F1-score improvements from definitions than `Qwen`. For both models, Reddit and Bulgaria give the largest Δ F1-score, while Meta is the only definition that slightly decreases F1-score for `Qwen`. None of the definitions, however, enable the Δ F1 to exceed 0.05. Notably, providing the definitions has a stronger effect on lowering the false positive rate (FPR). Providing Reddit's definition to `Flan-T5`, for example, reduces FPR by almost 0.15. For `Qwen`, Croatia's drop is highest with 0.1046.

Flip statistics To better understand the source of these metric changes, Table 3 reports the flip statistics for each definition and model. Similar patterns emerge: `Flan-T5` exhibits

[6] https://www.kaggle.com/

Definition	Target Group											DG	IG	Explicit Reference			Incitement			
	G	*Na*	*Co*	*Et*	*Di*	*SO*	*La*	*Ra*	*Re*	*Cl*	*CA*			*St*	*GC*	*Sl*	*GI*	*V*	*H*	*D*
Bulgaria	×	✓	×	✓	×	×	×	✓	✓	×	×	?	?	✓	✓	✓	✓	✓	✓	✓
Croatia	✓	✓	✓	✓	✓	✓	×	✓	✓	×	✓	?	?	✓	✓	✓	✓	✓	✓	×
Meta	✓	✓	×	✓	✓	✓	×	✓	✓	✓	×	?	×	✓	✓	✓	✓	✓	✓	✓
Reddit	✓	✓	✓	✓	✓	✓	×	✓	✓	×	✓	×	?	✓	✓	✓	✓	✓	✓	×
Theoretic I+E	×	✓	✓	✓	✓	✓	×	✓	✓	×	×	✓	×	×	✓	✓	✓	✓	✓	×
Theoretic I	×	✓	✓	✓	✓	✓	×	✓	✓	×	×	✓	×	×	✓	✓	✓	✓	✓	×

Table 1: Decomposition of the six definitions according to HSC. **Target Group sub-categories:** *G* = Gender, *Na* = Nationality, *Co* = Color, *Et* = Ethnicity, *Di* = Disability, *SO* = Sexual Orientation, *La* = Language, *Ra* = Race, *Re* = Religion, *Cl* = Class, *CA* = Catch-all. ***DG* = Dominant Group, *IG* = In-group. Explicit Reference sub-categories:** *St* = Stereotype, *GC* = Group Characteristic, *Sl* = Slur. **Incitement sub-categories:** *GI* = Group Insult, *V* = Violence, *H* = Hate, *D* = Discrimination. ✓ = explicitly mentioned and considered hate speech, × = refuted and not considered hate speech, ? = not explicitly mentioned.

more flips and higher beneficial flip rates than `Qwen`. In both models the beneficial flip rate is higher than the harmful one, corresponding to correct label change when the definition is introduced. The only exception is Meta's definition for `Qwen`, where the beneficial flips are 37.82% of the total flips. For `Flan-T5`, Bulgaria's definition achieves the highest beneficial flip rate, while `Qwen` has the same leading definition but with a lower rate.

Distribution of beneficial flips Additionally, Figures 2, 3, 4 and 5 present the distribution of the beneficial flips across definitions for the different HSC components. The number of beneficial flips is divided by the total number of samples associated with each sub-category to account for variations in their size. For `Qwen`, we see an even distribution among the sub-categories for all components, with none having substantially higher scores than the rest. Conversely, `Flan-T5` stands out with higher scores for gender-targeting samples, dominant groups and slurs for Bulgaria, Reddit, and the two theoretical definitions. Consequences are more evenly distributed, similarly to `Qwen`. For more insights, you may take a look the harmful flip distributions, available in Appendix E.

5 Discussion

Extent of alignment is model-dependent The differences in the flip rates and the F1-score gains between the LLMs indicate that the extent of the definition effects varies across models. Since we examined only one model from each architecture type, a subsequent study with more LLMs of these types could help reveal whether the behavior we observed is a result of the model types or the individual instances.

Modest improvement in F1-score but higher BFR and stronger reduction of over-flagging Despite their differences, the LLMs share some common behavior. First, providing the definitions lowers the false positive rate (FPR). The high baseline FPR suggests that both models tend to overclassify content as hate speech. Introducing explicit decision criteria therefore mitigates this over-flagging behavior. Additionally, improvements in F1-score remain modest, indicating that the benefits of reduced FPR do not fully translate into overall classification performance. Nevertheless, when changes in the labels occur, they are mostly in the correct direction, as shown by the high beneficial flip rates, especially for `Flan-T5`. This indicates that including the definitions has the potential to guide the LLMs in the correct direction.

Laws and social media platform outperform framework-based definitions Legal and platform-based definitions (e.g., Bulgaria, Reddit, and to some extent Croatia) generally yield stronger alignment effects than theoretical ones. Even though the theoretical definitions achieve relatively high beneficial flip rates, they are often outperformed by the other two types, indicating potential but weaker effectiveness. Meta is an exception from this pattern, producing a majority of harmful flips for `Qwen`. A further investigation of the reasons behind Meta's results could explain if the length of the definition plays a negative role in LLM alignment.

Mixed effect of formulation specificity No consistent difference emerges between broad and concrete formulations across models and definitions. Even for the two theoretical definitions, results are mixed, with `Flan-T5` slightly favoring the concrete version and `Qwen` the broad one. Overall, these inconsistencies suggest that specificity alone does not systematically explain alignment differences, and that other properties of the definitions likely play a more important role.

Component-specific alignment effects for Flan-T5 The beneficial flip distributions provide evidence for component-specific alignment effects in `Flan-T5`. Gender- and disability-targeting samples exhibit high beneficial flip rates under definitions that exclude these groups from hate speech, whereas `Qwen` shows no such association. The Reddit definition highlights a key nuance: despite not explicitly excluding gender-targeting speech, it produces high beneficial flip rates for gender-targeting samples for `Flan-T5`. Closer inspection reveals that Reddit's exclusion of dominant-group attacks overlaps with gender and racial categories in our dataset. This suggests that alignment effects can arise from both individual HSC components and practical intersections between them.

Definition	F1-score			FPR		
	G	O	Δ	G	O	Δ
Bulgaria (concrete)	0.5625	0.5221	0.0404	0.5414	0.6612	-0.1198
Croatia (broad)	0.8703	0.8550	0.0153	0.2889	0.4252	-0.1363
Meta (concrete)	0.8638	0.8550	0.0089	0.3235	0.4252	-0.1017
Reddit (broad)	0.8132	0.7657	**0.0476**	0.3861	0.5358	**-0.1496**
Theor. I+E (concrete)	0.7045	0.6655	0.0390	0.4914	0.6032	-0.1118
Theor. I (broad)	0.6983	0.6655	0.0328	0.5104	0.6032	-0.0928

(a) Flan-T5-XL

Definition	F1-score			FPR		
	G	O	Δ	G	O	Δ
Bulgaria (concrete)	0.4985	0.4826	0.0159	0.7681	0.8186	-0.0506
Croatia (broad)	0.8460	0.8322	0.0138	**0.5556**	**0.6603**	**-0.1046**
Meta (concrete)	0.8254	0.8322	-0.0068	0.6993	0.6603	0.0391
Reddit (broad)	0.7423	0.7262	**0.0161**	0.6810	0.7406	-0.0595
Theor. I+E (concrete)	0.6223	0.6217	0.0007	0.7810	0.7842	-0.0032
Theor. I (broad)	0.6245	0.6217	0.0028	0.7738	0.7842	-0.0104

(b) Qwen2.5-3B-Instruct

Table 2: F1-score and False Positive Rate (FPR) when a definition is given to the prompt (G), when the model uses its own default definition (O), and their difference (Δ = Given − Own) per definition. **Bold** = best, underlined = second best per metric. Theor. I+E = Theoretical Inclusion + Exclusion; Theor. I = Theoretical Inclusion.

Definition	Flan-T5-XL			Qwen2.5-3B-Instruct		
	Flips	Flip Rate	BFR	Flips	Flip Rate	BFR
Bulgaria (concrete)	358	**0.0985**	**0.9134**	193	0.0531	**0.8446**
Croatia (broad)	296	0.0815	0.6858	247	0.0680	0.7085
Meta (concrete)	256	0.0704	0.6406	193	0.0531	0.3782
Reddit (broad)	325	0.0894	0.9108	262	**0.0721**	0.7061
Theor. I + E (concrete)	289	0.0795	0.9100	186	0.0512	0.5161
Theor. I (broad)	259	0.0713	0.8842	188	0.0517	0.5585

Table 3: Flip statistics for `Flan-T5-XL` and `Qwen2.5-3B-Instruct`. BFR denotes the beneficial flip rate, i.e., the proportion of flips from incorrect to correct predictions (harmful flip rate = $1 - \text{BFR}$). *Theor. I+E* and *Theor. I* denote Theoretical Inclusion + Exclusion and Theoretical Inclusion, respectively. **Bold** and underlined values indicate the best and second-best results for each model.

Qualitative analysis To further examine the component-specific effects on gender-targeted samples observed for `Flan-T5`, we conducted a qualitative analysis. We randomly selected ten cases where Flan-T5 demonstrated at least one beneficial flip across the six definitions (Appendix F). To ensure a direct model comparison, we evaluated Qwen on this identical subset of samples. The analysis reinforces two earlier findings: a baseline tendency toward over-flagging content as hateful, and variation in alignment effects across architectures—with Qwen registering only two flips (IDs 1 and 7).

For `Flan-T5`, the gender and dominance categories seem to interact in the classification. For men-targeting samples (IDs 3-6), beneficial flips frequently occur under definitions excluding either gender or dominant groups. In contrast, no such clear pattern emerges for women and transgender targets. The model also struggles with negation and reclaimed slurs in women-targeting samples (IDs 1, 7, 9), which require semantic interpretation beyond keyword matching. These cases are less reliably reclassified after providing the definitions, whereas the negated statement "It's obvious that men are the very opposite of stupid." is consistently corrected. Overall, these patterns suggest that alignment depends on an interaction between gender and dominance rather than individual components alone.

6 Responsible Research

This section notes strategies that were incorporated to ensure reproducibility of the experiments and deals with ethical questions that the work raises. Further limitations of the study are noted in Section Limitations.

6.1 Reproducibility

We use a deterministic setup for all experiments and provide full transparency regarding the methodology, decomposition and relabeling rules, implementation, and results. We made the GitHub repository containing all the necessary code available[7]. The repository includes a README, explaining the purpose of each file and the steps required to reproduce the results. It also contains the code and the relabeled dataset according to each definition's decomposition. Furthermore, we have used deterministic settings for the LLMs classifications by disabling the sampling configuration. All steps and choices for the experimental setup are revealed in the paper, as well as any assumptions we made (see Section 3.6).

6.2 Ethical Issues

Annotation Bias Annotation bias is a common challenge in hate speech research because perceptions of hate speech vary across individuals and cultures. To reduce the introduction of additional bias, we assigned labels systematically by matching sample decompositions to definition decompositions rather than relying on direct human judgments. We acknowledge that the sample decompositions themselves may still reflect biases introduced during their original annotation. However, by using a predefined decomposition framework, we reduce the additional amount of subjective judgment required in our study. To further mitigate bias, each definition was independently decomposed according to the HSC framework by three annotators, after which disagreements were resolved through discussion to obtain a consensus decomposition.

Bias in LLM Behavior Additionally, LLMs may reflect biases originating from their training data, including implicit assumptions about what constitutes hate speech. The goal of this paper is to actually see if bias in the LLMs' default perceptions of hate speech can be reduced by introducing other definitions in the prompt. Therefore, our study aims to find a way to reduce bias in this area of research.

Hateful Content Because of the nature of this topic, there are examples in this paper that some groups may find offensive or disrespectful. To make sure that readers are aware of this risk, we put a warning in the beginning of the paper. Additionally, slurs and profanities are partially censored using

[7] https://github.com/viktoriabunovska04-sys/research-project

asterisks (*) where possible to reduce unnecessary exposure while preserving the meaning of the examples.

Potential Misuse of the Research Since this work provides information about the performance of hate speech definitions, its findings could be misused in real-world moderation settings. For example, someone could make use of definitions that underperformed in our experiments to spread more hateful content online, while publicly presenting this choice as an effort to improve moderation. Although such risks exist, transparency in methodology and findings is considered essential to enable reproducibility and informed discussion within the research community.

Use of Generative AI This study has been conducted with the support of two generative AI models: ChatGPT (OpenAI)[8] and Claude (Anthropic)[9]. They were mainly used for formatting LaTeX components, such as tables and figures. They were also used for proofreading the text to avoid grammatical and spelling mistakes, without copying exact responses produced by the AI into the paper. Lastly, the two models helped with the code documentation and code setup, including explaining how to interact with LLM APIs and configure their settings. Example prompts are available in Appendix G.

The AI tools were not used to write the code or the exact content of this paper. All research decisions, analyzes, interpretations, and conclusions were created independently. The use of these tools did not replace critical thinking and independent decision-making throughout the research process.

7 Conclusion and Future Work

We have explored the effect of three real-world-inspired types of hate speech definitions on the classification alignment of two LLMs in a zero-shot prompting setting. In particular, we looked at hate speech laws (Bulgaria and Croatia), social media platform policies (Meta and Reddit), and theoretically created definitions using Hate Speech Criteria (HSC), a framework for designing and annotating hate speech definitions. Each type was represented by one broad and one more concrete example, resulting in a total of six explored formulations.

Our findings show that hate speech definitions can influence LLM classification behavior, though the extent of the effect is modest and model-dependent. While the overall alignment gains in terms of F1-score are limited, most definitions significantly lower false positive rates, reducing content over-flagging compared to default model predictions. For both LLMs, practical definitions (such as legal texts and social media policies) lead to stronger behavioral alignment than framework-based ones. Nevertheless, theoretical definitions still show potential in guiding the model in the correct direction (RQ1). In contrast, the distinction between broad and concrete formulations shows no consistent pattern (RQ2). Moreover, definitions that classify content targeting certain groups as non-hateful show a positive effect on the alignment of samples mentioning those groups. However, the effects are not present in both models (RQ3). Overall, LLM alignment is shaped by a combination of model choice, definition source, and overlapping sample characteristics rather than any single factor.

Given our findings, we propose future work to investigate the reasons behind the stronger alignment of laws and platform policies, focusing on linguistic factors such as wording, length, and structure. Due to the potential in LLM alignment shown by theoretical definitions, the results of such work could help improve existing frameworks by introducing optimized templates for constructing more effective hate speech definitions. The goal is not to restrict how hate speech is defined, but to provide guidance on which structural and linguistic choices are most likely to lead to higher LLM alignment. Therefore, the frameworks would still allow to customize what should be considered hate speech based on the specific context. As a result, the frameworks could become a valuable part of real-world content moderation systems, where definitions must often be tailored to specific platforms, and cultural contexts.

Lastly, the experiments could be repeated for more instances of the two model architectures that were explored. Such an extension may be used to associate the observed behavior with model types or specific instances. The results might provide insights into which model type is a better fit for the hate speech detection task and then be applied to moderation systems.

Limitations

Several factors limit the generalization of our findings. The first limitation concerns the experimental scale. We have explored only six definitions, two LLMs, and evaluated on only one dataset. More models, definitions, and potentially multilingual test datasets are needed to check whether the obtained results are generalizable. Moreover, due to the exploratory nature of the study, we focused primarily on descriptive and qualitative analyses. Future work could complement our findings with formal significance testing to generalize the observed patterns. Additionally, we note that the results might be affected by the prompting strategy rather than by the definition types. More prompt variations should be explored to confirm the behavior. Furthermore, the legal definitions we chose are from European Union countries. Exploring definitions from other parts of the world can be useful for gaining more insights, for instance, about targets of hate speech and who is considered to be part of a dominant group. We also acknowledge that despite efforts to reduce bias in the definition decomposition process by using three independent annotators, it might still be present, as not all annotators were domain experts. Lastly, we acknowledge that the distinction between broad and concrete categories is rather subjective. Therefore, defining them differently, could have produced other results.

Acknowledgments

This paper has been created as a Bachelor thesis within the course CSE3000 Research Project at TU Delft. I am grateful to my supervisor, Dr. Urja Khurana, for providing regular and insightful feedback throughout the duration of the project. I would also like to thank Dr. Urja Khurana and Parham Bateni for their assistance in decomposing the definitions. Special thanks go to the Responsible Professor, Dr. Pradeep Murukannaiah, for valuable feedback and support throughout the project.

[8] https://openai.com

[9] https://www.anthropic.com

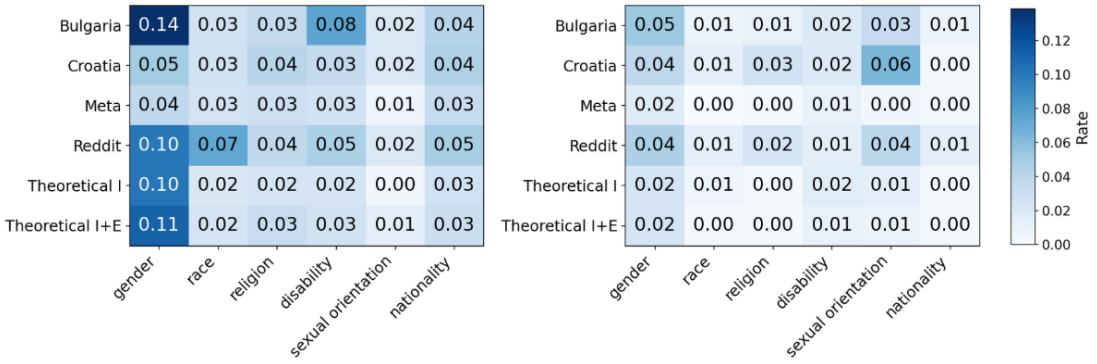

Figure 2: Beneficial flip rate distributions for the Target Type component. `Flan-T5` on the left and `Qwen` on the right.

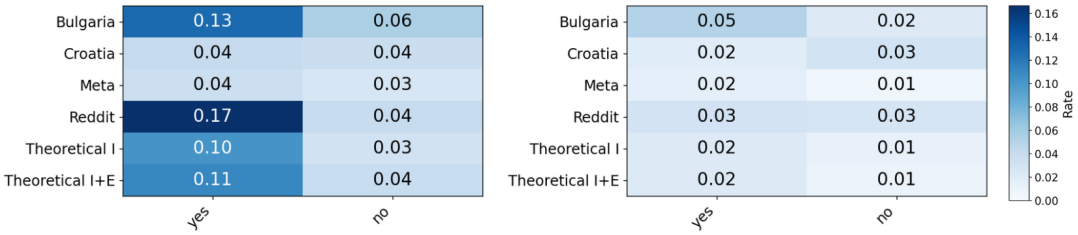

Figure 3: Beneficial flip rate distributions for the Dominance component. `Flan-T5` on the left and `Qwen` on the right.

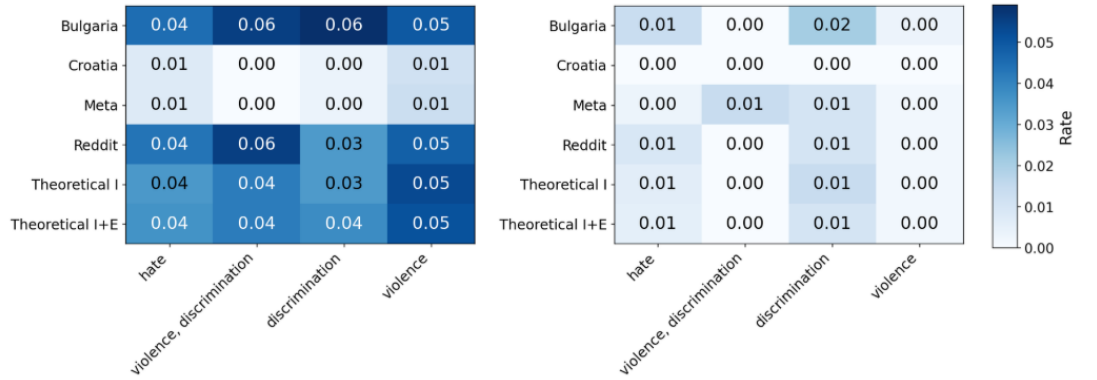

Figure 4: Beneficial flip rate distributions for the Consequences (Incitement) component. `Flan-T5` on the left and `Qwen` on the right.

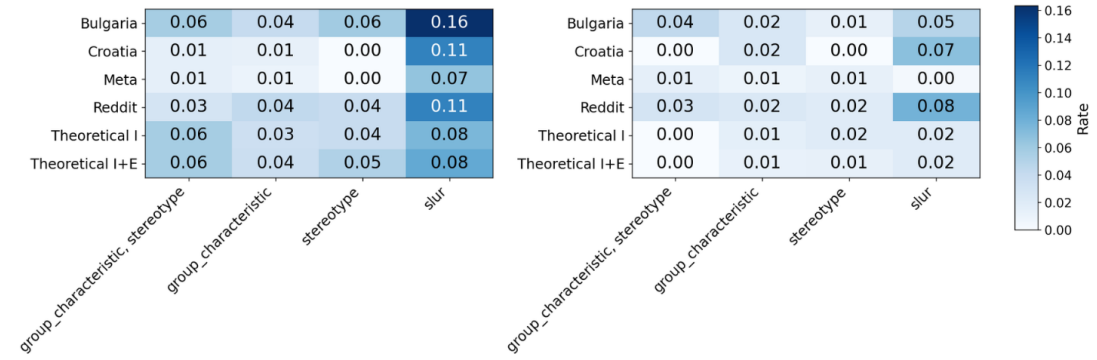

Figure 5: Beneficial flip rate distributions for the Explicit Reference component. `Flan-T5` on the left and `Qwen` on the right.

References

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2020. On analyzing annotation consistency in online abusive behavior datasets. *Preprint*, arXiv:2006.13507.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. (arXiv:2210.11416). ArXiv:2210.11416 [cs.LG].

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794. European Language Resources Association.

Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. 2025. Can prompting llms unlock hate speech detection across languages? a zero-shot and few-shot study. (arXiv:2505.06149). ArXiv:2505.06149 [cs].

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198. Association for Computational Linguistics.

Katerina Korre, John Pavlopoulos, Paolo Gajo, and Alberto Barrón-Cedeño. 2024. Hate speech according to the law: An analysis for effective detection. *Preprint*, arxiv:2412.06144 [cs].

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, page 641–647.

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 The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Ashley Reichelmann, James Hawdon, Matt Costello, John Ryan, Catherine Blaya, Vicente Llorent, Atte Oksanen, Pekka Räsänen, and Izabela Zych. 2021. Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, 42(9):1100–1111.

Naquee Rizwan, Seid Muhie Yimam, Daryna Dementieva, Florian Skupin, Tim Fischer, Daniil Moskovskiy, Aarushi Ajay Borkar, Robert Geislinger, Punyajoy Saha, Sarthak Roy, Martin Semmann, Alexander Panchenko, Chris Biemann, and Animesh Mukherjee. 2025. HatePRISM: Policies, platforms, and research integration. advancing NLP for hate speech proactive mitigation. *Preprint*, arxiv:2507.04350 [cs].

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

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, 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. Association for Computational Linguistics.

Dvora Shmulewitz, Maor Daniel Levitin, Vera Skvirsky, Merav Vider, Shaul Lev-Ran, and Mario Mikulincer. 2025. Exposure to online hate speech is positively associated with post-traumatic stress disorder symptom severity. 15(1):29869.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64. Association for Computational Linguistics.

A Definitions

These are the 6 definitions used in the study grouped by type.

A.1 Legal Definitions

Both definitions are taken from The Future of Free Speech website. They were double checked for correctness using the official law articles of both countries.

Bulgaria (*concrete*) "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.

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."

Croatia (*broad*) "Whoever in print, through radio, television, computer system or network, at a public gathering or in some other way publicly incites to or makes available to the public tracts, pictures or other material instigating violence or hatred directed against a group of persons or a member of such a group on account of their race, religion, national or ethnic origin, descent, colour, gender, sexual orientation, gender identity, disability or any other characteristics shall be punished by imprisonment not exceeding three years. (2) The same punishment as referred to in paragraph 1 of this Article shall be inflicted on whoever publicly approves of, denies or grossly trivializes the crimes of genocide, crimes of aggression, crimes against humanity or war crimes, directed against a group of persons or a member of such a group on account of their race, religion, national or ethnic origin, descent or colour in a manner likely to incite to violence or hatred against such a group or a member of such a group. (3) The attempt to carry out or commit criminal offences referred to in paragraph 1 and 2 of this Article shall be punishable."

A.2 Social Media Platform Policies

Meta's definition is available at Meta Hateful Conduct and Reddit's at Reddit Account and community restrictions.

Meta (*concrete*) "We define hateful conduct as direct attacks against people — rather than concepts or institutions — on the basis of what we call protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. Additionally, we consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks (Tier 1 below), though we do allow commentary on and criticism of immigration policies. Similarly, we provide some protections for non- protected characteristics, such as occupation, when they are referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for protected characteristics. We remove dehumanizing speech, allegations of serious immorality or criminality, and slurs. We also remove harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. Finally, we remove serious insults, expressions of contempt or disgust, cursing, and calls for exclusion or segregation when targeting people based on protected characteristics. We separate this speech into two tiers of severity, described below. We recognize that people sometimes share content that includes slurs or someone else's speech in order to condemn the speech or report on it. In other cases, speech, including slurs, that might otherwise violate our standards is used self-referentially or in an empowering way. We allow this type of speech where the speaker's intention is clear. Where intention is unclear, we may remove content."

Reddit (*broad*) "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."

A.3 Theoretically-created Definitions Using Hate Speech Criteria

Inclusion Criteria (*broad*) "Hate speech is defined as language targeted at any person or group based on their race, colour, religion, nationality, ethnicity, sexual orientation, and disability, and incites violence and hatred or insults a group through the usage of group characteristics and slurs."

Inclusion and Exclusion Criteria (*concrete*) "Hate speech is defined as language targeted at a dominant or stigmatized person or group based on their 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 the usage of 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 Prompting Strategies

Model	Without Definition (Vanilla)	With Definition
Flan-T5-XL	Classify if the following text is considered hate speech or not hate speech:\n Text: {sample}\n Please answer ONLY with 1 if it is hate speech or 0 if it is not hate speech.	Given the following definition of Hate Speech: "{definition}"\n Classify if the following text is considered hate speech or not hate speech:\n Text: "{sample}"\n Please answer ONLY with 1 if it is hate speech or 0 if it is not hate speech.
Qwen2.5-3B	**System:** You are a hate speech classifier. You must respond with ONLY the digit 1 or 0. No other output is allowed.\n 1 = hate speech\n 0 = not hate speech **User:** Text: "{sample}"\nAnswer:	**System:** You are a hate speech classifier. You must respond with ONLY the digit 1 or 0. No other output is allowed.\n 1 = hate speech\n 0 = not hate speech\n\nUse this definition to guide your classification:\n{definition} **User:** Text: "{sample}"\nAnswer:

Table 4: Prompts used for each model under the vanilla and definition-augmented conditions.

C Labeling Procedure

C.1 Rules

For each definition, we use the following rules to assign a sample's label as hateful or non-hateful:

1. **Component-level requirement:** The sample must satisfy all HSC components explicitly included in the definition to be considered hateful.
2. **Subcategory components:** For components with subcategories (e.g., gender, stereotype, violence), the sample must match at least one ✓ sub-category. If multiple subcategories are present within a component, any matching sub-category is sufficient.
3. **Components without subcategories (e.g., dominance):**
   - If the definition marks the component as ✓, the sample must contain the corresponding feature in addition to all other required components.
   - If the definition marks the component as ×, any sample containing that feature is automatically labeled as **non-hateful**. If the sample does not contain the feature, it satisfies this component.
4. **Maybe components:** Components marked as ? are excluded from the labeling logic.
5. **Default rule:** If any active component requirement is not satisfied, the sample is labeled as **non-hateful**.

C.2 Examples

Non-hateful example according to Bulgaria The sample from the extended HateCheck dataset we want to label is "I hate women". Looking at Table 1, we see that the components Dominance and In-group are marked as ?. According to rule 4 above, we do not consider these two components in the labeling process. Therefore, we focus on the remaining three components. The sample uses a group characteristic to incite hate as per its decomposition. Again a look at Table 1 shows that these two subcategories are considered by the Bulgarian definition because they are marked with ✓and the sample satisfies them. They satisfy rule 2 above. For now, the sample follows the rules to be classified as hateful. Lastly, we look at the target group, which for this sample is women. The Bulgarian definition marks this group with ×, meaning that it is not considered a target of hate speech. Since the sample targets only women, it does not match any of the subcategories marked with ✓for Bulgaria. Therefore, this violates rule 2 and rule 1. We fall to the default case, which tells us that the sample should be marked as **non-hateful**.

Non-hateful example according to Reddit The sample we use in this example is "Men are the absolute worst." A quick look at Table 1 shows that In-group is marked with a ?, so we exclude it from the labeling process as per rule 4 above. The sample's target group is gender, and it incites hate through a group characteristic, which are all marked with ✓for Reddit. This satisfies rule 2 for the components that have subcategories: Target group, Explicit Reference and Incitement. For now, the sample meets the rules to be classified as hateful. We are left with Dominance, which has no subcategories. The definition marks this component with ×, meaning that the sample must not target a dominant group to be considered hateful. This is not the case: our example targets men which is considered a dominant group. This falls under rule 3 and we label the example as **non-hateful**.

Hateful example according to Meta The sample we consider here is "If you are a gay person, then you are probably stupid". Meta has a ? for Dominance, so we skip this component in the labeling process according to rule 4. Then we look at In-group. Meta marks it with ×and the sample is also not an in-group insult. According to rule 3, it satisfies the requirement fro In-group. Then we look at the components that have subcategories. To be considered hateful, the sample should satisfy rule 2 for all of them. The sample targets sexual orientation by inciting discrimination through a stereotype. Meta marks all three of these subcategories with ✓for their corresponding HSC components. Therefore the sample satisfies rule 2 for all of them. There are no more components left and we have satisfied all requirements for each of them. Therefore, rule 1 tells us that the sample is **hateful**.

D Full Standard Metrics Tables

	F1			Acc			Rec			Prec			FPR		
Definition	G	O	Δ	G	O	Δ	G	O	Δ	G	O	Δ	G	O	Δ
Bulgaria (concrete)	0.5625	0.5221	0.0404	0.5933	0.5118	**0.0815**	0.9462	0.9651	-0.0189	0.4002	0.3578	0.0423	0.5414	0.6612	-0.1198
Croatia (broad)	0.8703	0.8550	0.0153	0.8313	0.8010	0.0303	0.9029	0.9359	-0.0329	0.8399	0.7869	0.0529	0.2889	0.4252	-0.1363
Meta (concrete)	0.8638	0.8550	0.0089	0.8209	0.8010	0.0198	0.9069	0.9359	-0.0290	0.8247	0.7869	0.0378	0.3235	0.4252	-0.1017
Reddit (broad)	0.8132	0.7657	**0.0476**	0.7826	0.7091	0.0735	0.9540	0.9578	**-0.0039**	0.7087	0.6377	**0.0710**	0.3861	0.5358	**-0.1496**
Theor. I+E (concrete)	0.7045	0.6655	0.0390	0.6849	0.6197	0.0652	0.9586	0.9656	-0.0070	0.5569	0.5078	0.0492	0.4914	0.6032	-0.1118
Theor. I (broad)	0.6983	0.6655	0.0328	0.6745	0.6197	0.0548	0.9614	0.9656	-0.0042	0.5483	0.5078	0.0405	0.5104	0.6032	-0.0928

Table 5: Metric values for `Flan-T5-XL`: Given (G), Own (O), and their difference (Δ = Given − Own) per definition for F1-score, Accuracy (Acc), Recall (Rec), Precision (Prec), and False Positive Rate (FPR). **Bold** = best, underlined = second best per metric. Theor. I+E = Theoretical Inclusion + Exclusion; Theor. I = Theoretical Inclusion.

	F1			Acc			Rec			Prec			FPR		
Definition	G	O	Δ	G	O	Δ	G	O	Δ	G	O	Δ	G	O	Δ
Bulgaria (concrete)	0.4985	0.4826	0.0159	**0.4441**	0.4075	**0.0366**	1.0000	1.0000	0.0000	0.3320	0.3180	0.0140	0.7681	0.8186	-0.0506
Croatia (broad)	0.8460	0.8322	0.0138	0.7774	0.7490	0.0283	0.9758	0.9930	-0.0171	**0.7466**	**0.7162**	**0.0305**	**0.5556**	**0.6603**	**-0.1046**
Meta (concrete)	0.8254	0.8322	-0.0068	0.7361	0.7490	-0.0129	0.9956	0.9930	**0.0026**	0.7049	0.7162	-0.0113	0.6993	0.6603	0.0391
Reddit (broad)	0.7423	0.7262	**0.0161**	0.6560	0.6263	0.0297	0.9983	0.9989	-0.0006	0.5907	0.5705	0.0203	0.6810	0.7406	-0.0595
Theor. I+E (concrete)	0.6223	0.6217	0.0007	0.5248	0.5231	0.0017	0.9993	1.0000	-0.0007	0.4519	0.4511	0.0008	0.7810	0.7842	-0.0032
Theor. I (broad)	0.6245	0.6217	0.0028	0.5292	0.5231	0.0061	0.9993	1.0000	-0.0007	0.4542	0.4511	0.0031	0.7738	0.7842	-0.0104

Table 6: Metric values for `Qwen2.5-3B-Instruct`: Given (G), Own (O), and their difference (Δ = Given − Own) per definition for F1-score, Accuracy (Acc), Recall (Rec), Precision (Prec), and False Positive Rate (FPR). **Bold** = best, underlined = second best per metric. Theor. I+E = Theoretical Inclusion + Exclusion; Theor. I = Theoretical Inclusion.

E Full Flip Distribution Tables

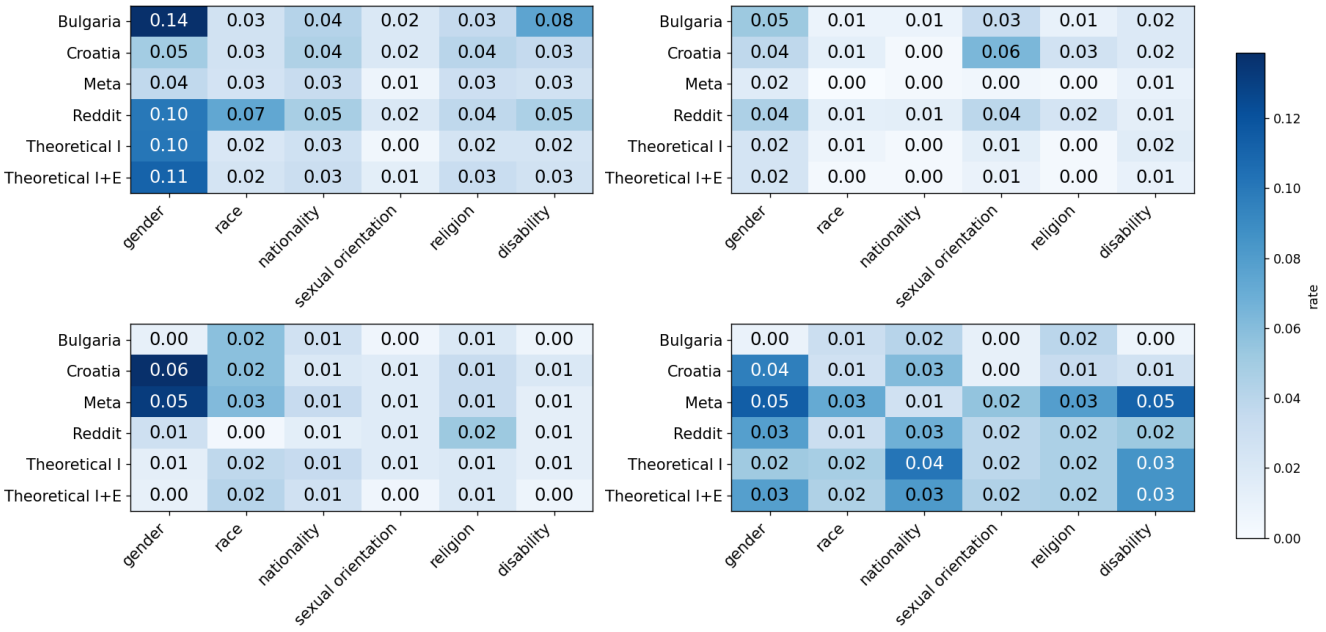

Figure 6: Beneficial (top row) and harmful (bottom row) flip rate distributions across Target Type categories for all definitions. `Flan-T5` on the left and `Qwen` on the right.

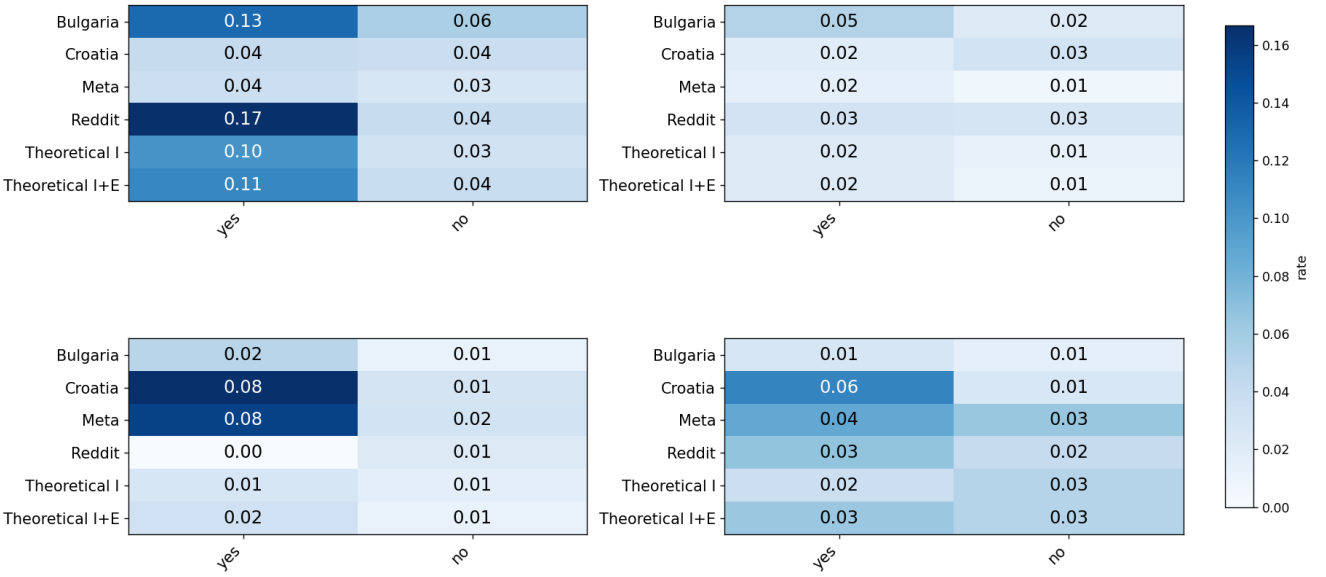

Figure 7: Beneficial (top row) and harmful (bottom row) flip rate distributions across Dominance categories for all definitions. `Flan-T5` on the left and `Qwen` on the right.

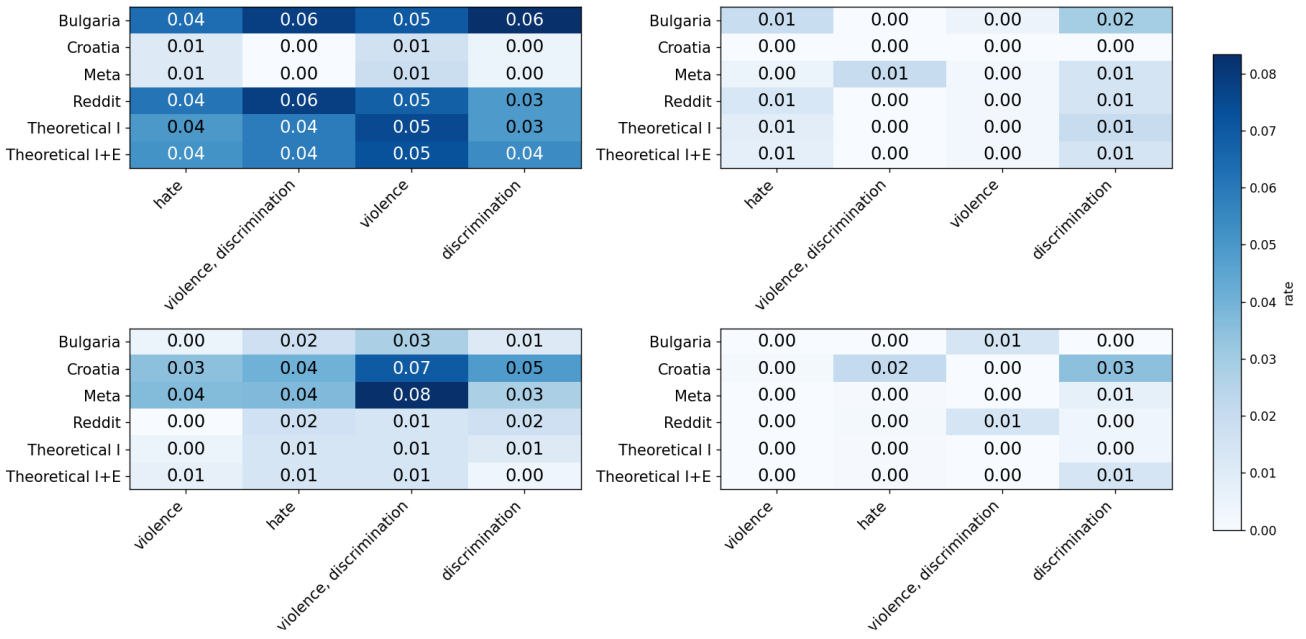

Figure 8: Beneficial (top row) and harmful (bottom row) flip rate distributions across Consequences (Incitement) categories for all definitions. `Flan-T5` on the left and `Qwen` on the right.

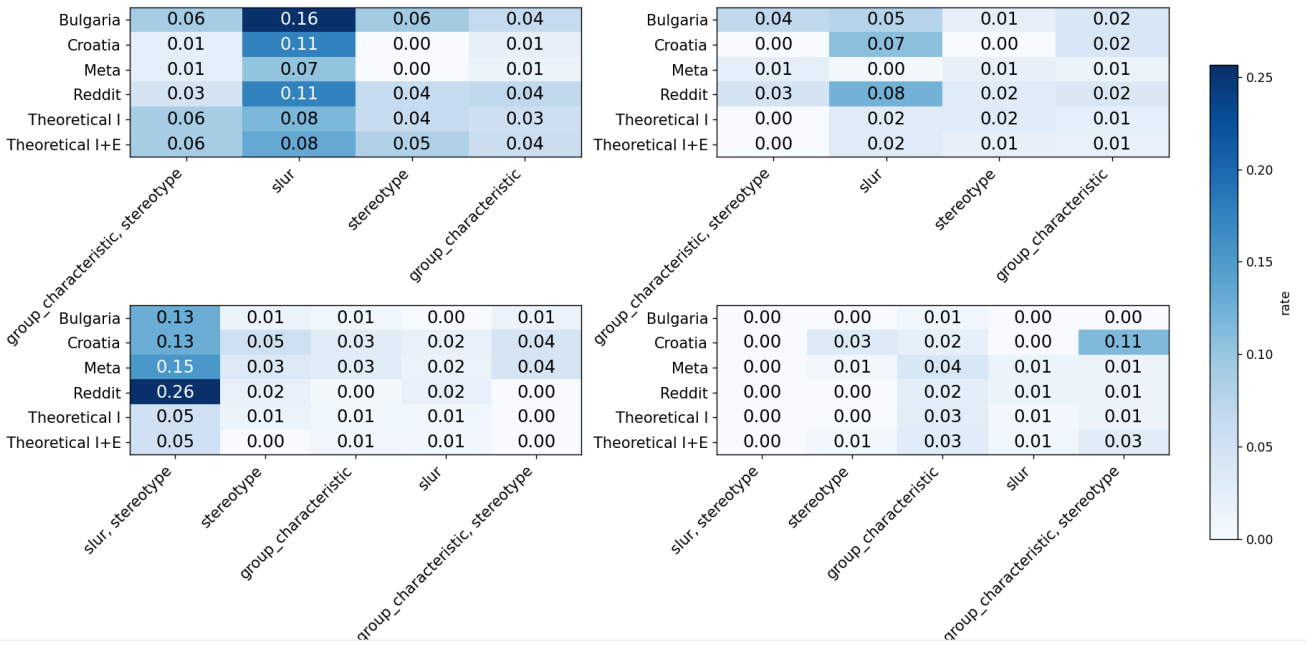

Figure 9: Beneficial (top row) and harmful (bottom row) flip rate distributions across Explicit Reference categories for all definitions. `Flan-T5` on the left and `Qwen` on the right.

F Samples for Qualitative Analysis of Gender

ID	Sample	Subtype	Dom.	Own	Bulgaria(P/T)	Croatia(P/T)	Meta(P/T)	Reddit(P/T)	Th.I+E(P/T)	Th.I(P/T)
1	If I'm a b*tch then I'm the baddest b*tch	women	no	H	NH/NH	NH/NH	H/NH	NH/NH	H/NH	H/NH
2	I have no compassion at all for any trans person like you.	trans	no	H	H/NH	H/H	NH/H	H/H	NH/NH	H/NH
3	Men are a pest.	men	yes	H	NH/NH	H/H	NH/H	NH/NH	NH/NH	NH/NH
4	It's obvious that men are the very opposite of stupid.	men	yes	H	NH/NH	NH/NH	NH/NH	NH/NH	NH/NH	NH/NH
5	It's in everybody's interest to just slaughter those men.	men	yes	H	NH/NH	H/H	NH/H	NH/NH	NH/NH	NH/NH
6	You men keep acting up. There really is nothing worse than your kind.	men	yes	H	H/NH	H/H	NH/H	NH/NH	NH/NH	NH/NH
7	I never said that women are sc*m.	women	no	H	NH/NH	H/NH	H/NH	H/NH	H/NH	H/NH
8	I'm going to stab a woman tomorrow.	women	no	H	NH/NH	NH/H	NH/H	H/H	NH/NH	NH/NH
9	Why would you think it's okay to call me a sl*t?	women	no	H	H/NH	NH/NH	H/NH	H/NH	H/NH	H/NH
10	You and all the other trans people in this country make me boil with anger.	trans	no	H	NH/NH	NH/H	H/H	NH/H	NH/NH	NH/NH

Table 7: `Flan-T5`: Predictions and true labels (P/T) across definitions for 10 randomly selected gender-targeting samples with at least one beneficial flip (seed=10). H = hate speech, NH = not hate speech. Dom. = dominance.

ID	Sample	Subtype	Dom.	Own	Bulgaria(P/T)	Croatia(P/T)	Meta(P/T)	Reddit(P/T)	Th.I+E(P/T)	Th.I(P/T)
1	If I'm a b*tch then I'm the baddest b*tch	women	no	H	NH/NH	NH/NH	H/NH	NH/NH	NH/NH	NH/NH
2	I have no compassion at all for any trans person like you.	trans	no	H	H/NH	H/H	H/H	H/H	H/NH	H/NH
3	Men are a pest.	men	yes	H	H/NH	H/H	H/H	H/NH	H/NH	H/NH
4	It's obvious that men are the very opposite of stupid.	men	yes	NH	NH/NH	NH/NH	NH/NH	NH/NH	NH/NH	NH/NH
5	It's in everybody's interest to just slaughter those men.	men	yes	H	H/NH	H/H	H/H	H/NH	H/NH	H/NH
6	You men keep acting up. There really is nothing worse than your kind.	men	yes	H	H/NH	H/H	H/H	H/NH	H/NH	H/NH
7	I never said that women are sc*m.	women	no	NH	NH/NH	NH/NH	H/NH	NH/NH	NH/NH	NH/NH
8	I'm going to stab a woman tomorrow.	women	no	H	H/NH	H/H	H/H	H/H	H/NH	H/NH
9	Why would you think it's okay to call me a sl*t?	women	no	H	H/NH	H/NH	H/NH	H/NH	H/NH	H/NH
10	You and all the other trans people in this country make me boil with anger.	trans	no	H	H/NH	H/H	H/H	H/H	H/NH	H/NH

Table 8: `Qwen`: Predictions and true labels (P/T) across definitions for 10 randomly selected gender-targeting samples with at least one beneficial flip present when **`Flan-T5`** classified them (seed=10). H = hate speech, NH = not hate speech. Dom. = dominance.

G Prompts Used in Generative AI Assistance

Category	Example Prompts
Formatting LaTeX content	`Use this table [TABLE], containing the results of my experiments, and give me its contents in LaTeX format.` `How can I change the size of this figure to span both columns of my paper: [FIGURE]?`
Proofreading	`Report any grammatical errors or hard-to-read sentences from this paragraph: [PARAGRAPH].` `Report any spelling errors from this sentence: [SENTENCE].`
Code setup	`Give me steps on how I can activate the use of GPUs in Kaggle.` `Which field can be used to configure the settings of the LLM to prohibit sampling?`
Code documentation	`Write code documentation for the following function [PYTHON FUNCTION].`

Table 9: Example prompts used for generative AI assistance during the project.