A Modular Taxonomy for Hate Speech Definitions and Its Impact on Zero-Shot LLM Classification Performance

Matteo Melis[1], Gabriella Lapesa[2,3], Dennis Assenmacher[2]
[1]Department of Linguistics, Cognitive Science and Semiotics - Aarhus University,
[2]GESIS - Leibniz Institute for the Social Sciences
[3]Heinrich-Heine University Düsseldorf
[1]mmls@cc.au.dk, [2]first.last@gesis.org

Abstract

Detecting harmful content is a crucial task in the landscape of NLP applications for Social Good, with hate speech being one of its most dangerous forms. But what do we mean by hate speech, how can we define it, and how does prompting different definitions of hate speech affect model performance? The contribution of this work is twofold. At the theoretical level, we address the ambiguity surrounding hate speech by collecting and analyzing existing definitions from the literature. We organize these definitions into a taxonomy of 14 Conceptual Elements—building blocks that capture different aspects of hate speech definitions, such as references to the target of hate (individual or groups) or of the potential consequences of it. At the experimental level, we employ the collection of definitions in a systematic zero-shot evaluation of three LLMs, on three hate speech datasets representing different types of data (synthetic, human-in-the-loop, and real-world). We find that choosing different definitions, i.e., definitions with a different degree of specificity in terms of encoded elements, impacts model performance, but this effect is not consistent across all architectures.

1 Introduction

In a world that is becoming increasingly online-based, detecting harmful content, specifically Hate Speech (HS), is crucial for maintaining the integrity of the democratic discourse and freedom of speech (Kiritchenko et al., 2021; Tsesis, 2009). The advent of Large Language Models (LLMs) paved the way for a variety of new methods for detecting (Roy et al., 2023) and countering HS (Bonaldi et al., 2023), and for the creation of new artificial benchmarking data (Jin et al., 2024; Sen et al., 2023).

In particular, novel methods for classifying harmful content diverge from conventional supervised learning that relies on input/output pairs, but uses only predetermined prompts without examples (Plaza-del arco et al., 2023) or adding further information on the task (Roy et al., 2023).

A crucial role in refining prompts for zero-shot classification is played by the definition of the target construct, i.e., in the focus of this paper, *the definition of hate speech*.[1] As typical of social constructs of the social sciences, the definition of HS is ambiguous (Plaza-del arco et al., 2023; Waseem and Hovy, 2016) and cannot be easily framed in a static dimension. This is a relevant issue for the community, because it affects the interoperability of resources annotated at high cost, and the comparability of the results (and insights) drawn from their modeling, when, for example, different definitions are used for equivalent concepts (Fortuna et al., 2020).

The contribution of our work is twofold: conceptual/theoretical and experimental.

At the conceptual level, we contribute to structuring the conceptual landscape of HS by collecting and qualitatively organizing various definitions for hate speech. The goal of this analysis is to identify a set of Conceptual Elements (CEs), i.e., the conceptual building blocks present in the definitions, which encode their key dimensions. For instance, all definitions highlight its problematic nature (CE = Problematic Content) and specify that the target is an individual or group (CE = Target). However, only some definitions include potential consequences of hate speech (CE = Possible Implications) or acknowledge that it can be implicit (CE = Implicit Hate). This taxonomy serves as a scaffold for constructing and analyzing definitions, which we believe is a novel and practically valuable contribution to both the NLP and social science communities.

With these Conceptual Elements, we create a

[1]A construct is defined as "an idea or theory containing various conceptual elements, typically one considered to be subjective and not based on empirical evidence"(Oxford Languages).

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 426–457

three-layer taxonomy (Fig. 1) that we complement with a curated collection of definitions that arise from their combination. The collection of definitions can be seen as a structured, modular summary of the original set of definitions we reviewed and constitute a resource that we make available to the community for further experimentation.

The second contribution is experimental: starting from the idea that LLMs already encode extensive knowledge due to their pre-training and instruction-tuning (Zhang et al., 2023) we employ our definitions from the collection to carry out zero-shot prompting experiments on three hate-speech datasets, representing different data types: HateCheck (Röttger et al., 2021b, synthetic), Learning from the Worst (Vidgen et al., 2021, human-in-the-loop) and Measuring Hate Speech (Sachdeva et al., 2022, real-world examples). We employ three different LLMs: `LLama-3`, `Mistral` and `Flan-T5`. We conduct an in-depth error analysis and exploit the HateCheck fine-grained annotation regarding types of hate.[2]

Our results demonstrate the usefulness of our modular approach to build Hate Speech definitions as prompts for zero-shot classification. We find that varying construct definitions affects model performance, but this effect is not consistent across all model architectures and datasets. In some cases, more detailed definitions reduce false negatives, in others they primarily decrease false positives; more specifically, our error analysis shows that more detailed definitions improve performance in cases requiring nuanced distinctions between hate categories (i.e., Implicit Hate).

2 Related Work

Zero-shot prompting: general evaluation issues and application to HS With no need for computationally expensive fine-tuning, zero-shot prompting allows researchers to "just ask" a LLM to perform a task (e.g., classification). Unsurprisingly, this strategy is very frequently employed in scenarios with low computational power (e.g., social scientists with no access to fine-tuning infrastructures). The evaluation challenges related to zero-shot prompting have recently been explored in depth by Beck et al. (2022), who reported differences in robustness and sensitivity when prompting diverse socio-demographic information along with the evaluated tasks.

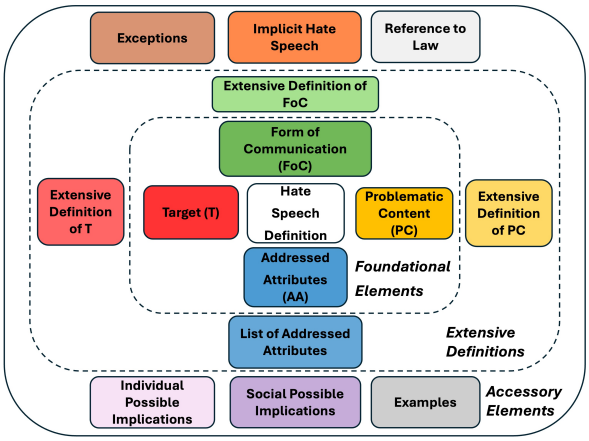


Figure 1: Taxonomy for Hate Speech definitions. To ease the readability of the work, the Conceptual Elements are color-coded. Refer to Appendix B for the full size figure.

Similar strategies have been observed in HS detection. Prompting LLMs with information on the task different from examples has improved performance in detecting HS (Roy et al., 2023; Plaza-del arco et al., 2023). Promising strategies involve prompting rationales or Chain-of-Thoughts alongside the task in zero-shot learning (ZSL), few-shot learning, or fine-tuning (Yang et al., 2023; Nghiem and Daumé III, 2024). These approaches have shown that in-context learning, particularly in the context of ZSL, is a worthwhile direction to explore (Ziems et al., 2024).

Construct Definition for HS detection Previous work explored how construct definitions can be utilized to obtain dataset-specific model-generated rationales (Nghiem and Daumé III, 2024). Other researchers explored how using a definition for an annotation task leads to more consistent answers among both human annotators (Ross et al., 2017) and LLMs (Li et al., 2024), affecting also their performance. Roy et al. (2023) investigated the effects of prompting different information (e.g., target, explanation) also among the HS construct definition. Their findings suggest definition-prompting led to mixed results, sometimes worsening and sometimes improving performance across various datasets.

Choosing an adequate definition to describe the construct of HS is challenging. There are overlapping and duplicate definitions (Fortuna et al., 2020), and sometimes conceptually different constructs are conflated, such as HS and Offensive

[2]The code can be found at: https://github.com/matteo-mls/Modular-Taxonomy-for-Hate-Speech-Definitions

Language (OL) (Davidson et al., 2017). In other cases, different constructs are put under the same umbrella, for example: HS, abusive and discriminatory language (Goldzycher et al., 2024). Furthermore, there seems to be minimal effort towards providing a more standardized definition. To the best of our knowledge, only Khurana et al. (2022) propose 5 criteria, taking also into account a legal perspective. In this work, we propose a taxonomy composed of 14 Conceptual Elements of which only three overlap with Khurana et al. (2022).[3]

3 A Taxonomy for Hate Speech Definitions

3.1 Procedure

We reviewed the HS literature over a substantial time span (2000–2021), with a focus on works that operationalized a definition to create datasets or corpora (9 definitions). Additionally, we selected two definitions from conceptual studies on HS (Tsesis, 2002; Nockleby, 2000) and two from works on HS detection (Mandl et al., 2021; Gao et al., 2017).

In total, we collected 20 HS definitions (see Appendix A) from the following sources:

- 13 definitions from literature (Sachdeva et al., 2022; Vidgen et al., 2021; Mandl et al., 2021; Röttger et al., 2021b; Basile et al., 2019; Gibert et al., 2018; Founta et al., 2018; Davidson et al., 2017; Gao et al., 2017; Nobata et al., 2016; Warner and Hirschberg, 2012; Tsesis, 2002; Nockleby, 2000);
- 3 definitions from social networks policies (Twitter/𝕏, Facebook, Youtube);
- 2 definitions automatically generated by LLMs (ChatGPT, Gemini);
- 2 definitions from official documents (UN Strategy and Plan of Action on Hate Speech, Code of Conduct between European Union Commission and companies, Wigand and Voin, 2017);

Using these definitions, we inductively identified 14 CEs (building blocks of the HS construct) which we organize in three layers (see Appendix B for a visual representation).

Defining the taxonomy presents two key challenges. First, distinguishing Offensive Language (OL) from Hate Speech (HS) is complicated by a confounding effect noted by Davidson et al. (2017) and Waseem and Hovy (2016), where OL and HS overlap. We clarify that while OL can exist without being HS, any content classified as HS must also be considered OL. Second, avoiding circular definitions is crucial (i.e., a definition that relies on another definition to be understood). While 'protected groups' are often used to differentiate HS from OL, and this approach has legal relevance (Khurana et al., 2022), using this as a defining criterion, would mean defining HS by using another definition, which varies in relation to culture, laws and people's sensitivity. Definitions in prior work often rely on the identification of protected groups (Gibert et al., 2018). However, based on the researcher's choice, the protected groups can be listed in the definition,[4] but we do not recommend to use them as defining factor. Instead, our approach shifts the focus from only enumerating categories to explicitly describing the dynamic of attacking a target based on some "inherent characteristics that are attributed to that group and shared among its members".

3.2 Taxonomy

Below, we describe each of the three layers of CEs. A detailed description of each CE can be found in Appendix C. Table 1 illustrates the different Conceptual Elements and corresponding abbreviations.

Foundational Elements We label the most common and therefore most important CEs as Foundational Elements, being essential for constructing a meaningful definition of HS. These elements include: Form of Communication (FoC), Target (T), and Problematic Content (PC).

However, considering the challenge of distinguishing HS from OL (Davidson et al., 2017; Waseem and Hovy, 2016), we added another Foundational Element, the Addressed Attributes (AA). This element reflects the explicit relationship between the target and the inherent or perceived characteristics being attacked (e.g., attacking someone based on the belief they follow a specific religion).

These four Conceptual Elements—FoC, T, PC, and AA—together form the basis of a foundational HS definition, which from now on we will refer to as the Hate Speech Base (HSB) definition, and represent the minimal conceptual units that are consistently present in almost all hate speech definitions

Extensive Definitions of the Foundational Elements Within the second layer, four Conceptual

[3] Two of them are what we call the "Target" and the "Problematic Content " and the third is "Possible Implications".

[4] In what we later define as List of Addressed Attributes (LAA)

Elements provide additional detail about the core components, including: Extensive Definitions of Form of Communication (EDFoC), Target (EDT), and Problematic Content (EDPC), as well as the List of Addressed Attributes (LAA). These CE capture richer or more granular information about the same dimensions present in the previous layer.

Accessory Elements The remaining six elements are categorized in the third layer and provide different information from the core components of the construct, in other words, new information: social Possible Implications (sPI), individual Possible Implications (iPI), Exceptions (Exc), Implicit Hate Speech (IHS), Examples (Exa), Reference to Laws (Law).

3.3 Building definitions from the taxonomy

Based on the CEs and their modular arrangement within the taxonomy, we generated a collection of definitions by recombining them according to the criteria outlined below.

First, we created content reflecting each CE. For example, the CE: *Target* would be mapped in the natural language expression "toward a group or an individual" while the corresponding, more informative CE: *Extensive Definition of Target* would map into "toward a group or an individual" followed by "which is thought to be a member of that group".

Second, we combined these elements to create definitions with varying conceptual compositions, aiming to represent different levels of informativeness (level of details of the definition[5]) and types of information (i.e., the specific mention of implicit HS). When combining the CEs to create definitions, we made sure they would not differ in style or wording: for instance, the textual span representing the CE Target is exactly the same in all the definitions.

Table 1 lists all the CEs, their abbreviations, and how they are reflected in the definitions we created. In Appendix D we showcase the presence or absence of CEs in all the definitions. The full set of definitions contained in our collection is reported in Appendix E.

While building the collection of definitions, which was designed for the goal of prompting, we consolidated various forms of potential implications into a single CE: PI (Possible Implications). Additionally, we excluded two CEs—Examples and Reference to Law. The former was omitted to preserve the zero-shot learning (ZSL) condition, as including examples would shift the setup toward few-shot learning. The latter was excluded because assessing models' legal domain knowledge falls outside the scope of this study.

We emphasize that there is no direct one-to-one correspondence between the original definitions used to develop the taxonomy and the definition collection derived from it. Instead, our collection serves as a structured summary of existing definitions, with carefully curated wording to ensure that variation stems solely from different combinations of CEs. This makes it an ideal starting point for the prompting experiments presented in the next section.

4 Zero-shot prompting

4.1 Experimental setup

Datasets In our zero-shot experiments, we use three different datasets reflecting different data types:

1. HateCheck (Röttger et al., 2021b): synthetically generated functional test-suite for HS;
2. Learning from the Worst (LFTW, Vidgen et al., 2021): curated collection of challenging HS through a human-in-the-loop process;
3. Measuring Hate Speech (MHS, Sachdeva et al., 2022): real-world instances of HS collected from various social media;

These three datasets not only represent different types of data points but also adopt operational definitions of hate speech that align with the Foundational Elements outlined in our taxonomy, ensuring a meaningful and consistent interpretation of hate speech. For reasons of better comparability and to avoid unnecessary computational costs, we randomly sampled from LFTW and MHS the same amount of data-points (3901) with the same distribution among classes (68.16% Hate Speech, and 31,84% Not-Hate Speech) of HateCheck. Which we have taken as a reference point due to its structure, which differentiates between all these different functionalities (challenging types of hate), enabling us to investigate them in our error analysis.

Models In our experiments, we employ three open-source, instruction-based LLMs of small to medium sizes from different model families: `Meta-Llama-3-8B-Instruct`, `Mistral-7B-Instruct-v0.2`, and `Flan-T5-XL`, all sourced

[5]We assume that adding more Conceptual Elements leads to higher level of detail

Conceptual Element	CE	Example in definition
Foundational Conceptual Elements		
Form of Communication + Target + Problematic Content = **Offensive Language**	FoC + T + PC = **OL**	Hate Speech is considered any kind of content that conveys malevolent intentions toward a group or an individual.
Form of Communication + Target + Problematic Content + Addressed Attributes = **Hate Speech Base**	FoC + T + PC + AA = **HSB**	Hate Speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members.
Extensive Definition of the Foundational Elements (Step 1)		
Hate Speech Base + Extensive Definition Form of Communication	**HSB** + EDFoC	Hate speech is considered any kind of content or communication expressed using language (written or spoken) or actions, that convey malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members.
Hate Speech Base + Extensive Definition Target	**HSB** + EDT	Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual which is, or thought to be, a member of that group, and motivated by inherent characteristics that are attributed to that group and shared among its members.
Hate Speech Base + Extensive Definition Problematic Content	**HSB** + EDPC	Hate speech is considered any kind of content that conveys malevolent intentions such as statements of inferiority, aversion, cursing, calls for exclusion, threaten, harass or violence, toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members.
Accessory Elements (Step 2)		
Hate Speech Base + List of Addressed Attributes	**HSB** + LAA	Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics.
Hate Speech Base + Possible Implications	**HSB** + PI	Hate speech is... The outcome of Hate Speech could be the promotion of division among people, undermining of social cohesion in communities, inciting others to commit violence or discrimination, and could have consequences for individuals' health and safety.
Hate Speech Base + Exception	**HSB** + Exc	Hate speech is... However, even if it is offensive, it is not considered Hate Speech any content that attacks a person's personality traits, ideas, or opinions.
Hate Speech Base + Implicit Hate Speech	**HSB** + IHS	Hate speech is... Hate Speech can also be implicit, portrayed as an indirect or coded language that uses Irony, Stereotypes, or Misinformation.

Table 1: Colour coded Conceptual Elements and examples in the definitions prompted in the HS detection task.

from HuggingFace. While for `LLama-3` and `Mistral` (Jiang et al., 2023), there is no clear information about their pre-training and fine-tuning data, we are only certain that `Flan-T5` (Chung et al., 2022) is the only model which was not exposed to any of these particular datasets (though being instruction-tuned on some other hate speech/toxicity datasets, Wang et al., 2022)[6]

[6]Only for `Mistral`, due to numerous instances in which it refused to answer, we used `outlines` by Willard and Louf (2023), a library that allows the user to retrieve a structured generation from the LLMs. This has set the model's temperature to its default value, 0.7, while for the models the temperature was set to 0.95.

Prompting Strategy We framed the task as a binary classification task (HS/No Hate Speech (NHS)), keeping the instruction as brief and concise as possible (Weber et al., 2023; Chang et al., 2024), Appendix F showcase the resulting prompts.

To systematically reduce the number of CE combinations, we followed a two-step approach. First, in **Step 1**, we refine the definition of Hate Speech Base (HSB), recognizing its central role in our study. We focus on identifying which of the Extensive Definitions of—Form of Communication (FoC), Target (T), Problematic Content (PC), and Addressed Attributes (AA)— provide the most informative input for the models. Secondly, in **Step 2**, we test the best-performing definition from the

Step 1 (highest macro-F1 score) by incorporating additional Accessory Elements: the List of Addressed Attributes (LAA), Possible Implications (PI), Implicit Hate Speech (IHS), and Exceptions (Exc).

For evaluation, we also include: a) each dataset's respective construct definition (referred to as "Own"), as we expected these definitions to be most reflective of the dataset's specific data points, and b) a condition in which no definition is given ("NO"), but the model is only asked to classify if the data-point is Hate Speech or not.

5 Results

As outlined in Sec. 4.1, our experiments followed a two-step approach:[7]

Step 1: Which Extensive Definitions provide the most informative refinement of the Hate Speech Base definition?

Step 2: How does incorporating additional Conceptual Elements impact the results from Step 1?

5.1 Step 1: What is the Best Base Refined Definition for Hate Speech?

Table 2 presents macro-F1 scores for different models and datasets, along with correlation values between performance and definition informativeness.

LLama-3 performs best without any definition (NO) in two out of three datasets, suggesting potential data leakage from HateCheck. While in the LFTW dataset, we encounter the only instance in which the best definition is the one of the dataset itself (Own). Among the crafted definitions, HSB + EDT performs best for HateCheck and LFTW, while HSB + EDFoC + EDPC is optimal for the MHS dataset.

Mistral achieves its highest scores with either NO definition or Offensive Language (OL), implying an internalized concept of hate speech that aligns with offensive language. Among crafted definitions, HSB + EDFoC performs best in two datasets, while HSB is most effective in HateCheck.

Flan-T5, unlike the other models, benefits consistently from definition prompting. Performance improves as definitions become more detailed, with HSB + EDT yielding the highest results in HateCheck and MHS, while the most extensive definition (HSB + EDFoC + EDPC + EDT) is optimal for LFTW.

[7]To ensure stability, each experiment was repeated three times.

5.2 Step 2: Adding more Conceptual Elements to the optimal base definition

Table 3 presents the results of combining accessory elements with the best-crafted definition from step one.

LLama-3 improves in performance on the crafted definitions only in LFTW and Measuring Hate Speech, with the former surpassing the best performing definition (Own) of the previous step with +LAA.

Mistral contrary to the previous step, is the most positively affected, improving its performance in different conditions over all the datasets concerning not only the crafted definitions. Reaching its new best performance in HateCheck with and LFTW both with +LAA + PI + IHS.

Flan-T5 shows an opposite trend compared to the previous step, where definition prompting has always led to an improvement in performance, here we do not observe in any condition a further increase in performance, though all the results are still higher than the condition without definition.

Ultimately, we observe two consistent trends across the three datasets. `Mistral` improves only on the second step, when additional elements are added to the construct of HS, or in other words, some specificity of information is added to the definition. While `Flan-T5` shows improvement only in the first step, being thus more sensitive to the level of detail/informativeness of the definition, being also the only model which shows a positive correlation between performance and length of the definition (Table 2).

Performance-wise, we observe that while all models behave differently, their trends remain consistent across datasets. `LLama-3` generally does not show improvement, with a performance increase occurring only once on the LFTW dataset. In contrast, `Mistral` consistently improves in the second step, while `Flan-T5` shows gains in the first step, indicating that these models are more responsive to different types of information. `Mistral` benefits from more specific details, such as references to implicit HS, whereas `Flan-T5` responds to broader definitional refinements.

As a sanity check, we include additional analyses in the Appendix: robustness, which examines the stability of model performance across different runs (Appendix G), and sensitivity, which measures how much model responses vary when different definitions are applied (Appendix H).

Definitions	HateCheck			LFTW			MHS		
	LLama3	Mistral	FlanT5	LLama3	Mistral	FlanT5	LLama3	Mistral	FlanT5
NO	**84.82**	**78.57**	72.18	72.07	56.05	60.99	**75.94**	**79.12**	74.21
Own	76.72	75.10	75.95	**73.86**	53.83	62.43	74.17	77.10	74.79
OL	77.62	**78.57**	74.40	71.75	**57.28**	62.54	69.08	76.80	74.63
HSB	80.02	*78.20*	74.91	72.63	55.78	63.66	70.72	75.81	74.30
HSB_EDFoC	80.04	77.77	75.18	72.87	*55.82*	63.41	72.00	*77.14*	75.21
HSB_EDPC	78.90	76.40	75.11	71.95	54.72	63.32	73.24	76.09	74.77
HSB_EDT	*80.14*	77.17	***76.29***	73.42	54.19	63.83	72.04	75.59	***75.54***
HSB_EDFoC_EDT	79.99	76.66	75.66	73.31	54.78	63.65	72.61	75.98	75.38
HSB_EDFoC_EDPC	80.01	76.44	74.71	72.33	55.04	62.85	*73.99*	76.40	74.85
HSB_EDT_EDPC	79.59	75.58	75.97	72.52	53.17	63.76	73.77	75.76	74.58
HSB_EDFoC_EDPC_EDT	80.06	75.54	76.21	72.64	53.46	***64.19***	73.94	76.70	75.15
Pearson Corr. (tokens)	-0.05	-0.96	0.62	-0.10	-0.59	0.67	0.70	-0.26	0.35
Best Conceptual Elements	EDT	-	EDT	EDT	EDFoC	EDs	EDFoC_EDPC	EDFoC	EDT

Table 2: Step 1, F1-macro: In **bold** the highest score, the underlined score is the chosen one for the second step. The Correlation Coefficients do not consider the condition without definition (NO). (Own = Definition of the dataset the model is being tested on, OL = Offensive Language, HSB = Hate Speech Base, ED = Exstensive Definition, FoC = Form of Communication, PC, Problematic Content, T = Target).

Definitions	HateCheck			LFTW			MHS		
	LLama3	Mistral	FlanT5	LLama3	Mistral	FlanT5	LLama3	Mistral	FlanT5
+LAA	79.70	76.87	75.69	**74.24**	56.05	63.04	73.12	77.91	74.96
+LAA_PI	77.39	78.44	**75.95**	73.16	58.40	**63.84**	72.09	77.71	**75.30**
+LAA_Exc	**79.72**	75.95	75.30	72.96	54.22	62.31	**74.67**	76.16	74.61
+LAA_IHS	77.42	80.74	75.66	72.97	60.97	63.27	71.53	**78.22**	74.91
+LAA_PI_Exc	76.65	73.88	75.38	73.22	53.00	62.99	73.60	75.56	75.14
+LAA_Exc_IHS	78.17	78.27	75.76	73.72	56.38	63.34	74.55	77.61	74.81
+LAA_PI_IHS	76.03	**81.69**	75.37	72.06	**62.17**	62.95	71.43	78.06	74.48
+LAA_PI_IHS_Exc	77.48	78.62	75.70	72.71	57.92	62.80	72.30	77.88	75.00

Table 3: Step 2, F1-macro: in **bold** the highest score in the step, the underlined scores are those which are higher than the chosen crafted definitions of Step 1. Scores underlined twice are higher than the best performing definition of Step 1. (+ = best performing definition from Step 1, LAA = List of Addressed Attributes, PI = Possible Implications, Exc = Exception, IHS = Implicit Hate Speech).

6 Error Analysis

Hate Speech vs. Not Hate Speech A distinct model-dependent trend is evident in both the HS and NHS classes across all datasets. As shown in Fig. 2, `LLama-3` frequently misclassifies NHS instances (i.e., non-hateful content) as HS, resulting in a higher false positive rate. This tendency appears to intensify when a definition is added to the prompt (e.g., changing from NO to Own). At the same time, introducing any definition reduces the number of misclassified HS instances (false negatives), suggesting that the model follows a more conservative classification approach. `Mistral`, on the other hand, exhibits the opposite tendency, frequently misclassifying HS instances as NHS. However, adding a definition to the prompt reduces the number of false positives (misclassified NHS instances). `Flan-T5` maintains a more balanced classification pattern but shows a higher false negative rate, especially when tested on human-in-the-loop data points.[8]

Analysis of HateCheck functionalities HateCheck is a test suite for HS with a comprehensive labeling structure which defines each data point's functionality— the specific type of hate conveyed. In this section, we conduct a micro-analysis on these HS functionalities to examine how classification performance changes across them when prompted with different CEs. The authors of HateCheck (Röttger et al., 2021b) identify 29 distinct functionalities. For easier comparability, we have grouped them into five macro classes: HS, NHS,

[8]Appendix I contains detailed graphs of the error distribution across all conditions.

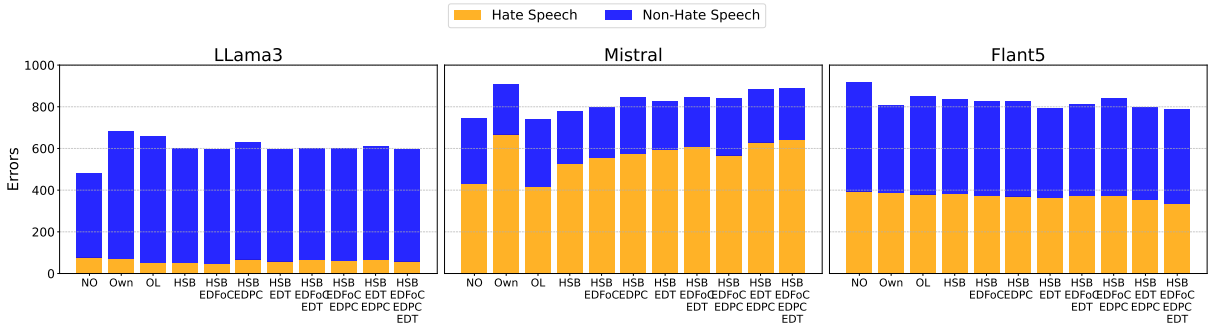


Figure 2: Distribution of errors across the three models on HateCheck.

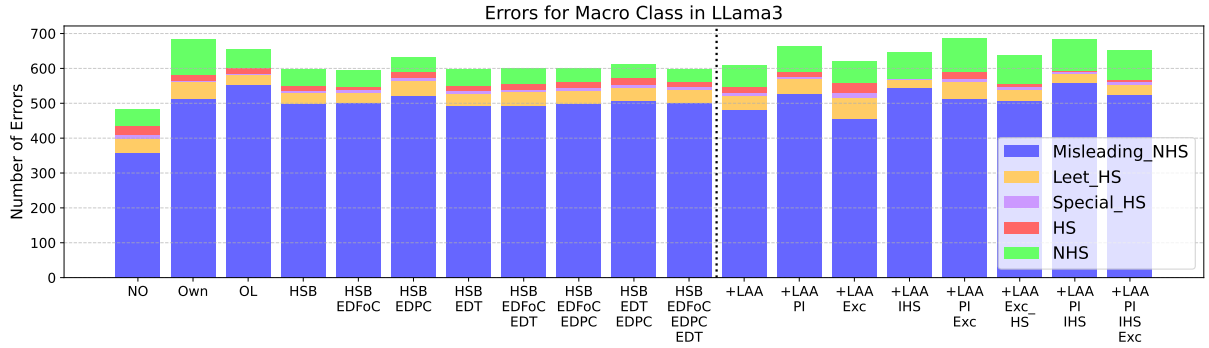


Figure 3: Distribution of errors in LLama-3 across Hatecheck Classes

Misleading NHS, Leet HS, and Special HS. Appendix J provides explanations and details on this grouping.

The error distribution across all functionalities [9] reveals that models struggle the most with instances designed to counter HS, such as statements like *"If you say 'I hate gay people,' then you are a bigot!"*. Even the best-performing model, `LLama-3`, misclassifies 93.75% of these cases, suggesting that LLMs still rely on specific words or sentence structures when making classifications.

Next, we examine how different definitions influence classification performance across the Macro Classes. Figure 3 illustrates the variation in classification errors for `LLama-3` across all definitions. Overall, a more detailed definition tends to improve classification accuracy for general HS and NHS, with a slight positive effect on Leet HS and Special HS. However, it also leads to an increase in errors for the Misleading NHS class (Appendix L presents results for all models). Furthermore, we find that adding a CE specific to a class of instances reduces classification errors for that class. For example, explicitly informing the model that some statements may be offensive but not hateful (i.e., defining exceptions) improves performance in the Misleading NHS class. A similar effect is observed for implicit hate speech: clarifying that hate speech can be conveyed through coded language, irony, or sarcasm leads to performance gains in the Special HS class. This effect is most pronounced in `Mistral`, the only model that consistently improves in the second step (see Table 3). We also observe a partial effect in `Flan-T5`, though it never improves in the second step [10]. Table 4 presents a detailed breakdown of these effects. These findings have a potential relevance for content moderation, which we discuss in the conclusion.

Mistral	NO	+IHS	+Exc	+Exc+IHS
Misleading NHS	34.81%	28.36%	**23.26%**	26.04%
Leet HS	20.81%	**19.76%**	29.23%	24.24%
Special HS	18.75%	**15.71%**	27.14%	21.84%

Table 4: Error percentage, Conceptual Elements & macro classes in Mistral.

7 Conclusion

In this work, we explored the conceptualization of the construct definition of HS and its influence on zero-shot prompting on three datasets. Our starting point has been the review of existing HS

[9] An overview of all errors are presented in Appendix J.

[10] Appendix M provides detailed results for all models.

definitions, from which we inductively derived a set of Conceptual Elements. We then combined the different elements in the taxonomy to build a collection of definitions that lend themselves as prompts for LLM modeling. Thus, the taxonomy and the collection definition are not just a conceptual contribution of our work, but also a concrete resource that can and should be used by researchers to structure their operationalization of the HS construct, thereby contributing to a clearer research landscape. Furthermore, the three-layers taxonomy, allows for combinations reflecting different levels of detail, which can be employed in annotation tasks in the descriptive vs. prescriptive paradigms (Röttger et al., 2021a).

In our experiments, we exploited the definition collection for a series of zero-shot experiments, with the definitions serving as a series of curated prompts with increasing level of details.

Our results show that varying construct definitions affects model performance, in a complex constellation of patterns. Some models benefit from detailed construct definitions by reducing false negatives, while others primarily decrease false positives. Our micro-analysis of different HateCheck functionalities shows that incorporating specific Conceptual Elements targeting particular types of hate improves model performance, especially in cases requiring nuanced distinctions between hate categories. Given our findings, we recommend that such a modular inspection of possible definitions should be employed for other complex constructs, beyond HS. Moreover, our findings do have practical implications for the usage of LLMs in production. Models that benefit from detailed construct definitions by reducing false negatives are, for example, better suited for high-recall moderation strategies, ensuring that fewer instances of hate speech go undetected. Conversely, models that primarily lower false positives are more appropriate for high-precision approaches, minimizing the risk of over-flagging benign content. By strategically refining definitions and incorporating targeted Conceptual Elements, moderation systems can be optimized to balance recall and precision according to platform-specific goals.

8 Limitations

Our study is not without limitations. A first one stems from computational restrictions. We were unable to test the largest model variants and assess their stability when prompted with different construct definitions. Furthermore, due to these computational constraints, we did not experiment with all possible construct definition combinations and settled on one fixed extensive definition. There is a possibility that different variants could have led to better performance.

We also acknowledge that semantically different realization of the Conceptual Elements could have had a different impact on the models' performance. In other words surface-level phrasing, even when underlying CEs are held constant, can influence model behavior, an example of this can be seen on the MHS dataset, where the Own definition contains the exact same CEs of the HSB definition, though leading to different results.

Another limitation is tied to the effect we have found in Sec. 6. This is limited to the HateCheck datasets, to actually prove if this is a general effect, further studies should be conducted in annotated datasets. Our work only investigates performance differences in a zero-shot setting. It would be interesting to explore how carefully selected few-shot examples adhering to the given construct definition might impact stability and performance. Finally, we acknowledge that even the formulation of the prompt, without considering the construct definition itself, may influence the model's final performance.

9 Ethical Considerations

In our experiments, we do not collect any data, we instead use publicly available resources, so to ensure data protection. We also acknowledge that using Large Language Models to detect Hate Speech is not safe from issues, potentially not filtering appropriately and ending up spreading even more biases and discrimination. Furthermore, we made every effort to minimize content that could be disturbing or offensive, ensuring that any necessary reporting is handled appropriately and responsibly.

10 Acknowledgments

The authors thank the funding received from the European Union's Horizon Europe Marie Skłodowska-Curie Actions Doctoral Networks, under Grant Agreement No. 101167978. This work was also conducted as part of the project Digital Dehumanization: Measurement, Exposure, and Prevalence (DeHum), supported by the Leibniz Association Competition (P101/2020).

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2022. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 13–28, Prague, Czechia. Association for Computational Linguistics.

K. Chang, S. Xu, C. Wang, Y. Luo, T. Xiao, and J. Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

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

M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

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.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

S. Ghosh, M. Suri, P. Chiniya, U. Tyagi, S. Kumar, and D. Manocha. 2023. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. *arXiv preprint arXiv:2303.03387*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

J. Goldzycher, P. Röttger, and G. Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from gahd, a german hate speech dataset. *arXiv preprint arXiv:2403.19559*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Yiping Jin, Leo Wanner, and Alexander Shvets. 2024. GPT-HateCheck: Can LLMs write better functional tests for hate speech detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7867–7885, Torino, Italia. ELRA and ICCL.

Urja Khurana, Iris Vermeulen, Eric Nalisnick, Marcel 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.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Lei Li, Lianyang Fan, Subin Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

H. Nghiem and H. Daumé III. 2024. Hatecot: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. *arXiv preprint*, arXiv:2403.11456.

Chikashi Nobata, Joel Tetreault, Andrea Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

John T. Nockleby. 2000. Hate speech. In Leonard W. Levy, Kenneth L. Karst, et al., editors, *Encyclopedia of the American Constitution*, 2nd edition, pages 1277–1279. Macmillan, New York.

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, Toronto, Canada. Association for Computational Linguistics.

Benedikt Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nikolay Kurowsky, and Max Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint*, arXiv:1701.08118.

P. Röttger, B. Vidgen, D. Hovy, and J. B. Pierrehumbert. 2021a. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021b. 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.

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, Singapore. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Alexander Tsesis. 2002. *Destructive messages: How hate speech paves the way for harmful social movements*, volume 27. NYU Press.

Alexander Tsesis. 2009. Dignity and speech: The regulation of hate speech in a democracy. *Wake Forest Law Review*, 44:497.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, and D. Khashabi. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

L. Weber, E. Bruni, and D. Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. *arXiv preprint arXiv:2310.13486*.

C. Wigand and M. Voin. 2017. Speech by commissioner jourová—10 years of the eu fundamental rights agency: A call to action in defence of fundamental rights, democracy and the rule of law.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, and S. Y. Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. *arXiv preprint*, arXiv:2311.00321.

Shuohang Zhang, Li Dong, Xiang Lisa Li, Shuming Ma, Xiaodong Sun, Shaohan Huang, Furu Wei, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *arXiv preprint*, arXiv:2308.10792.

C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Collection of Hate Speech definitions

Authors	Definition
Nockleby, 2000	Any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.
Tsesis, 2002	Hate speech provides the "vocabulary and grammar depicting a common enemy," and establishes a "mutual interest in trying to rid society of the designated pest."
Warner and Hirschberg, 2012	Hate speech is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.
Nobata et al., 2016	An act that attacks or demeans a group/individual based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity.
Davidson et al., 2017	Hate speech is language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group.
Gao et al., 2017	Hateful speech is defined as the language which explicitly or implicitly threatens or demeans a person or a group based upon a facet of their identity such as gender, ethnicity, or sexual orientation.
Founta et al., 2018	Hate speech is language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.
Gibert et al., 2018	Hate speech is any communication that disparages a target group of people based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.
Basile et al., 2019	Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics.
Mandl et al., 2021	Hate speech includes ascribing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). This class combines any hateful comments toward groups because of race, political opinion, sexual orientation, gender, social status, health condition, or similar.
Röttger et al., 2021b	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, which broadly reflects international legal consensus (particularly the UK's 2010 Equality Act, the US 1964 Civil Rights Act, and the EU's Charter of Fundamental Rights).
Vidgen et al., 2021	Hate is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.

Sachdeva et al., 2022	Hate speech, defined as "bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics, especially when the group is unnecessarily labeled"
UN Strategy and Plan of Action on Hate Speech	Any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor.
Code of Conduct between European Union Commission and companies	All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic.
ChatGPT's definition	Hate speech typically refers to any form of communication – whether spoken, written, or expressed through actions – that seeks to demean, intimidate, discriminate against, or incite violence or prejudice against individuals or groups based on characteristics such as race, ethnicity, nationality, religion, gender identity, sexual orientation, disability, or any other immutable characteristic.
Gemini's definition	Hate speech is basically language that attacks a person or group based on things they can't control, like their race, religion, gender, sexual orientation, or disability.
Facebook	We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.
Twitter/X	You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
Youtube	We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran Status.

B Taxonomy: visual representation

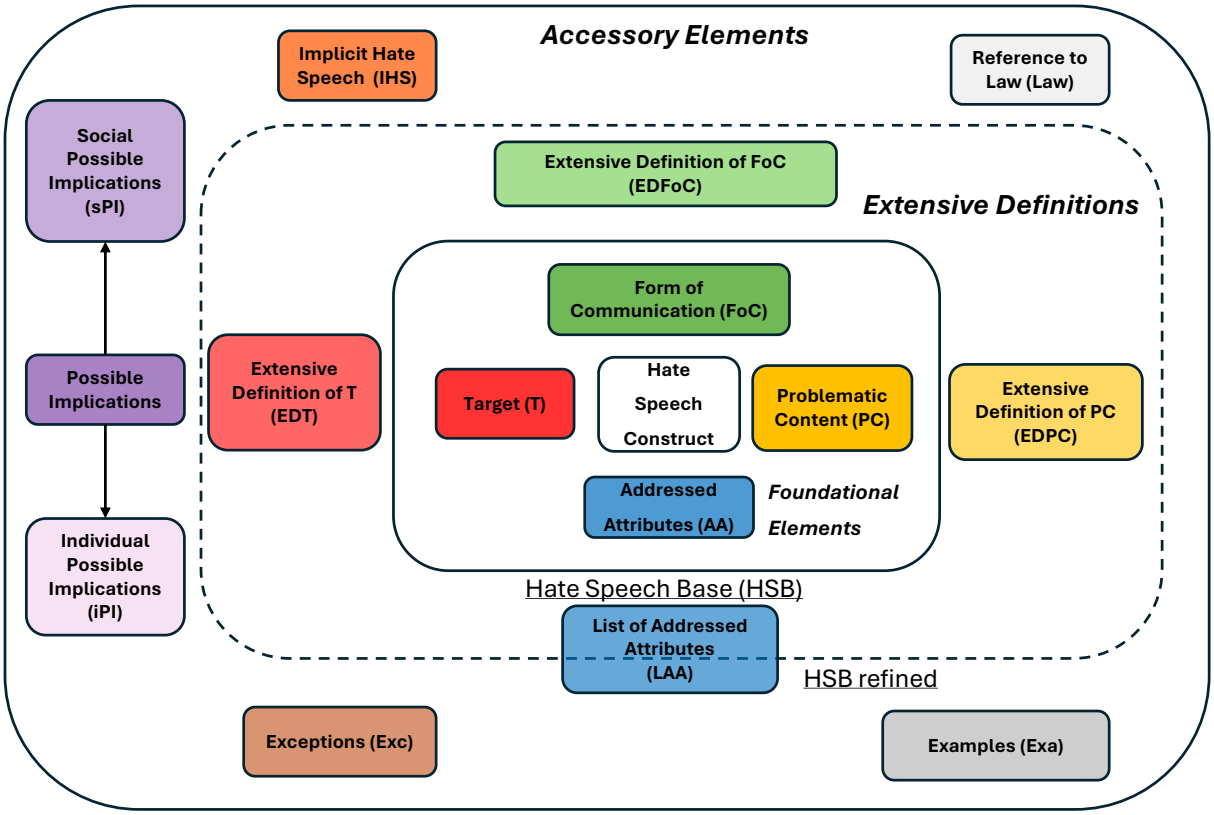


Figure 4: Graphic visualization of the taxonomy.

C Conceptual Elements in our Taxonomy: Definitions

FOUNDATIONAL ELEMENTS: We define Foundational Elements as those Conceptual Elements that are required to build a meaningful definition of Hate Speech. In this category we find:

Form of Communication (FoC): represents how the message is expressed. Refers to the modality of transmission, it can be text, speech, image, or video, . . . This element is independent from all the others, it is what grounds the HS to the real world and it is the whole "container" of the HS.

Target (T): represents toward whom the FoC is directed (individual or group). It describes the real word entity that is addressed by the FoC. We can identify it as the object of the message. In the FoC it is often identified as a social category ("black people are. . . "), a slur that identifies a member of that category or the category itself ("n-word").

Problematic Content (PC): represents realisation of the malevolent communicative intent conveyed by a specific FoC. It describes that part of the FoC that has a negative connotation and it is implied to be a derogatory descriptor of T. It is the form (in our case, linguistics) in which the malevolent communicative intent is expressed in the FoC. It can be a sentiment ("I hate..", "I can't bear. . . ") a slur ("gay people are all dumb"), or anything that implies negativity toward the T.

PC and T can assume multiple forms and sometimes overlap (e.g., n-word, f-word), and they are both dependent form the FoC — without it there cannot be PC and T.

Addressed Attributes (AA): represents that part of the FoC that is specific to the Hate Speech and explicitly describes the relation between PC and T. In other words it describes which are the aspects of the

T that motivate the malevolent communicative intent and thus the creation of a PC. It describes that the malevolent communicative intent has to specifically aim to a group or a person that belongs to a group and to the inherent characteristics that the group and the individual share or are thought to share. Thought being part of the definition, it can also be found in HS comment: being explicitly expressed in the FoC (”I hate black people [skin color]”) or take the form of a generalisation (”[All] disabled people are stupid”) or can be left implicit overlapping with the other elements (for instance i with the Target: ”[Affirmative action] means we get affirmatively second rate doctors and other professionals”).

Being these the Foundational Elements of the construct definition of Hate Speech, different combination of them will lead to constructs different than HS, here below we provide four examples of different combination.

1. If PC and AA are missing, the communication (FoC, T) is not Hate Speech, but it is just communication.
2. If T and AA are missing, the communication (FoC, PC) **can** still be offensive (or toxic), but not categorized as Hate Speech (i.e., “this is bul***it”, “Cauliflowers are fu**ing disgusting”).
3. If AA is missing, the communication (FoC, T, PC) it is not Hate Speech but Offensive Language (“[POLITICIAN NAME] is the dumbest politician in the US").
4. There are no cases in which there are only PC (FoC, PC, AA) and AA or T and AA (FoC, T, AA). This makes AA dependent from PC and T (other than from the FoC). It comes that, when it seems to have a case of this kind, actually AA overlaps with the apparent “missing Conceptual Element" (i.e., AA overlaps with PC “you are a [f-word]”).

EXTENSIVE DEFINITIONS: Are those elements that provide further information about the construct, and can be used go implement further levels of details/informativeness of the construct definition. First we have identified a group of Conceptual Elements, Extensive Definitions (EDs) that add further information to the Foundational elements, In other words they do not provide different pieces information from those already in the definition (HSB), but only describe more in details the pieces of information provided by the Foundational Elements.

.

Extensive Definition Form of Communication (EDFoC): other ways to describe the FoC, in our case, it is important that it is explained as text or language or communication, however, this Accessory element provides another way in which the Hate Speech can be transmitted (e.g.,“Hate speech can manifest in various forms including but not limited to verbal attacks","any form of communication – whether spoken, written, or expressed through actions").

Extensive Definition Target (EDT): specifies the relation between a person and being a member of a group, the idea of “belonging to a group”.

Extensive Definition Problematic Content (EDPC): it gives more information and better describes PC, providing examples of what is considered PC; (i.e., ”We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation.”).

List of Addressed Attributes (LAA): it provides a list of characteristics/attributes of the T that can be object of the PC (i.e., "such as race, gender, religion, ..").

ACCESSORY ELEMENTS: Finally, we define as Accessory Elements those elements that provide different information on the construct of HS, namely, information that it is not present in the HSB definition and describes other aspects of the HS construct. In

Possible Implications (PI): part of the FoC that refers to the possible consequences of a particular combination of PC and T. It can be divided into two sublevels:

1. *social (sPI):* it refers to the implication on the social level of one (or more) PC toward a T (i.e., "undermines social cohesion, promotes division ... in communities").

2. *individual (iPI):* it refers directly to the effects that one (or more) PC can have on the T (i.e., "can have serious consequences for individuals, often perpetuating discrimination, hostility, and violence").

Exceptions (Exc): provide information on what is not considered HS (i.e., "attacks on people's personality traits, ideas, or opinions").

Implicit Hate Speech (IHS): Hate speech is not always explicit, this conceptual element describes what is considered Implicit Hate Speech, conceptually a communication that is missing a conceptual element among Target, Problematic Content and Addressed Attributes. To define this conceptual element we have been inspired by Ghosh et al., 2023 and ElSherief et al., 2021.

The following two Conceptual Elements were not implemented in our experiment. The first in order to maintain a Zero-Shot-Learning condition, while the second would have implied an to investigate if the models actually knows the laws that we are referring to, and this was not in the scope of our research.

Examples (Exa): the information provided by this CE is simply an instance of a sentence that it is considered Hate Speech.

Reference to laws: (Law): part of the definition that provide information in regards to specific laws that regulate Hate Speech.

D Conceptual Elements in HS the literature: overview

Author	FoC	T	PC	AA	EDFoC	EDT	EDPC	LAA	sPI	iPI	Exc	IHS	Exa	Law
Nockleby, 2000	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Tsesis, 2002	✓	?	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Warner and Hirschberg, 2012	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Nobata et al., 2016	?	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Davidson et al., 2017	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
Gao et al., 2017	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Founta et al., 2018	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
Gibert et al., 2018	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Basile et al., 2019	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Mandl et al., 2021	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Röttger et al., 2021b	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
Vidgen et al., 2021	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Sachdeva et al., 2022	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Twitter/X	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗
Facebook	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗
YouTube	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
ChatGPT	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗
Gemini	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓	✗	✓
UN Strategy & Plan of Action on HS	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
Code of Conduct EU (Wigand and Voin, 2017)	✓	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗

Table 5: Outline of the Conceptual Elements in the collected definitions.
✓= present in the definition, ✗= absent, ? = present but we consider it too vague to be part of a definition.

E Collection of Definition Prompted in the Experiment

OL - Offensive Language
Hate Speech is considered any kind of content that conveys malevolent intentions toward a group or an individual.
HSB - Hate Speech Base
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDFoC - Hate Speech Base + Extensive Definitions of Form of Communication
Hate speech is considered any kind of content or communication expressed using language (written or spoken) or actions, that conveys malevolent intentions toward a group or an individual, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDPC - Hate Speech Base + Extensive Definitions of Problematic Content
Hate speech is considered any kind of content that conveys malevolent intentions such as statements of inferiority, aversion, cursing, calls for exclusion, threats, harassment, or violence, toward a group or an individual, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDT - Hate Speech Base + Extensive Definitions of Target
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual who is, or is thought to be, a member of that group, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDFoC_EDT - Hate Speech Base + Extensive Definitions of: Form of Communication and Target
Hate speech is considered any kind of content or communication expressed using language (written or spoken) or actions, that conveys malevolent intentions toward a group or an individual who is, or is thought to be, a member of that group, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDFoC_EDPC - Hate Speech Base + Extensive Definitions of: Form of Communication and Problematic Content
Hate speech is considered any kind of content or communication expressed using language (written or spoken) or actions, that conveys malevolent intentions such as statements of inferiority, aversion, cursing, calls for exclusion, threats, harassment, or violence, toward a group or an individual, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDT_EDPC - Hate Speech Base + Extensive Definitions of: Problematic Content and Target
Hate speech is considered any kind of content that conveys malevolent intentions such as statements of inferiority, aversion, cursing, calls for exclusion, threats, harassment, or violence, toward a group or an individual who is, or is thought to be, a member of that group, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
HSB_EDFoC_EDPC_EDT - Hate Speech Base + Extensive Definitions of: Form of Communication, Problematic Content, and Target

Continued on next page

(Continued) Definitions of Offensive and Hate Speech
Hate speech is considered any kind of content or communication expressed using language (written or spoken) or actions, that conveys malevolent intentions such as statements of inferiority, aversion, cursing, calls for exclusion, threats, harassment, or violence, toward a group or an individual who is, or is thought to be, a member of that group, and is motivated by inherent characteristics that are attributed to that group and shared among its members.
For the second step, we take the HSB (Hate Speech Base) definition as exemplification and we add the combinations of the Accessory Elements.
HSB_LAA - Hate Speech Base + List of Addressed Attributes
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics.
HSB_LAA_PI - Hate Speech Base + List of Addressed Attributes + Possible Implications
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. The outcome of Hate Speech could be the promotion of division among people, undermining of social cohesion in communities, inciting others to commit violence or discrimination, and could have consequences for individuals' health and safety.
HSB_LAA_Exc - Hate Speech Base + List of Addressed Attributes + Exception
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. However, even if it is offensive, it is not considered Hate Speech any content that attacks a person's personality traits, ideas, or opinions.
HSB_LAA_IHS - Hate Speech Base + List of Addressed Attributes + Implicit Hate Speech
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. Hate Speech can also be implicit, portrayed as an indirect or coded language that uses Irony, Stereotypes, or Misinformation.
HSB_LAA_PI_Exc -Hate Speech Base + List of Addressed Attributes + Possible Implication + Exceptions
Continued on next page

(Continued) Definitions of Offensive and Hate Speech
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. The outcome of Hate Speech could be the promotion of division among people, undermining of social cohesion in communities, inciting others to commit violence or discrimination, and could have consequences for individuals' health and safety. However, even if it is offensive, it is not considered Hate Speech any content that attacks a person's personality traits, ideas, or opinions.
HSB_LAA_Exc_IHS - Hate Speech Base + List of Addressed Attributes + + Exceptions + Implicit Hate Speech
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. Hate Speech can also be implicit, portrayed as an indirect or coded language that uses Irony, Stereotypes, or Misinformation. However, even if it is offensive, it is not considered Hate Speech any content that attacks a person's personality traits, ideas, or opinions.
HSB_LAA_PI_IHS - Hate Speech Base + List of Addressed Attributes + Possible Implications + Implicit Hate Speech
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. The outcome of Hate Speech could be the promotion of division among people, undermining of social cohesion in communities, inciting others to commit violence or discrimination, and could have consequences for individuals' health and safety. Hate Speech can also be implicit, portrayed as an indirect or coded language that uses Irony, Stereotypes, or Misinformation.
HSB_LAA_PI_IHS_Exc - Hate Speech Base + List of Addressed Attributes + Possible Implications + Implicit Hate Speech + Exception
Hate speech is considered any kind of content that conveys malevolent intentions toward a group or an individual, and motivated by inherent characteristics that are attributed to that group and shared among its members such as race, color, ethnicity, gender, sexual orientation, nationality, religion, disability, social status, health conditions, or other characteristics. The outcome of Hate Speech could be the promotion of division among people, undermining of social cohesion in communities, inciting others to commit violence or discrimination, and could have consequences for individuals' health and safety. Hate Speech can also be implicit, portrayed as an indirect or coded language that uses Irony, Stereotypes, or Misinformation. However, even if it is offensive, it is not considered Hate Speech any content that attacks a person's personality traits, ideas, or opinions.

Table 6: Collection of definition prompted

F Prompt Templates

Without definition	With definition
Classify if the following text is considered hate speech or not hate speech:\n [TEXT]\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]\n Please answer ONLY with 1 if it is hate speech or 0 if it is not hate speech.*

Table 7: Prompts used for the classification task

G Robustness

We measure robustness by checking how many times the models answer in the same way under the three runs. All the results are reported below in Table 8.

In general, we observe a high consistency which stays relatively stable across models and datasets. Given the overall similarity between the scores, we identified outliers using the Interquartile Range (IQR), we observe that on the first step, 6 out of 9 times the outliers are the scores obtained in the condition without definition (NO). While on the second step, the definitions which contain the CE of exception generally lead to less robustness. Finally, we do not observe particular trends related to the highest value obtained in either of the steps.

Definitions	HateCheck			Learning from the Worst			Measuring Hate Speech		
	LLama3	Mistral	FlanT5	LLama3	Mistral	FlanT5	LLama3	Mistral	FlanT5
NO	91.28%	96.00%	70.34%	82.26%	93.92%	57.88%	88.75%	95.87%	72.78%
Own	94.03%	96.46%	75.44%	86.77%	95.95%	65.34%	89.64%	96.15%	76.95%
OL	94.44%	**96.87%**	75.11%	87.85%	94.59%	64.24%	85.82%	94.69%	75.85%
HSB	93.95%	96.51%	74.67%	88.29%	95.31%	65.65%	88.57%	96.46%	76.31%
HSB_EDFoC	95.08%	96.49%	75.67%	88.29%	95.33%	66.91%	90.11%	96.39%	77.26%
HSB_EDPC	94.77%	96.41%	74.06%	87.46%	**95.98%**	65.62%	89.11%	**96.69%**	77.39%
HSB_EDT	94.26%	96.56%	**76.83%**	88.23%	95.67%	67.03%	89.39%	96.23%	76.72%
HSB_EDFoC_EDT	94.23%	95.69%	76.06%	**88.75%**	95.57%	65.24%	90.16%	95.85%	77.77%
HSB_EDFoC_EDPC	95.21%	96.18%	74.44%	87.23%	95.49%	66.50%	88.95%	96.44%	**78.19%**
HSB_EDT_EDPC	94.77%	95.95%	75.85%	87.26%	95.85%	**67.55%**	89.03%	96.54%	76.16%
HSB_EDFoC_EDPC_EDT	**95.36%**	96.33%	76.65%	87.23%	95.72%	67.50%	**90.59%**	96.23%	78.11%
Avg. Step 1	**94.31%**	96.31%	**75.01%**	**87.23%**	95.39%	65.40%	89.10%	96.14%	**76.68%**
+LAA	94.23%	96.41%	**76.19%**	87.64%	96.08%	66.21%	90.41%	96.64%	**76.85%**
+LAA_PI	94.46%	96.82%	74.47%	88.21%	**96.28%**	**67.91%**	92.26%	**97.36%**	76.29%
+LAA_Exc	92.03%	96.62%	74.31%	82.62%	95.39%	64.47%	86.29%	96.44%	74.26%
+LAA_IHS	95.16%	97.23%	73.93%	88.13%	95.39%	65.42%	90.75%	96.69%	74.96%
+LAA_PI_Exc	92.54%	97.03%	74.72%	84.54%	96.15%	66.78%	88.39%	96.23%	76.72%
+LAA_Exc_IHS	94.05%	97.15%	74.31%	87.16%	95.44%	67.14%	90.39%	96.46%	74.08%
+LAA_PI_IHS	**95.49%**	**97.46%**	74.37%	**88.31%**	95.59%	65.34%	**92.44%**	97.15%	75.70%
+LAA_PI_IHS_Exc	94.41%	96.80%	75.06%	87.11%	96.23%	66.42%	90.57%	96.56%	75.19%
Avg. Step 2	94.04%	**96.94%**	74.67%	86.71%	**95.81%**	**66.21%**	**90.18%**	**96.69%**	75.50%

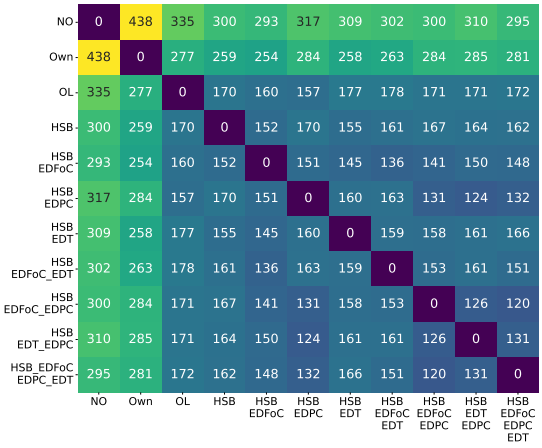
Table 8: Scores in consistency within the same definition. In **bold** the highest values observed (per step), underlined the outliers identified with the Interquartile Range method.

H Sensitivity

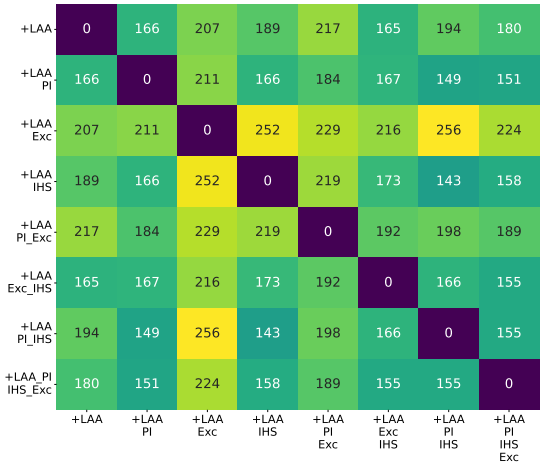
In this analysis, instead of comparing the results produced by each run, we are comparing how the answers change definition by definition, in other words, how sensitive is the model to different definitions.

We represent this trough confusion matrices reflecting the average non-consistent answers between each definition. Through this sensitivity analysis we observe that generally all the models tend to be less and less sensitive as more information is added to the definition.

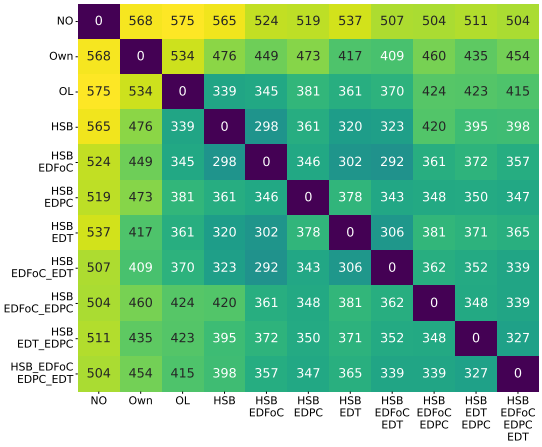
The same does not apply to the second step, when more specific information are added to the definition (i.e., notion of implicit HS, or possible implications), we instead observe more sensitivity when we are comparing definitions with different Conceptual Elements (i.e., definition with CE of implicit HS vs. definition with CE of Exception), and vice versa when the these CEs are shared by the compared definitions. Especially, this results coherent with what we observe in Sec. 6, we observe more non-consistent answer when we are comparing definitions with different CEs. For instance, when we are comparing the definition with the CE of exception and the definition with the CE of implicit HS, we observe an higher number different responses, hinting that the model is classifying data-points in a different way, exactly how we saw in our error analysis. Even though in Sec. 6 we could test it only for HateCheck, we observe the same non-consistent pattern in the second step across all three the datasets.



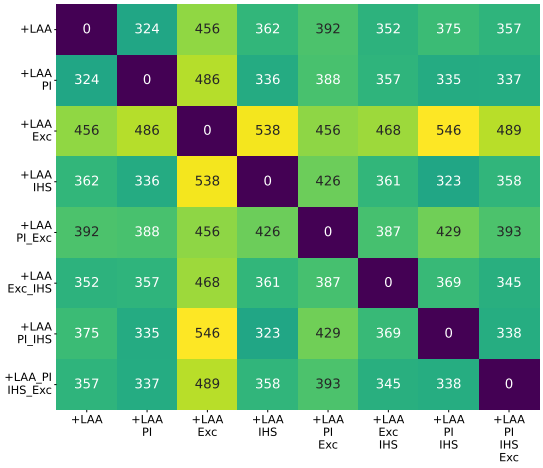
(a) HateCheck Dataset Step 1



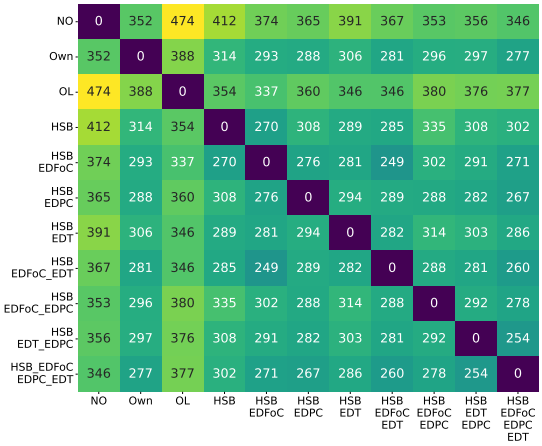
(b) HateCheck Dataset Step 2



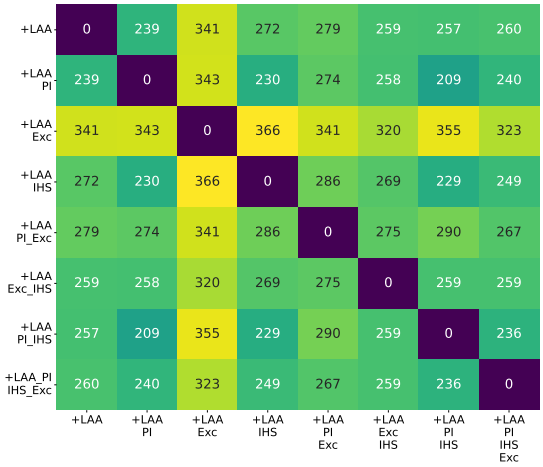
(c) Learning from the Worst Dataset Step 1



(d) Learning from the Worst Dataset Step 2

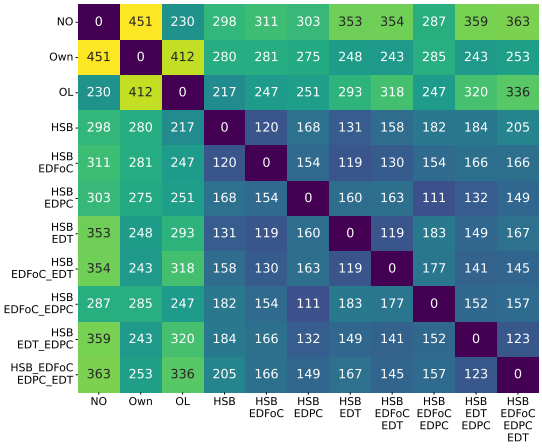


(e) Measuring Hate Speech Dataset Step 1

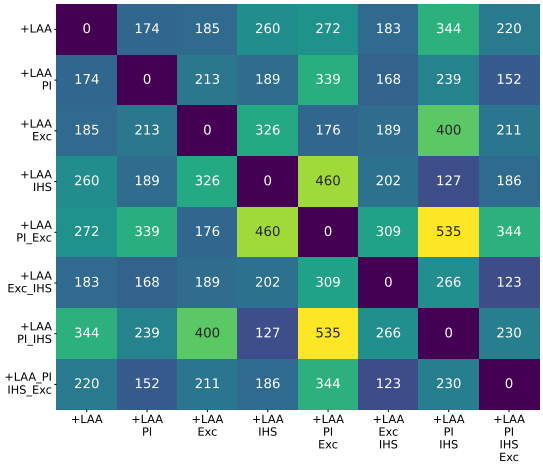


(f) Measuring Hate Speech Dataset Step 2

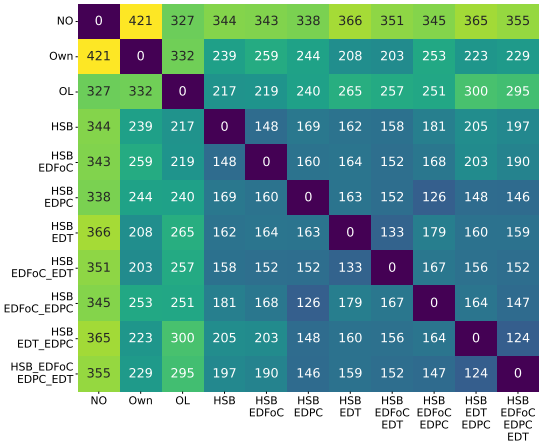
Figure 5: Confusion matrices of non-consistent answer between definitions in **LLama3**



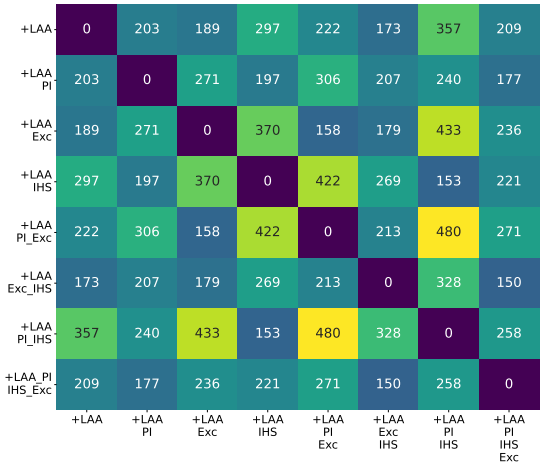
(a) HateCheck Dataset Step 1



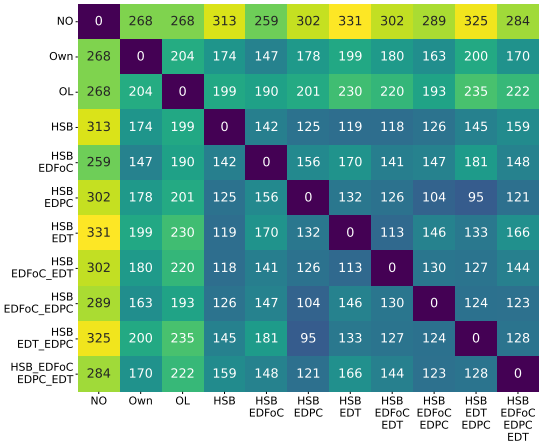
(b) HateCheck Dataset Step 2



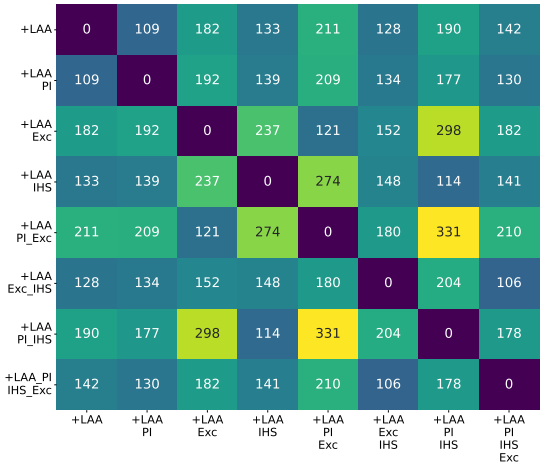
(c) Learning from the Worst Dataset Step 1



(d) Learning from the Worst Dataset Step 2

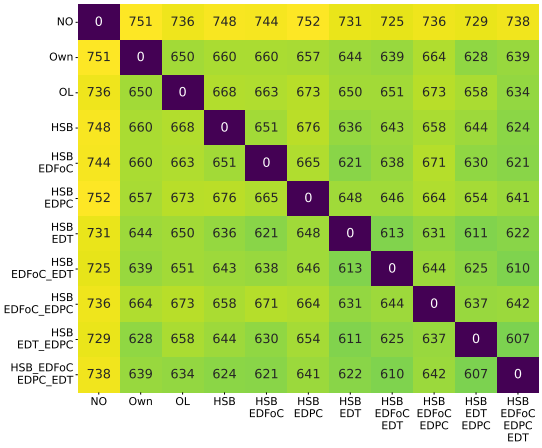


(e) Measuring Hate Speech Dataset Step 1

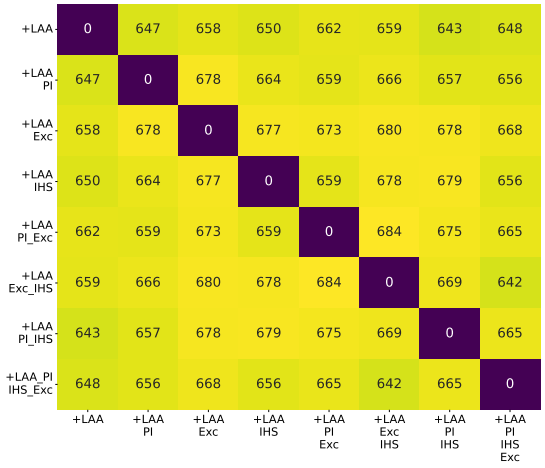


(f) Measuring Hate Speech Dataset Step 2

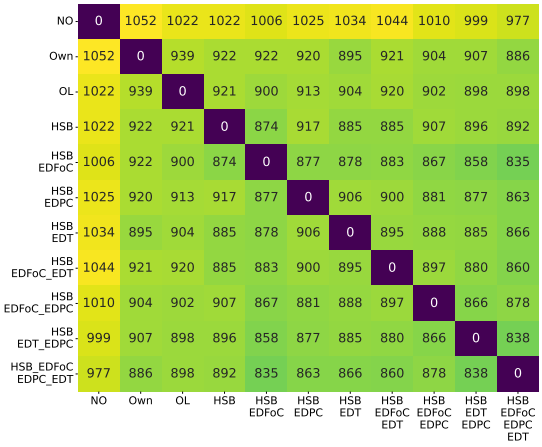
Figure 6: Confusion matrices of non-consistent answer between definitions in **Mistral**



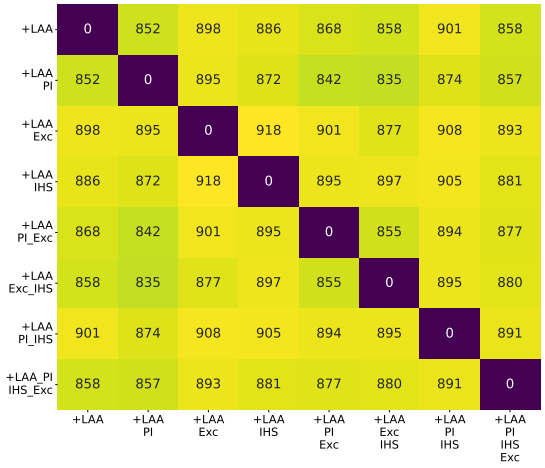
(a) HateCheck Dataset Step 1



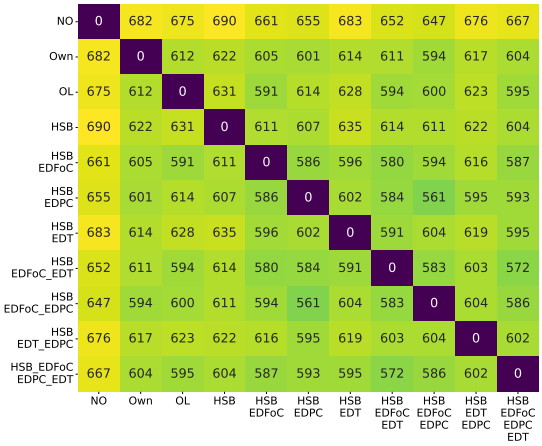
(b) HateCheck Dataset Step 2



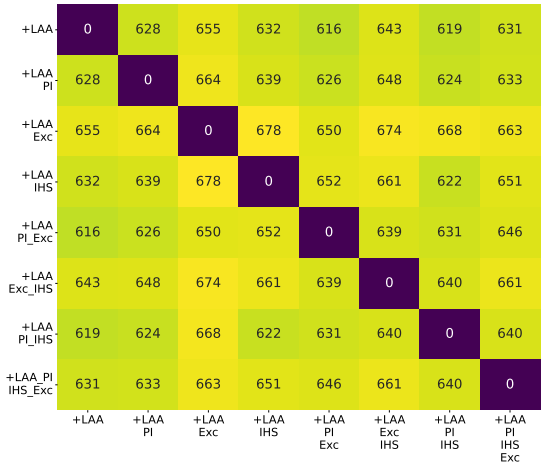
(c) Learning from the Worst Dataset Step 1



(d) Learning from the Worst Dataset Step 2



(e) Measuring Hate Speech Dataset Step 1



(f) Measuring Hate Speech Dataset Step 2

Figure 7: Confusion matrices of non-consistent answer between definitions in **Flan-T5**

I Error Distribution Based on Classes

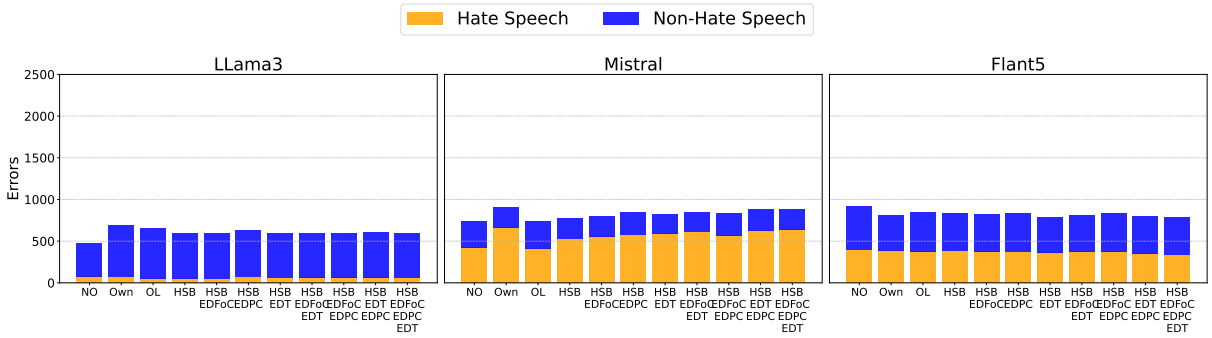


Figure 8: Distribution of errors across the three models on HateCheck (Step 1).

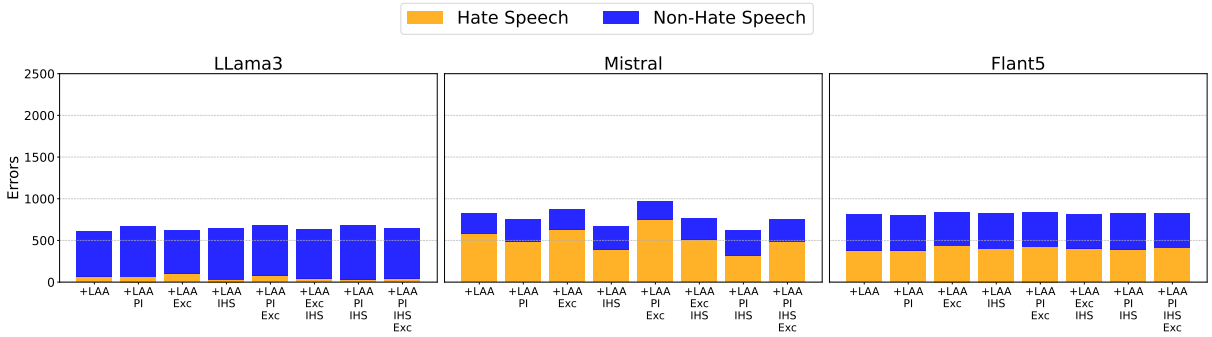


Figure 9: Distribution of errors across the three models on HateCheck (Step 2).

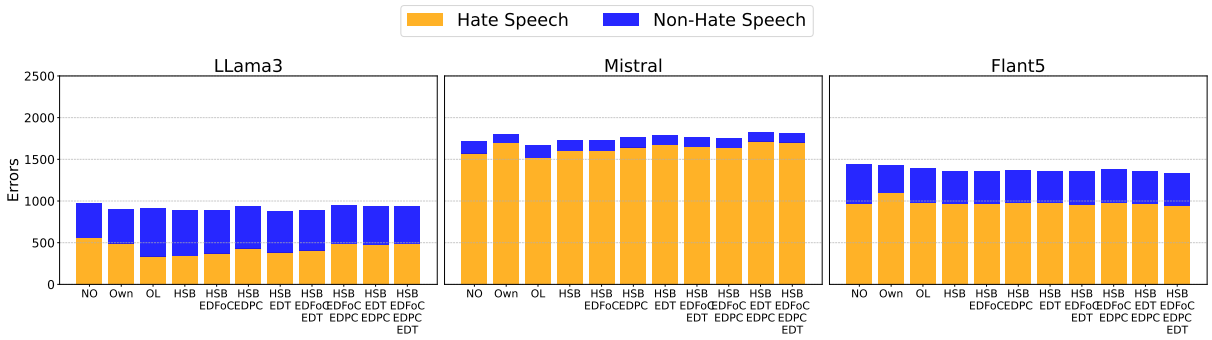


Figure 10: Distribution of errors across the three models on Learning from the Worst (Step 1).

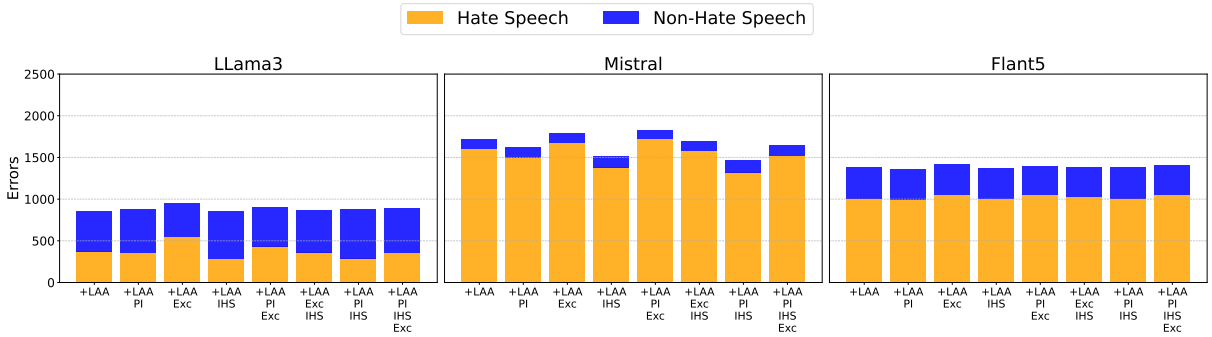


Figure 11: Distribution of errors across the three models on Learning from the Worst (Step 2).

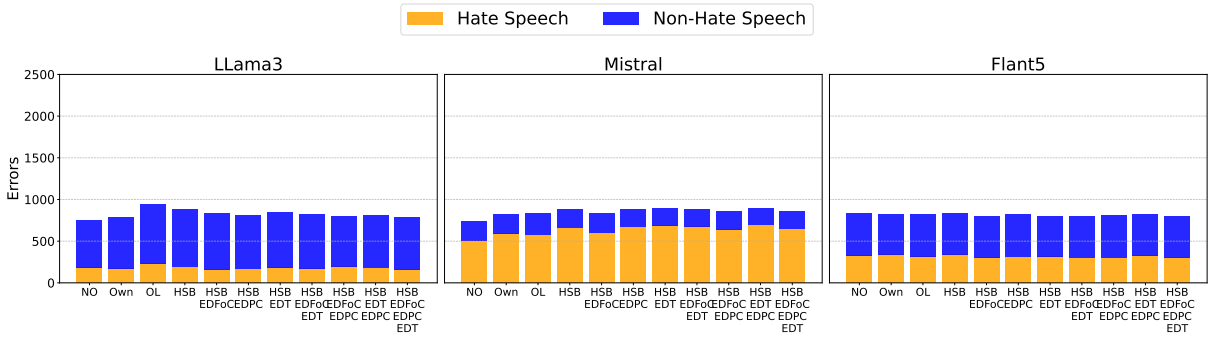


Figure 12: Distribution of errors across the three models on Measuring Hate Speech (Step 1).

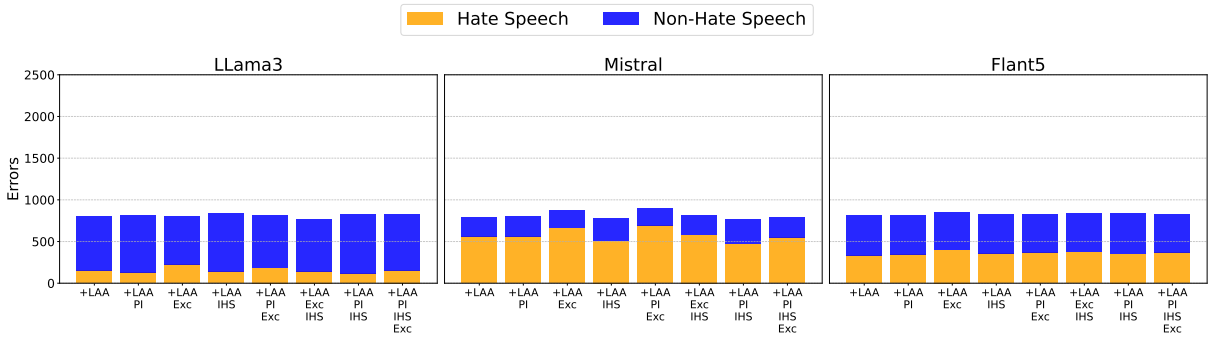


Figure 13: Distribution of errors across the three models on Measuring Hate Speech (Step 2).

J Grouping functionalities in Macro Classes

HateCheck dataset presents 29 functionalities (or type of hate) "motivated by a review of previous research and a series of interviews with civil society stakeholders" (Röttger et al., 2021b). In order to investigate trends after the inference phase, we decided to cluster these into 5 Macro Classes reflecting:

- Hate Speech: data-points that simply reflect direct and evident Hate Speech, such as: (slur, dehumanization, threats, etc...)
- Non Hate Speech: data-points that reflect general sentences without conveying hate (i.e. "I am a trans person", "No fucking way")
- Leet Hate Speech: data-points build to resemble *leetspeak*, so containing different spelling errors, which makes them more difficult to be detected, for instance: "Ihate trans people" or "I h4te immigrants".
- Misleading Not Hate Speech: data-points which do not convey HS but are built in a way which resemble an Hate Speech comment such as: "Saying that you hate women shows how misguided you are" or "She ought to be executed".
- Special Hate Speech: data-points which are more implicitly phrased such as: "Never trust a single word said by a woman" or "I really do believe black people are just lesser beings".

Macro-classes	Functionalities
Hate Speech (HS)	derog_dehum_h derog_neg_attrib_h threat_dir_h negate_pos_h derog_neg_emote_h slur_h ref_subs_sent_h threat_norm_h profanity_h
Non Hate Speech (NHS)	ident_pos_nh ident_neutral_nh profanity_nh target_obj_nh
Leet HS	space_deleted, leet_hate_speech, character_swap, space_add, character_deleted
Misleading NHS	counter_ref_nh negate_neg_nh counter_quote_nh target_indiv_nh target_group_nh slur_reclaimed_nh slur_homonym_nh
Special HS	derog_impl_h ref_subs_clause_h phrase_question_h phrase_opinion_h

Table 9: 29 functionalities (type of hate) grouped in 5 Macro Classes.

K HateCheck Errors - Functionalities

Below, we report the relative average model errors across all HateCheck functionalities.

LLama-3		Mistral		Flan-T5	
Functionality	Error	Functionality	Error	Functionality	Error
counter_quote_nh	93,75%	counter_quote_nh	76,63%	counter_quote_nh	84,81%
couter_ref_nh	80,47%	derog_impl_h	42,45%	couter_ref_nh	68,86%
slur_reclaimed_nh	70,71%	couter_ref_nh	42,13%	target_group_nh	54,22%
target_indiv_nh	62,05%	slur_h	38.97%	slur_reclaimed_nh	48,26%
target_group_nh	58,93%	target_indiv_nh	33,61%	target_indiv_nh	44,43%
negate_neg_nh	36,95%	spell_space_add_h	33,04%	slur_homonym_nh	37,27%
slur_homonym_nh	25,86%	spell_space_del_h	29,58%	derog_impl_h	28,90%
ident_pos_nh	14,66%	spell_leet_h	26,97%	profanity_nh	26,24%
ident_neutral_nh	8,03%	derog_neg_emote_h	25,37%	spell_space_add_h	24,82%
spell_space_add_h	7,78%	phrase_question_h	25,22%	spell_leet_h	22,18&
spell_leet_h	5,90%	profanity_h	24,00%	negate_neg_nh	21,95%
profanity_nh	5,73%	spell_char_del_h	23,98%	derog_neg_emote_h	19,91%
spell_char_del_h	5,52%	spell_char_swap_h	22,45%	spell_char_del_h	18,68%
slur_h	5,18%	derog_neg_attrib_h	22,12%	spell_char_swap_h	18,29%
derog_impl_h	4,59%	target_group_nh	19,95%	slur_h	18,03%
target_obj_nh	3,64%	ref_subs_sent_h	15,84%	negate_pos_h	16,80%
spell_space_del_h	3,43%	ref_subs_clause_h	14,57%	spell_space_del_h	16,34&
derog_neg_emote_h	2,27%	negate_pos_h	8,85%	threat_norm_h	9,37%
threat_norm_h	2,21%	phrase_opinion_h	8,40%	target_obj_nh	8,90%
threat_dir_h	0,35%	negate_neg_nh	8,31%	phrased_question_h	8,85%
phrase_question_h	0,22%	slur_reclaimed_nh	5,27%	ref_subs_sent_h	8,25%
derog_neg_attrib_h	0,19%	slur_homonym_sh	4,14%	ref_subs_clause_h	8,22%
negate_pos_h	0,06%	ident_pos_nh	3,43%	profanity_h	7,23%
spell_char_swap_h	0,06%	derog_dehum_h	3,35%	derog_neg_attrib_h	7,10%
phrase_opinion_h	0,06%	threat_dir_h	1,95%	threat_dir_h	5,32%
profanity_h	0,02%	threat_norm_h	1,13%	ident_pos_nh	3,94%
derog_dehum_h	-	profanity_nh	0,51%	derog_dehum_h	3,92%
ref_subs_clause_h	-	target_obj_nh	0,47%	ident_neutral_nh	3,42%
ref_subs_sent_h	-	ident_neutral_nh	0,32%	phrase_opinion_h	2,94%

Table 10: Average error per functionality across definition per models. We colour coded each functionality based on the Macro Class in which it belongs: Hate Speech (HS), Non Hate Speech (NHS), Leet HS, Misleading NHS, Special HS

L Graph of Errors by Macro Classes

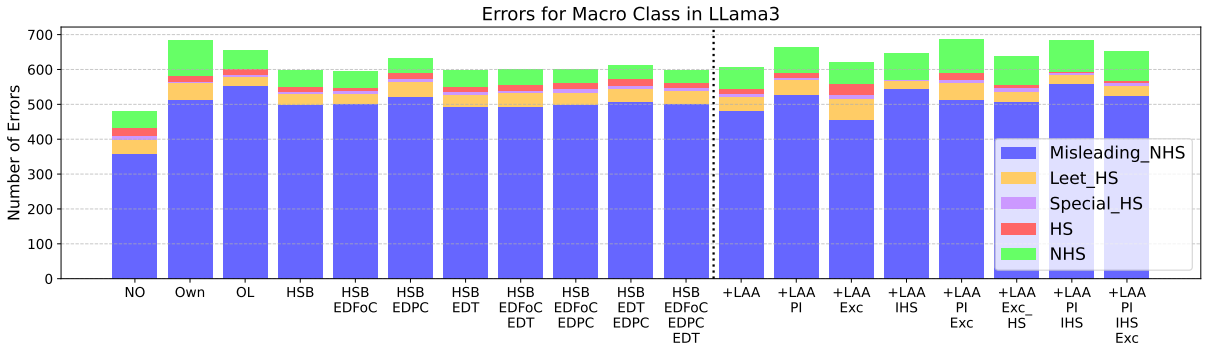


Figure 14: Distribution of errors across Macro Classes

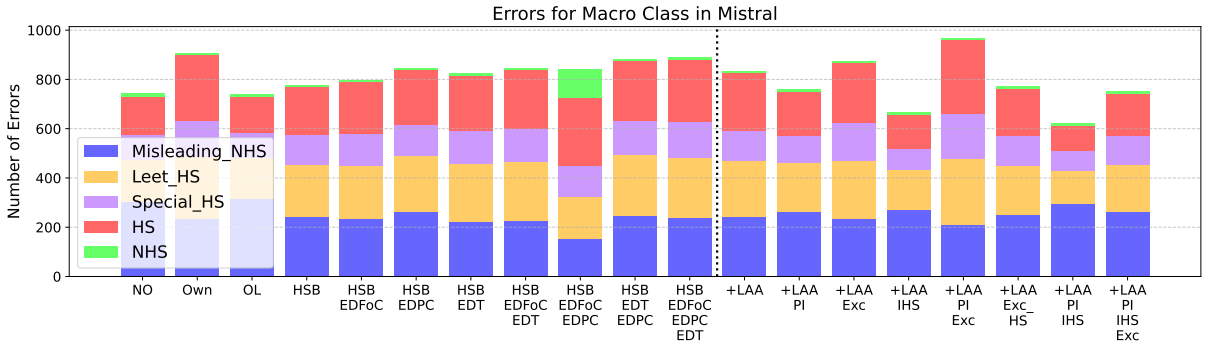


Figure 15: Distribution of errors across Macro Classes.

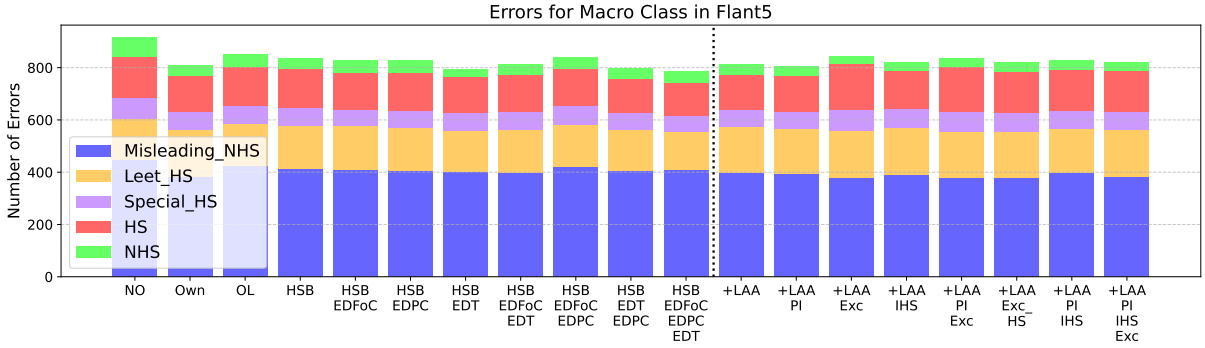


Figure 16: Distribution of errors across Macro Classes.

M Impact of Conceptual Elements on functionality Macro Classes

LLama-3	No Def	HSB_EDT	+LAA_IHS	+LAA_Exc	+LAA_Exc_IHS
Misleading NHS	**47,59%**	60,00%	67,18%	*53,92%*	60,28%
Leet HS	5,07%	4,18%	**2,78%**	5,81%	3,66%
Special HS	1,73%	1,61%	**0,36%**	1,61%	1,61%

Table 11: Error percentage, Conceptual Eelements & Macro Classes LLama-3. In **bold** the best result per Macro Class, in *italic* the best result considering only the second step.

Mistral	No Def	HSB	+LAA_IHS	+LAA_Exc	+LAA_Exc_IHS
Misleading NHS	34.81%	25.64%	28.36%	**23.26%**	26.04%
Leet HS	20.81%	26.17%	**19.76%**	29.23%	24.24%
Special HS	18.75%	21.31%	**15.71%**	27.14%	21.84%

Table 12: Error percentage, Conceptual Elements & Macro Classes Mistral. In **bold** the best result per Macro Class, in *italic* the best result considering only the second step.

Flan-T5	No Def	HSB_EDT	+LAA_IHS	+LAA_Exc	+LAA_Exc_IHS
Misleading NHS	58,24%	49,03%	49,36%	**47,03%**	48,46%
Leet HS	**19,51%**	19,58%	22,14%	22,12%	21,55%
Special HS	14,11%	**11,84%**	*12,98%*	14,64%	13,51%

Table 13: Error percentage, Conceptual Elements & Macro Classes T5. In **bold** the best result per Macro Class, in *italic* the best result considering only the second step.