Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083252

ARTICLE

From Binary to Multi-Class: LLM-Judged Synthetic Annotation Applied to Hate Speech Detection

Antonio Moreno-Cediel, Antonio Garcia-Cabot and Eva Garcia-Lopez*

Departamento de Ciencias de la Computación, Universidad de Alcalá, Alcalá de Henares, Madrid, Spain

*Corresponding Author: Eva Garcia-Lopez. Email: eva.garcial@uah.es

Received: 31 March 2026; Accepted: 11 June 2026; Published: 13 August 2026

ABSTRACT: The increasing prevalence of hate speech on social media platforms has spurred research aimed at mitigating this societal harm. However, the development of effective machine learning solutions is hindered by a lack of labelled hate speech data in languages beyond English, particularly when attempting granular, multi-class classification. This research aims to address this data scarcity by introducing a novel methodology leveraging the 'Large Language Model as a judge' paradigm to transform existing binary-labelled hate speech data into multi-class datasets. Our approach aims to generate balanced datasets and enables classification across seven identity groups: race, religion, origin, gender, sexuality, age, and disability. The methodology has been applied to the Spanish Hate Speech Superset, and it has been validated using the Measuring Hate Speech dataset, demonstrating significant efficacy and broad applicability. Specifically, our approach obtains a higher match rate with human labels and a lower number of mismatches when compared with prompt-only strategies. In addition, Cohen's Kappa scores demonstrate that our approach exhibits a moderate strength of agreement with human annotators, outperforming prompt-only strategies' scores by 5%. As a result of the application of the proposed strategy to the Spanish Hate Speech Superset dataset, a multi-class version is obtained, comprising 6325 hateful samples classified among the seven identity groups. The proposed strategy offers a versatile solution for nuanced classification tasks beyond hate speech, providing a valuable technique for detailed categorisation in various domains.

KEYWORDS: Hate speech; synthetic data annotation; dataset; LLM-as-a-judge; multi-class classification

1 Introduction

The term 'hate speech' was first coined by a group of legal scholars in the late 1980s. Specifically, Matsuda applied this term to refer to trials where the law does not adequately protect victims of racist hate speech [1]. Since then, the term 'hate speech' has been taken up in both legal and ordinary contexts [2]. However, it is widely acknowledged that there is no universal definition of hate speech, and that its boundaries with other terms, such as 'hate crime', are not clearly defined.

International organisations such as the Council of Europe [3], the European Union Council [4] and the United Nations [5] have already proposed some definitions with the purpose of establishing a unified framework to address this societal scourge. In summary, these definitions align with the understanding of hate speech as a form of communication that incites or promotes hatred against a group of individuals based on identity-related factors. Therefore, with the objective of promoting this unified framework, this conceptualisation will serve as the working definition adopted in the present study.

The increasing prevalence of hate speech in contemporary society is closely linked to the expansion of social media platforms. The advent of social media has had both beneficial and detrimental effects on society. On the one hand, social media has improved overall communication, and is currently used as a communication medium in addition to its inherent social functions [6]. On the other hand, this progress has brought about an increase in the propagation of hate speech [7–10]. In fact, the influence of social media is now so prominent that several authors consider the internet to be the main source of hate speech [11].

However, the adverse effects of hate speech on individuals' psychological well-being, the erosion of social relationships, and the incitement of real-world violence are among the many consequences that extend beyond the digital sphere [8,12,13]. Hence, the detection and classification of hate speech within social networks are essential to curb its dissemination. Given the impracticality of manually processing the large volumes of data generated by social media, computer science techniques are being applied to address this challenging global issue [14].

In recent years, a variety of approaches have been employed to address the task of automatic hate speech classification. Despite methodological differences, spanning from traditional machine learning models to recent Transformer-based architectures [15], these approaches share a fundamental reliance on the availability of large, diverse, and representative datasets to enable effective training. However, such datasets remain scarce, a limitation that becomes even more pronounced when the classification task is framed as a multi-class problem and applied to languages other than English. Consequently, there is a clear need for a technique that automatically converts existing binary datasets into multi-class ones.

Due to the high time and economic costs of human annotation, there is a pervasive consensus in favour of the adoption of LLMs for the automated annotation of data within the domain of computational social science [16]. Nevertheless, there is no consensus regarding the potential impact of employing these models for synthetic annotation on scientific integrity. While some studies support the deployment of LLMs in a broad spectrum of text annotation applications [17,18], others question the reliability of the use of LLMs for annotation purposes [16,19]. As a result, this uncertainty underscores the necessity for additional research in the domain of automated text annotation.

The use of Large Language Models (LLMs) as automated evaluators, often referred to as judges, has emerged as a result of the remarkable performance achieved by LLMs in instruction following, query understanding, and response generation [20]. The efficacy of using LLMs as judges has been demonstrated in a wide range of fields, such as ranking [21], mathematical reasoning, or medicine. However, to the best of our knowledge, the LLM-as-a-judge paradigm has not yet been proven effective when handling hate speech synthetic data annotation tasks.

Objective and contribution. In this paper, we aim to develop an effective methodology for automatically converting binary hate speech datasets into multi-class ones. The proposed approach leverages the use of LLMs as judges to improve annotation reliability and inter-annotator agreement with human labels. To this end, we seek answers to the following research questions.

- **RQ1:** Can the LLM-as-a-judge paradigm be effectively integrated into a synthetic annotation architecture to improve existing state-of-the-art synthetic annotation techniques?
- **RQ2:** To what extent does the proposed LLM-based judging methodology align with human annotations?
- **RQ3**: Which label-reduction strategy provides the optimal balance between annotation quality and class distribution fairness when converting multi-label datasets to multi-class?
- **RQ4:** Which combinations of LLM architectures and scales optimize the synergy between the classifier and the judge to achieve the highest annotation quality?

- **RQ5:** To what extent does the proposed methodology maintain consistent performance when applied across English and Spanish datasets?

The contributions of this paper to the domain of hate speech and synthetic data annotation are as follows:

- The proposal of a novel synthetic annotation methodology leveraging an LLM-as-a-judge approach. This approach enables the automatic translation of binary-labelled datasets into multi-class ones, while also providing an intermediate multi-label dataset.
- The conversion of the Spanish Hate Speech Superset binary dataset into a multi-class hate speech dataset useful for target-group detection in hate speech. Hateful samples have been annotated according to the following target identity groups: race, religion, national origin or citizenship status, gender, sexual orientation, age, and disability status.
- An empirical analysis of the performance of existing LLMs in synthetic labelling. This analysis reveals that model scale does not correlate linearly with annotation quality; instead, it highlights a critical distinction between 'classification power' and 'judging ability', thereby identifying the best classifiers and judges.
- The validation of the methodology's cross-lingual robustness, demonstrating that the proposed LLM-as-a-judge paradigm maintains consistent fidelity with human annotators across both English and Spanish datasets.

The remainder of this paper is organised as follows: Section 2 reviews related work in text and hate speech classification, including relevant datasets and synthetic data annotation techniques. Section 3 details the methodology employed in this research. Section 4 presents the results obtained. Finally, Sections 5 and 6 present the discussion and conclusions of the research, respectively.

2 Related Work

The present study focuses on translating binary hate speech datasets into multi-class ones by means of synthetic annotation using LLMs. Therefore, this section presents a thorough review of the state-of-the-art regarding automated text classification, including automated hate speech classification and synthetic annotation with LLMs.

2.1 Automated Text Classification

Text classification is a classical problem in Natural Language Processing (NLP) that aims to assign labels to textual units, such as sentences, queries, paragraphs, and documents [22], and it is a field of significant applicability in areas such as sentiment analysis, topic labelling, and question answering [23]. Depending on the label nature of the labels, classification tasks are categorized as binary (two mutually exclusive categories), multi-class (more than two mutually exclusive classes), or multi-label (multiple non-exclusive categories) [24].

Over the years, text classification has evolved from early rule-based methods [25] to statistical and machine learning techniques, such as Naïve Bayes [26], Support Vector Machines [27], and K-Nearest Neighbours [28]. Subsequently, deep learning introduced architectures such as Recurrent Neural Networks (RNNs) [29,30] and Convolutional Neural Networks (CNNs) [31] to handle sequential and semantic patterns. The field was further revolutionised by the Transformer architecture [15], specifically encoder-based models such as the Bidirectional Encoder Representations from Transformers (BERT) [32–34] which significantly improved performance [23]. Currently, the field is shifting towards the use of LLMs for zero- or few-shot classification [35,36].

Automated Hate Speech Classification

Narrowing the scope to hate speech textual classification, research activity in this field has experienced a noticeable growth in recent years [10,37]. Recent work on automated hate speech classification consistently reports the superiority of Transformer-based models over traditional machine learning and earlier deep neural architectures. Comparative studies, such as that by Malik et al. [38], show that Transformers (e.g., BERT [32], ELECTRA [39], ALBERT [40]) outperform shallow classifiers and recurrent models across multiple binary benchmarks. Similarly, Abusaqer et al. [41] evaluate 38 models ranging from traditional machine learning classifiers to Transformers, concluding that architectures such as RoBERTa [42] and BERT consistently achieve the highest performance.

These findings extend to low-resource and multilingual settings. Fetahi et al. [43] demonstrate that a fine-tuned XLM-RoBERTa [44] surpasses classical machine learning and standard deep learning approaches for Albanian hate speech detection, while More et al. [45] report that a fine-tuned BERT model outperforms baselines in political hate speech classification. In the comparatively scarce multi-label scenario, Angger Saputra and Sibaroni [46] demonstrate that hybrid BERT-based architectures (e.g., BERT-CNN) improve performance over standalone BERT.

Using the binary MetaHate dataset, Chapagain et al. [47] evaluate ELECTRA against traditional baselines (linear SVM with TF-IDF and a CNN) and multiple Transformer variants. ELECTRA achieves the highest performance, with lighter models such as DistilBERT remaining competitive. The authors also highlight potential label noise in the dataset and advocate for semi-automated relabelling with explainable AI, as well as hybrid fine-tuning strategies for low-resource and cross-platform scenarios.

Finally, Albladi et al. [48] synthesise recent advancements in LLM-based hate speech detection. They emphasise that fine-tuned LLMs such as BERT, GPT-3 [49], and RoBERTa consistently outperform traditional machine learning methods, particularly in identifying implicit, sarcastic, or context-dependent hate speech that often escapes keyword-based systems.

Overall, the literature converges on deep learning, and especially LLM-based methods, as the state of the art in hate speech detection. However, these strategies typically require substantial amounts of annotated data for effective fine-tuning. In practice, this requirement directly interacts with a structural limitation of the field: the nature and availability of existing datasets.

In particular, the vast majority of hate speech datasets employed for model training are binarily annotated [50,51]. Consequently, research commonly addresses hate speech detection as a binary classification problem, identifying whether a text is considered to be hateful or not [13,52]. Regarding multi-class and multi-label datasets, their availability tends to decrease as the number of class labels increases. Datasets containing only three classes are the most prevalent [53,54], whereas those with a greater number of labels are significantly more difficult to obtain [55].

From a methodological perspective, multi-label classification is encouraged in the literature because, albeit more complex, it yields more nuanced outcomes. However, as is also the case with the definition of hate speech, there is no globally adopted categorisation of hate speech subtypes, and special attention is needed to establish a balance between performance and granularity [56,57].

Data-related constraints further exacerbate these issues. Rarely do these datasets contain a substantial volume of samples, and they often suffer from class imbalance and a low proportion of offensive content. Some studies attempt to mitigate these limitations by combining multiple datasets [58], although such strategies are generally only feasible for high-resource languages, such as English. When working with other languages, the scarcity of annotated data becomes even more pronounced [52]. Additionally, dataset

aggregation in multi-class or multi-label settings requires compatible label schemes, which is often difficult due to the lack of a standardised hate speech categorisation.

2.2 Synthetic Annotation with Large Language Models

High-quality, annotated data is paramount to ensure the effective training of deep learning models. Traditionally, human annotators have undertaken the task of manually annotating data, shaping datasets for supervised learning. However, this time-consuming process entails exorbitant costs [59], thereby driving the research community to further explore alternative annotation strategies [60]. The remarkable capabilities achieved by LLMs over the last few years, which approximate human-level performance [35], have shed light on the use of these models as a tool for synthetically annotating large data corpora across the computational social sciences [16].

Automatic labelling using LLMs has become an emerging area of research. Kazemi et al. [61] systematically evaluate the role of LLM-generated data and labels in cyberbullying detection under four distinct scenarios. Of particular interest is their Scenario 4, in which LLMs generate synthetic labels for unlabelled authentic data, a setting closely aligned with our work. Their results show that a BERT-based classifier trained on synthetic labels achieves 79.1% test accuracy, compared to 81.5% when trained on fully human-labelled data, leaving a 2.4% performance gap. This small difference demonstrates the practical viability of synthetic annotation as a substitute for manual labelling.

However, an important methodological distinction must be made. Kazemi et al. [61] evaluate annotation quality indirectly, by measuring downstream classifier performance. While this approach captures the utility of synthetic labels for model training, it may partially obscure the intrinsic quality of the annotations themselves, as final performance depends not only on label correctness but also on the generalization capacity and biases of the classifier. Nevertheless, the persistent 2.4% gap observed in Scenario 4 underscores that synthetic labelling remains imperfect and highlights the need for improved annotation strategies.

Furthermore, several studies conceptualise LLMs as substitutes for human annotators through direct prompting. Horych et al. [60] demonstrate that LLM-annotated datasets for media bias detection can train downstream classifiers that perform on par with models trained on human-labelled data. Furthermore, Parfenova et al. [19] report competitive performance of LLMs in inductive coding, particularly for simpler textual instances, although humans retain an advantage in complex, interpretative cases.

A complementary line of research focuses on label error detection and quality control. Nahum et al. [62] use ensembles of LLMs via prompting to flag high-confidence disagreements with existing labels, identifying 6%–21% of mislabelled instances and improving downstream performance after correction. Their findings further indicate that LLM agreement can approach expert-level consistency, while surpassing crowd workers in efficiency.

Despite broad enthusiasm for LLM-based annotation, concerns remain regarding robustness and scientific validity. Baumann et al. [16] demonstrate that LLM-generated annotations can lead to incorrect downstream statistical conclusions in a substantial proportion of cases, with performance being highly sensitive to prompt formulation. Moreover, LLM reliability has been assessed in the hate speech detection domain. While LLMs' annotation reliability remains below human-levels, when used as evaluators rather than annotators, LLMs generate more reliable outputs [63]. These findings highlight the unpredictability of purely prompt-based pipelines, underscoring the need for principled safeguards while also revealing the opportunity to exploit improved reliability when LLMs are used as evaluators.

In summary, existing research establishes that LLMs can serve as scalable and cost-effective annotators, with performance approaching human benchmarks. However, synthetic labelling still exhibits measurable

gaps relative to gold-standard annotations, and current approaches rely on zero or few-shot prompting. Recently, the LLM-as-a-judge approach [64] has gained popularity to meet the increasing demand for LLM evaluation, and it is considered useful for obtaining high-quality annotated data [65]. Consequently, our research moves beyond straightforward LLM annotations and instead leverages the observed reliability gains of LLMs in evaluator roles by adopting an iterative LLM-as-a-judge approach. This strategy ensures inter-model agreement and considered annotations, while preventing straightforward, non-grounded classifications.

3 Objectives

Data availability is crucial for the development of effective and reliable artificial intelligence systems. In the hate speech domain, a notable lack of multi-class datasets is hindering the advancement of multi-class hate speech classification systems. Synthetic data labelling with LLMs has emerged as a promising remedy to data scarcity issues. Nevertheless, concerns regarding the reliability of synthetic labelling techniques remain a matter of ongoing debate [16], meaning that the objective is to obtain synthetic labels that are as closely aligned as possible with those of human annotators [62]. Our study identifies a clear gap between the ongoing research efforts to enhance data labelling with LLMs and the data scarcity problem for multi-class hate speech classification, highlighting the need to establish an improved, automatic, multi-class labelling methodology. As a result, this research is guided by four main objectives:

- To propose a novel synthetic annotation methodology based on the LLM-as-a-judge paradigm.
- To apply this strategy to effectively convert binary-labelled hate speech datasets into multi-class labelled datasets with target-group labels.
- To demonstrate that the proposed strategy can improve upon synthetic classification approaches based on prompting a single LLM.
- To analyse the level of agreement between labels obtained using the proposed strategy and those provided by human annotators.

4 Methodology

The present section illustrates the proposed methodology for automatically labelling binary hate speech datasets using the LLM-as-a-judge approach for multi-class classification purposes. It describes the dataset employed, defines the different classes used for hate speech classification, specifies the architecture and implementation details, and explains the synthetic annotation evaluation framework.

4.1 Hate Speech Dataset

The dataset employed in this work is the Spanish Hate Speech Superset (SHSS) [66], a Spanish corpus aligned with the United Nations definition of hate speech. This dataset is composed of approximately thirty thousand posts obtained from the datasets proposed in the following studies: *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* [50], *Detecting and Monitoring Hate Speech in Twitter* [67], *Multilingual Resources for Offensive Language Detection* [68], *Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection* [69], *HOMO-MEX: A Mexican Spanish Annotated Corpus for LGBT+phobia Detection on twitter* [70], and *Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings* [71].

Since some of the datasets that constitute this superset do not contain binary labels, Tonneau et al. [66] pre-processed them to binarize the data into hateful and non-hateful samples, creating the dataset used in this study (SHSS). The SHSS dataset comprises a total of 29,855 posts, divided into 22,590 non-hateful and 7265 hateful posts. Table 1 includes three illustrative samples of the dataset.

Table 1: SHSS dataset sample. The left column represents post content, while the right column represents the hate speech binary label (1 denotes hateful and 0 denotes non-hateful). Offensive expressions have been masked using angle brackets to omit potentially sensitive or harmful content, ensuring neutrality and avoiding unintended ideological or political implications.

Text (In Spanish)	Labels
Eran tan pero tan feministas que invisibilizaban constantemente a las trabajadoras sexuales, haciéndole creer al mundo que eran incapaces de decider y que cada vez que ejercían su derecho a hacerlo estaban siendo violadas.	0
@USER @USER @USER Me carga en lo que se convirtió la <*political reference*>. . .una <*class-based insult*>.	0
¿Sabrán las <*an insult directed at feminists*> como @USER y todo el <*derogatory expression targeting Indigenous people and lower-income groups*>, que si hay una cultura y sociedad absolutamente hetero patriarcal, de un machismo extremo, es justamente la <*name of the society*>?	1

The geographical distribution shown in Fig. 1 highlights the inherent complexity of working with the Spanish language, which is an international language spoken by over 500 million native speakers across 21 countries [72]. As observed, the dataset is predominantly composed of posts from Chile and Spain, though it integrates various other Latin American variants. This diversity is crucial because Spanish exhibits a complex array of regional dialects: while Spain has several geographically delimited varieties [73], Latin American Spanish is characterised by a high variability driven by language contact, internal migrations, and rural-urban polarisation [73].

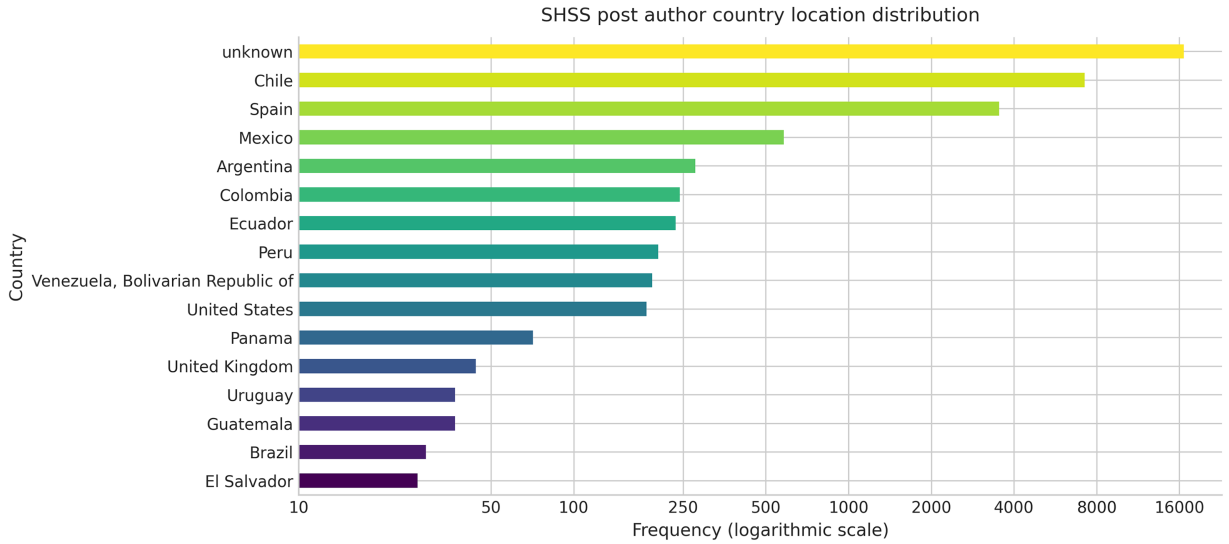

Figure 1: Distribution of the top 15 post author countries in the SHSS dataset. The 'unknown' category represents samples for which Tonneau et al. [66] state that inferring the author's country location was not possible.

Such linguistic diversity is particularly critical in hate speech detection, as offensive force is highly sensitive to lexical, pragmatic, and sociocultural variations. Research suggests that a phrase considered derogatory in one culture may be benign in another, leading to significant disparities in annotation consensus

and increased rates of false positives or negatives [74]. Furthermore, models trained on a single Spanish variant often experience performance decay when applied to others, as the terms used to denote hate may differ significantly across regions [69]. This "cultural gap" can lead to a substantial decrease in F1 scores and a dramatic increase in false negative rates, underscoring the risk of cultural insensitivity in language models [75].

By aggregating resources from multiple origins, including specific work on Mexican Spanish (HOMO-MEX) and other regional corpora, the superset built by Tonneau et al. [66] follows the recommendation to incorporate as many variants as possible during the training phase to develop models capable of handling the linguistic diversity of global online communities [69,75]. While we acknowledge the imbalance in geographical distribution, this multi-variant approach aims to mitigate the risk of cultural insensitivity and provide a more robust foundation for the subsequent classification.

4.2 Hate Speech Classes

Multi-class classification requires defining the classes in which text will be classified. Since our approach will convert binary-labelled hate speech datasets into multi-class ones, it is necessary to establish these classes beforehand. In the context of hate speech, a universally accepted categorisation of hate speech types remains elusive, mirroring the ongoing debate surrounding its precise definition. However, as the definition of hate speech itself indicates, this discourse is fundamentally targeted towards groups of individuals based on identity-related factors. Consequently, categorising hate speech by the targeted identity group is a common and logical approach.

The Measuring Hate Speech (MHS) dataset [55,76], a significant English-language hate speech corpus, underscores the importance of identifying the target identity group. Accordingly, the samples within this dataset are classified into seven identity groups: race/ethnicity, religion, national origin or citizenship status, gender, sexual orientation, age, and disability status. It should be noted that while the original MHS dataset provides a richer set of annotation dimensions (including categories such as sentiment, insult, humiliation, dehumanisation, violence, and genocide), the present work focuses exclusively on the target-group dimension.

In this work, the MHS dataset identity groups are used as the target hate speech classes for synthetic label generation, thereby classifying hate speech by identity group. This design choice is based on the premise that identifying the targeted group is a fundamental first step in hate speech detection. This approach provides a stable, categorical basis for evaluating the efficacy of synthetic labelling. While the resulting labels in this study are limited to target-group identities, the underlying methodology is designed to be agnostic to the specific categories used, making it applicable to any multi-class schema. Using this categorisation offers two key advantages: first, it employs a recognized hate speech target-group framework; and second, it allows performance evaluation by comparing our results to the human annotations provided within the MHS dataset. It should be noted that data from this dataset are used solely for evaluation purposes.

4.3 Automatic Annotation Architecture

This study aims to develop a strategy for classifying hateful texts by their target identity group, ultimately enabling the transformation of binary-labelled datasets into multi-class ones. Leveraging the SHSS dataset binary classification scheme (hateful/non-hateful), the annotation process therefore focuses on categorising the instances of hate speech contained within the dataset.

The automatic annotation architecture employs an LLM-as-a-judge approach. Specifically, the single answer grading variation is adopted, a strategy in which an LLM judge ($\mathcal{G}_{judge}$) is asked to assign a score

to a single answer [64]. In this context, the answer consists of one of the predefined identity groups and is generated beforehand by another LLM ($\mathcal{G}_{classifier}$). Regarding the possible scores, $\mathcal{G}_{judge}$ is configured to assign a binary score (accept/reject) to the classification generated by $\mathcal{G}_{classifier}$.

Leveraging the explainability of the LLM-as-a-judge approach, $\mathcal{G}_{classifier}$ will not only receive a classification score, but also LLM-generated feedback from $\mathcal{G}_{judge}$ when the classification is rejected. As a result, the proposed methodology is outlined as follows. Firstly, $\mathcal{G}_{classifier}$ receives the text and classifies it within the predefined identity groups. Then, given the text in question and the classification provided by $\mathcal{G}_{classifier}$, $\mathcal{G}_{judge}$ is asked to evaluate whether the classification is considered correct, justifying the answer. If the text is correctly classified, no further action is needed. Conversely, should the text not be correctly classified, this justification is sent as feedback to $\mathcal{G}_{classifier}$ to improve the classification outcome. This iterative process is limited to ten attempts; if $\mathcal{G}_{judge}$ does not accept $\mathcal{G}_{classifier}$ classification within this limit, the classification is finally rejected and, therefore, excluded from the dataset. This constraint has been implemented to prevent infinite loops and to limit computational costs. Fig. 2 illustrates this process.

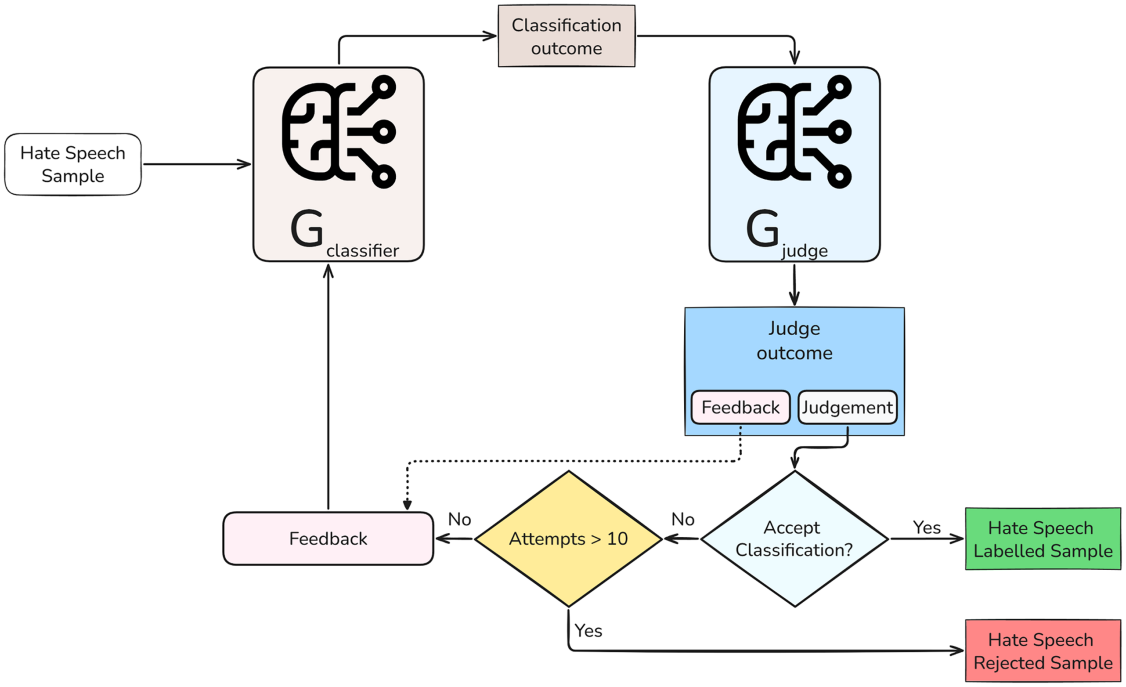

Figure 2: LLM-as-a-judge hate speech annotation strategy. First, a hate speech sample is fed into $\mathcal{G}_{classifier}$, which outputs a classification outcome. This outcome is then sent to $\mathcal{G}_{judge}$, which provides an answer including the judgement (accept or reject) and the corresponding feedback. If the judgement is "accept" the classification, the labelled hate speech sample is obtained. Conversely, if the judgement is "reject" the classification, the feedback provided is fed into $\mathcal{G}_{classifier}$, triggering a new iteration of the classification loop.

Fig. A1 illustrates the system prompt used to instruct the $\mathcal{G}_{classifier}$ model ($\mathcal{G}_{classifier}$) for the classification task. The prompt is structured into four distinct sections. Firstly, the model receives general contextual information regarding the task. Secondly, definitions for each target identity group are presented. The third section details specific instructions for the classification process, including constraints on the number of categories, an emphasis on objectivity and justification, and the encouragement to learn from previous errors. Finally, the expected output JSON format is specified.

Additionally, Fig. A2 presents the system prompt used to instruct the $\mathcal{G}_{judge}$ model ($\mathcal{G}_{judge}$) for the evaluation task. Similarly, the prompt is structured into four sections. The first two sections serve the same purpose as in the classification task, with the exception of the task definition. However, the instructions in the third section differ significantly, incorporating constraints on the number of categories, guidance against making judgements about the hateful nature of the comments, an explanation of the expected evaluation output, and directives to prioritise objectivity, justify responses, and select the most representative categories.

It should be noted that, while both prompts encourage the models to use a single identity group, they allow the annotation of samples with up to two identity groups. This decision was made because limiting annotation to a single group proved problematic, leading to endless disagreements between $\mathcal{G}_{classifier}$ and $\mathcal{G}_{judge}$ for samples concerning multiple identity groups. To mitigate this while remaining as close as possible to a multi-class structure, a maximum of two labels was permitted; this provided the necessary flexibility to resolve conflicts without allowing the annotations to become overly broad or fragmented. However, this decision entails the creation of a multi-label dataset instead of a multi-class one. Consequently, the resulting dataset is not suitable for multi-class classification, as a unique label is needed for every sample. However, this dataset is considered an intermediate outcome that will be further processed to be converted into a multi-class dataset. The strategy used for converting this multi-label dataset into a multi-class one is explained in Subsection Label Reduction Strategies in Section 4.3.

Fig. 3 illustrates the annotation process from a conversational perspective. As shown, the process begins with the aforementioned system prompts. Subsequently, a conversation begins between $\mathcal{G}_{classifier}$ and a simulated 'user' providing a hate speech sample. Afterwards, the context shifts to a conversation between $\mathcal{G}_{judge}$ and the classification produced by $\mathcal{G}_{classifier}$ in the initial conversation. Then, only when $\mathcal{G}_{judge}$ accepts this classification do both conversations conclude, resulting in an annotated sample. As an example, since the first evaluation in Fig. 3 is rejected, the judge's justification is fed into the conversation with $\mathcal{G}_{classifier}$. Afterwards, the process is repeated, the classification is accepted, and the annotated sample is obtained.

Finally, it is crucial to emphasise that this annotation methodology could be applied to any domain requiring multi-class data classification. Regardless of the nature of the data, this annotation strategy is suitable for converting any binary labelled dataset into a multi-class one, given the definitions of the different categories.

Label Reduction Strategies

While hate speech can target multiple identity groups simultaneously, this study aims to provide a dataset prepared for multi-class classification. This decision is driven by the objective of facilitating subsequent classification tasks for the research community. Given that hate speech detection is already a highly challenging problem due to dialectal variations and semantic ambiguity, a multi-label formulation can introduce additional noise and complexity that may hinder the performance of downstream models. Since the LLM-as-a-judge annotation outcomes may contain more than one label per instance, the application of a label reduction procedure is needed to obtain a single label representation. This ensures a cleaner signal for model training while maintaining the intermediate multi-label dataset as a resource for those interested in the transversal nature of the discourse.

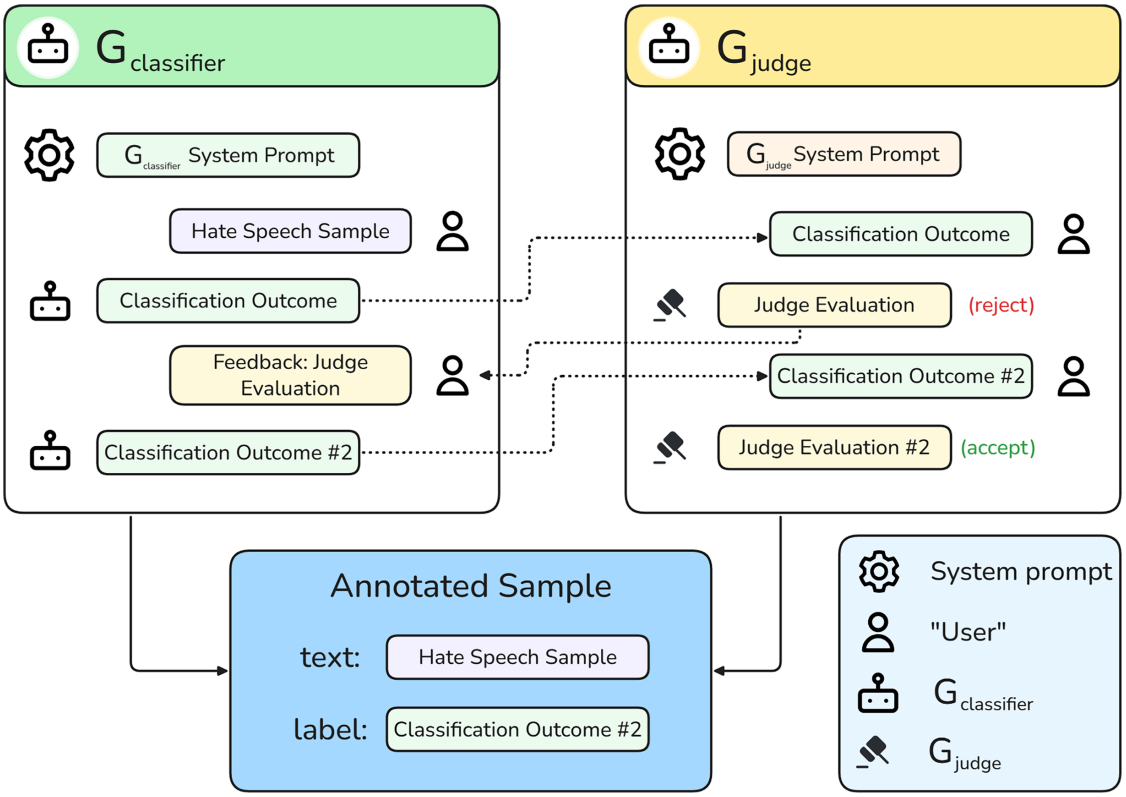

Figure 3: Conversation history for annotating a hate speech sample using an LLM-as-a-judge approach. First, both $\mathcal{G}_{classifier}$ and $\mathcal{G}_{judge}$ receive system prompt instructions. Subsequently, leveraging the user role, the hate speech sample is fed into $\mathcal{G}_{classifier}$, whose answer is the classification outcome. Then, the process shifts to the $\mathcal{G}_{judge}$'s conversation. The classification outcome is sent to $\mathcal{G}_{judge}$ using the user role. The judge ($\mathcal{G}_{judge}$) provides the evaluation outcome, which in this case, rejects the classification outcome. Therefore, the feedback contained in the judge's evaluation is then sent to $\mathcal{G}_{classifier}$. $\mathcal{G}_{classifier}$ considers $\mathcal{G}_{judge}$'s feedback and provides a new classification outcome. Finally, $\mathcal{G}_{judge}$ accepts this new classification outcome, and the annotated sample is obtained.

To determine the most effective method for resolving label ambiguity, three different label reduction strategies are evaluated. The first, minority-class assignment, is described as follows: let $y_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$ denote the set of labels associated with instance i, where $k \in \{1, 2\}$ in our setting. When $k = 1$, no modification is required. When $k = 2$, a single label must be selected. The selection rule for this strategy is based on the distribution of instances that are uniquely assigned to a single class within the considered dataset. Let $g(c)$ denote the number of instances labelled exclusively with class c (i.e., instances for which $y_j = \{c\}$). For an instance associated with two labels $y_i = \{c_a, c_b\}$, the selected label c^* is defined as:

$$c^* = arg \min_{c \in y_i} g(c)$$

That is, the label corresponding to the class with fewer uniquely assigned instances is selected. Equivalently, if the number of single-label instances is higher for class c_a than for class c_b, the instance is assigned to class c_b. In the event of a tie, the first class (c_a) is chosen. This criterion prioritises the comparatively less represented class among the already unambiguous (single label) instances, thereby discouraging systematic bias toward the more prevalent class and fostering class balance when resolving label ambiguity.

The second strategy is majority-class assignment, where for an instance associated with two labels $y_i = \{c_d, c_e\}$, the selected label c^* is defined as $c^* = arg\max_{c \in y_i} g(c)$. In this case, the comparatively most represented class is selected among the already single-labelled instances. Finally, the third strategy is random-class assignment, where the selected label c^* is chosen at random. To mitigate the variance inherent in this stochastic process, a resampling approach was used. Specifically, we generated 100 independent versions of the dataset using 100 random seeds (sampled from 0 to 1,000,000 via the *numpy. randint* function with a base seed of 16).

Each of these strategies is applied independently to the previous obtained multi-label data. By comparing these three strategies (minority, majority, and random), the trade-off between class distribution preservation and the potential distortion of true labels is evaluated, as detailed in the results (Section 5.1).

4.4 Implementation Details

The proposed architecture has been implemented using state-of-the-art LLMs, frameworks, and libraries for model management and deployment. A critical component of this implementation is the choice of models for $\mathcal{G}_{classifier}$ and $\mathcal{G}_{judge}$.

In our primary configuration, we employed the Llama 3.1 70B model [77] (quantized FP8 version, accessed via Hugging Face[1]) as $\mathcal{G}_{classifier}$ and the Gemma 3 27B model [78] as $\mathcal{G}_{judge}$. This deliberate decision to use two different models aims to ensure robust and objective annotation. By using separate LLMs, we intend to prevent reliance on a single model's inherent biases and encourage nuanced classification through divergent perspectives, thereby avoiding self-enhancement bias [64].

To address the potential impact of model selection and the asymmetry in model size (e.g., a smaller judge evaluating a larger classifier), a comprehensive ablation study has been conducted. This study evaluates several combinations of model families and sizes—including Qwen3.5 27B [79], Gemma 4 31B [80], and Llama 3.1 70B [77]—to verify the best configuration for the proposed architecture.

The annotation strategy has been implemented using the OpenAI agents SDK framework, configuring both $\mathcal{G}_{classifier}$ and $\mathcal{G}_{judge}$ as tool-less agents. To enhance response consistency and reliability, structured outputs were employed to enforce adherence to a predefined JSON schema for each agent. The classification schema incorporates fields for the identified class(es) and the corresponding justification, thereby prompting the LLM to ground its classifications in explicit reasoning. A separate schema governs the evaluation phase, including fields for the evaluation outcome and feedback. Specifically, this feedback field is the one fed into the $\mathcal{G}_{classifier}$ conversation when a classification is deemed incorrect.

For the deployment of these models, the vLLM serving system (version 0.8.3) [81], configured with a maximum sequence length of 4096 tokens, has been used. All remaining parameters have been maintained at their default settings. Each model has been instantiated on a single Nvidia H100 GPU equipped with 96 GB of VRAM.

4.5 Synthetic Annotation Evaluation Framework

Evaluating the proposed multi-class synthetic annotation strategy necessitates addressing specific data constraints. Notably, the SHSS is inherently binary and, therefore, lacks the ground truth labels required for multi-class validation. Consequently, to verify the reliability of the proposed methodology, it is necessary to use an alternative, human-labelled, multi-class dataset. In this work, the English Measuring Hate Speech (MHS) dataset is employed for this purpose.

[1] https://huggingface.co/RedHatAI/Meta-Llama-3.1-70B-Instruct-FP8

It should be noted that this cross-lingual validation introduces certain limitations. While the proposed methodology relies on state-of-the-art LLMs that are trained with multilingual data, performance may vary between English and Spanish. Therefore, the performance metrics obtained for the MHS dataset should be interpreted as a proxy for the method's effectiveness rather than a direct reflection of its performance on Spanish data.

Critically, the evaluation with the MHS evaluation data is performed completely independently of the SHSS data: the synthetic annotation strategy is applied directly to the evaluation data, whose multi-class ground truth labels are then used to validate the resulting classifications. Nevertheless, to mitigate cross-lingual limitations and ensure the method's applicability to the target resource, a small-scale human evaluation on a sample of the SHSS is also conducted. The remainder of this section is structured as follows: (1) a description of the evaluation data, (2) the definition of the evaluation metrics employed, (3) an overview of the established baseline, and (4) an explanation of the evaluation protocol.

4.5.1 Evaluation Data

The MHS dataset [55,76] is a corpus aligned with the United Nations definition of hate speech, as stated by Piot et al. [58]. The MHS dataset contains approximately 40,000 English comments, ranging from 4 to 600 characters in length, sourced from Twitter (40%), Reddit (40%), and YouTube (20%) between March and August 2019. The dataset incorporates 135,556 annotations, generated by 7912 annotators, with each comment reviewed by multiple individuals. Each comment in the dataset contains a 'hatespeech' label representing the annotator binary classification indicating whether the comment is considered hateful or not. In addition, as previously mentioned, each comment is also annotated across seven distinct target identity groups: race or ethnicity, religion, national origin or citizenship status, gender identity, sexual orientation, age, and disability status. Notably, the classification schema used for the ground truth is non-exclusive, as each sample is treated as a multi-label instance, where annotators may assign multiple identity groups where applicable. Consequently, the intersection of this multi-label structure with the fact that each comment is evaluated by multiple annotators makes the application of this dataset within our multi-class framework non-trivial. Table 2 shows three examples from the MHS dataset, including only the columns used in this study.

Firstly, to ensure inter-annotator agreement, it is necessary to unify the annotators' criteria. This process is divided into two separate phases. In the first phase, annotations are reconciled to determine whether each comment constitutes hate speech. To resolve annotation discrepancies, majority voting (using the 'hatespeech' label) across annotators is used to determine whether a comment is considered hateful or not, with ties resulting in exclusion from the dataset. This resulted in a final corpus containing 8089 hateful samples. The second phase focuses on merging the different specific target identity groups assigned by the annotators for the comments classified as hateful. For this purpose, an OR operation is applied to combine the identity groups identified by each annotator, allowing for multiple target identities per comment. These processing steps generate a hateful, multi-label subset (hereafter, Hate-MHS) of the original MHS dataset.

Secondly, since Hate-MHS samples originally follow a multi-label schema, a label reduction strategy must be applied. Accordingly, each strategy explained in Subsection Label Reduction Strategies in Section 4.3 is applied, separately, to the Hate-MHS dataset. As a result of this process, three different multi-class datasets derived from the HATE-MHS dataset are obtained: (1) Minority-MC-Hate-MHS, derived from minority-class assignment; (2) Majority-MC-Hate-MHS, derived from majority-class assignment; and (3) Random-MC-Hate-MHS, derived from random-class assignment. It should be noted that these techniques, which have been used here for converting multi-label evaluation data into multi-class, align with the methodology previously used for the unification of labels synthetically generated by both the proposed

approach and the baseline. This ensures consistency across our synthetically generated data and that used for evaluation.

Table 2: MHS dataset sample. The left column represents the post content. The 'hatespeech' column indicates whether the text is labelled as hateful (2) or non-hateful (0). The rest of the columns indicate the targeted identity group(s) of the comment. Ellipses are used in the examples within the table to omit potentially sensitive or context-specific content, ensuring neutrality and avoiding unintended ideological or political implications.

Text	Hatespeech	Target Race	Target Religion	Target Origin	Target Gender	Target Sexuality	Target Age	Target Disability
Yes indeed. She sort of reminds me. . . WE should be proud of our race no matter what it is!	0	True	False	False	False	False	False	False
The trans women reading this tweet right now is beautiful.	0	False	False	False	True	False	False	False
Get rid of them and their <derogatory term for children of undocummented immigrants> and grand babies. <profanity> worthless <origin-related insult>	2	True	False	True	False	False	False	False

In summary, processing the original MHS corpus resulted in four distinct datasets. The first, Hate-MHS, is a multi-label dataset containing all hateful samples identified through a majority voting consensus among annotators. The remaining three are multi-class variants of the Hate-MHS dataset, derived by consolidating these multi-label annotations into mutually exclusive categories using different label reduction strategies: (i) Minority-MC-Hate-MHS, (ii) Majority-MC-Hate-MHS, and (iii) Random-MC-Hate-MHS. These datasets are used independently for evaluation purposes. Specifically, Hate-MHS is employed to assess classification performance prior to label unification, whereas the multi-class variants allow for an analysis of how different unification processes impact overall performance.

The distribution of identity labels within the Hate-MHS dataset is illustrated in Fig. 4. As shown in Fig. 4a, the most prevalent categories are *gender* and *race*, while *age* has the lowest representation. To capture the complexity of the identities represented, the overlap between labels is also analysed. The co-occurrence heatmap in Fig. 4b reveals the intersectional patterns of the dataset, showing that the most frequent combinations occur between *gender* and *sexuality*, *gender* and *race*, and *origin* and *race*. Other categories, such as *age* and *disability*, exhibit minimal overlap.

Turning to multi-class variants, Fig. 5 depicts the identity label distribution for each of the three label-reduction strategies: minority-class assignment (Fig. 5a), majority-class assignment (Fig. 5b), and random-class assignment (Fig. 5c). While the former two follow deterministic unification strategies, the distribution for the random variant has been estimated by averaging the results of 100 randomly resampled datasets to ensure statistical stability. Comparing the three variants, the minority-class assignment leads to a more balanced distribution, effectively increasing the representation of under-represented groups. In contrast, the majority-class assignment tends to amplify the dominance of the most frequent labels, further skewing the dataset. The random approach provides a neutral distribution that highlights the intentional bias introduced by deterministic strategies.

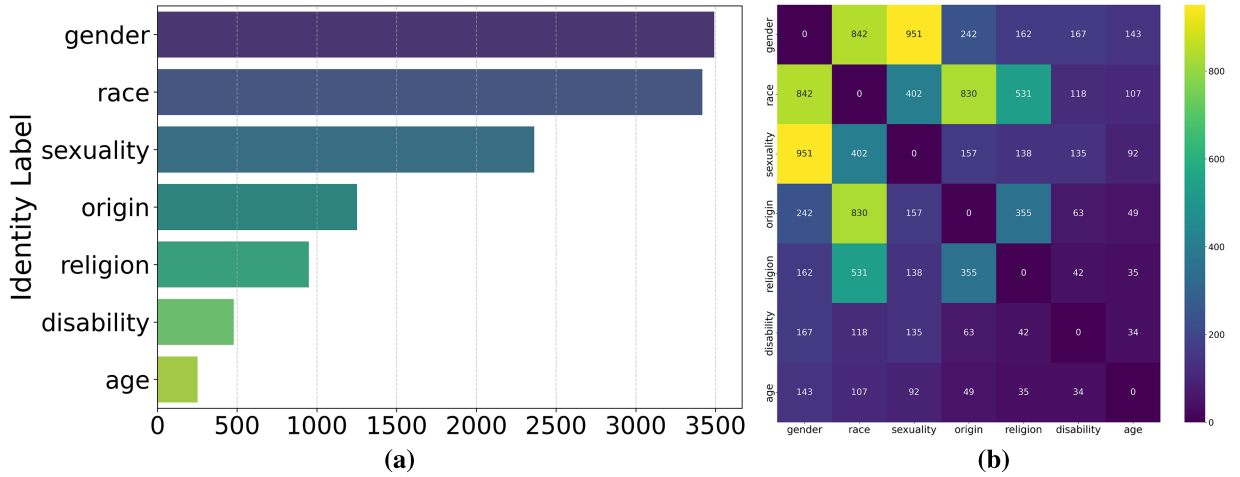

Figure 4: Distribution and intersectionality of identity labels in the Hate-MHS dataset. (**a**) presents the aggregate label distribution, showing the total frequency of each identity group. To account for multi-label instances, occurrences are summed independently across all combinations. (**b**) shows the label co-occurrence heatmap, where the matrix illustrates the intersectional overlap between categories.

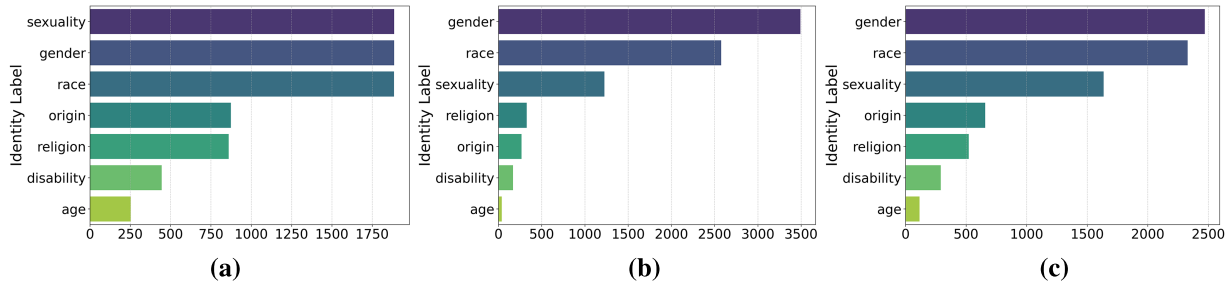

Figure 5: Identity label distributions for multi-class Hate-MHS unification variants. (**a**) minority-class assignment (Minority-MC-Hate-MHS), (**b**) majority-class assignment (Majority-MC-Hate-MHS), and (**c**) random-class assignment (Random-MC-Hate-MHS). The random distribution is reported as the mean of 100 resampled datasets.

The entire workflow for obtaining evaluation data is illustrated in Fig. 6, using minority-class assignment as the label-reduction strategy for the example. For instance, for a given comment (Comment A) present in the MHS dataset, the different annotations are considered. Firstly, majority voting is performed using the 'hatespeech' labels. Since two annotators classify this comment as hateful (two green, dotted hatespeech boxes) and only one as non-hateful (one solid red hatespeech box), the sample is considered hateful. If the sample were considered non-hateful, it would be discarded at this point. Next, since the comment in question is considered hateful, an OR operation is applied across all the groups identified by the annotators. This leads to "Multi-label comment A", which is included in the Hate-MHS dataset. Finally, from the different identity groups present in "Multi-label comment A", the least represented group in Hate-MHS is chosen, thereby creating "Multi-class comment A", which is included in Minority-MC-Hate-MHS.

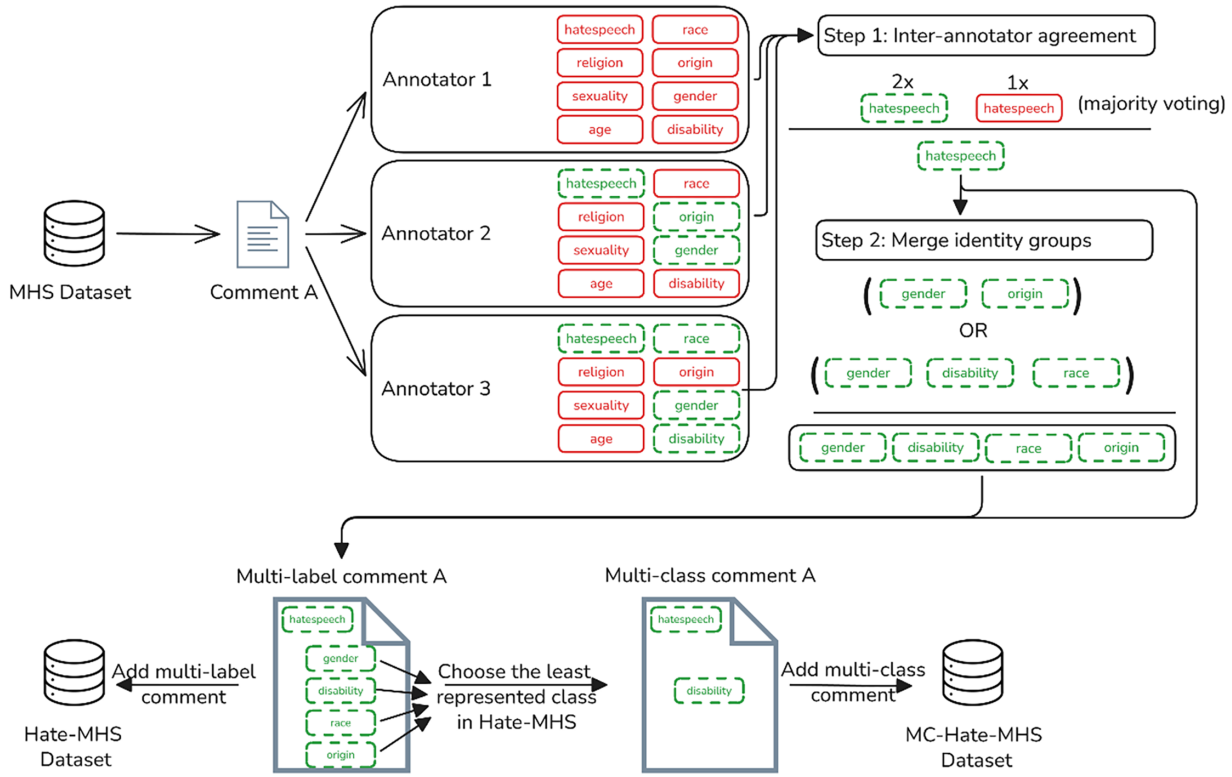

Figure 6: Complete workflow for obtaining evaluation data (Minority-MC-Hate-MHS).

4.5.2 Evaluation Metrics

Given the non-trivial nature of evaluating our synthetic annotation strategy, some granular performance metrics have been defined, alongside Cohen's Kappa coefficient, to measure the inter-annotator agreement between synthetic and human labels. Let N denote the total number of samples in the corresponding evaluation dataset. For each sample i, let $\hat{Y}_i$ represent the set of predicted classes and Y_i represent the set of ground truth labels. The indicator function $\mathbb{1}(\cdot)$ is used, which outputs 1 if the condition is true and 0 otherwise.

Set-Level Agreement

These metrics evaluate the correspondence between the predicted and ground truth sets of labels.

- **Full Match:** counts instances where the predicted labels are identical to the complete set of ground truth labels. See Eq. (1).

$$\text{Full Match} = \sum_{i=1}^{N} \mathbb{1}\left(\widehat{Y}_i = Y_i\right) \tag{1}$$

- **Match:** represents instances where at least one of the labels in the predicted set is contained within the ground truth set. See Eq. (2).

$$\text{Match} = \sum_{i=1}^{N} \mathbb{1}\left(\widehat{Y}_i \cap Y_i \neq \varnothing\right) \tag{2}$$

- **Full Mismatch:** counts instances where the predicted labels are entirely absent from the ground truth set. See Eq. (3).

$$\text{Full Mismatch} = \sum_{i=1}^{N} \mathbb{1}\left(\widehat{Y}_i \cap Y_i = \varnothing\right) \tag{3}$$

Label-Level Agreement

To account for the information loss inherent in multi-class systems, label-level metrics are defined to track the total number of labels captured or missed.

- **Individual Match:** quantifies the total number of ground truth labels that were captured by the model. See Eq. (4).

$$\text{Individual Match} = \sum_{i=1}^{N} \left|\widehat{Y}_i \cap Y_i\right| \tag{4}$$

- **Individual Mismatch:** quantifies the total number of ground truth labels that were not captured by the model. This accounts for both incorrect predictions and 'ignored' labels. See Eq. (5).

$$\text{Individual Mismatch} = \sum_{i=1}^{N} \left|Y_i \backslash \widehat{Y}_i\right| \tag{5}$$

Per-Class Distribution

For a more granular analysis of model bias, the latter metrics are extended to a per-class basis for each identity group c.

- **Individual Match per Class:**

$$\text{Individual Match}_c = \sum_{i=1}^{N} \mathbb{1}\left(c \in \widehat{Y}_i \wedge c \in Y_i\right) \tag{6}$$

- **Individual Mismatch per Class:**

$$\text{Individual Mismatch}_c = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}\left(\mathrm{c} \in \mathrm{Y_i} \wedge \mathrm{c} \notin \widehat{\mathrm{Y}}_\mathrm{i}\right) \tag{7}$$

Cohen's Kappa Coefficient

Cohen's Kappa Coefficient (κ) [82] measures the agreement between two annotators beyond chance. Hence, this metric is of paramount importance for this evaluation, since the ability of the proposed architecture to substitute human annotations is being assessed. Eq. (8). presents its formula.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

where p_o is the proportion of samples on which the raters agree, and p_e is the proportion of agreement that would be expected by chance. For k categories and N observations, where n_{ki} is the number of times rater i predicted category k, p_o and p_e are defined as follows:

$$p_o = \frac{Number\ of\ aggrements}{N}; p_e = \sum_k \frac{n_{k1}}{N}\frac{n_{k2}}{N} \tag{9}$$

According to Landis and Koch [83], values of this metric ranging between 0.41 and 0.60 denote a moderate agreement, while values between 0.61 and 0.80 indicate substantial agreement.

4.5.3 Baseline

The proposed LLM-as-a-judge synthetic annotation approach is evaluated against a prompt-only classification baseline. This baseline employs a single LLM guided by a descriptive classification prompt, mirroring established methodologies in previous research [60,84,85]. To ensure a controlled experimental environment, this baseline is implemented using the same LLM used in $\mathcal{G}_{classifier}$ and the specific prompt outlined in Fig. A3. Furthermore, consistent with the annotation guidelines used for $\mathcal{G}_{classifier}$, the model is also permitted to assign samples to a maximum of two identity groups. This alignment is maintained to ensure the highest degree of comparability and reliability between the two approaches.

4.5.4 Evaluation Protocol

Although the target task is formulated as a multi-class classification problem, both the proposed and the prompt-only approaches are modelled in a way that allows them to assign up to two labels to a given instance. These multi-label outputs should be understood as intermediate results that are subsequently processed, using the label reduction strategies explained in Section 4.3, to be converted into the final multi-class results. Accordingly, the evaluation protocol is designed to assess both: (i) the raw synthetic classification outputs prior to label reduction, and (ii) the final single-label predictions used for the multi-class setting.

Evaluation of Raw Predictions (Multi-Label Setting)

In the first stage, the synthetic classification outcomes are evaluated without any post-processing. Since both synthetic classification approaches (LLM-as-a-judge and prompt-only) may assign up to two labels to a given instance, predictions are treated as multi-label outputs. These predictions are compared against the annotations of the multi-label evaluation dataset Hate-MHS, in which each text can be associated with one or more classes.

This evaluation provides insight into the model's ability to identify all relevant categories associated with an instance before enforcing the single-label constraint required for multi-class classification. No label reduction strategy is applied at this stage.

Evaluation after Label Reduction (Multi-Class Setting)

In the second stage, synthetic classification outcomes are transformed to comply with the multi-class classification setting. When more than one label is assigned to an instance, a single label is selected by applying the label reduction strategies described in Section 4.3. This is the same procedure used to construct the three multi-label variants of Hate-MHS, ensuring methodological consistency between prediction, post-processing, and ground-truth preparation prior to computing the multi-class evaluation metrics.

The resulting single-label predictions of both approaches (LLM-as-a-judge and prompt-only) are then evaluated against the three versions of multi-class ground truth (Minority-MC-Hate-MHS, Majority-MC-Hate-MHS, and Random-MC-Hate-MHS). This evaluation reflects the performance of different approaches under the task for which they are ultimately intended (multi-class classification).

5 Results

Table 3 describes the results obtained using raw multi-label predictions for both the LLM-as-a-judge and prompt-only approaches. The LLM-as-a-judge approach slightly outperformed the baseline, achieving higher *Match* (94.09% vs. 93.89%) and *Individual Match* (77.24% vs. 76.75%) rates. In addition, it yielded

lower percentages (where lower values indicate the better performance) for the *Full Mismatch* (5.08% vs. 6.11%) and *Individual Mismatch* (22.76% vs. 23.25%) metrics. To ensure a fair comparison, all percentages for the *Full Match, Match*, and *Full Mismatch* metrics are calculated using the total sample size (N = 8089) as the denominator for both approaches. Consequently, any unlabelled samples are treated as incorrect predictions, thereby penalising such instances.

Table 3: Evaluation results using the LLM-as-a-judge (Judge) and prompt-only approaches with the Hate-MHS dataset.

Metric	Hate-MHS			
	Judge		Prompt-Only	
	n	%	n	%
Matches				
Full Match	3587	44.34	3605	**44.57**
Match	7611	**94.09**	7595	93.89
Individual Match	9354	**77.24**	9367	76.75
Mismatches				
Full Mismatch	411	**5.08**	494	6.11
Individual Mismatch	2756	**22.76**	2838	23.25
Total Samples	8022	–	8089	–

Note: The best results comparing the Judge and prompt-only approaches are highlighted in bold.

Regarding the number of labelled instances, the prompt-only approach successfully labelled the 8089 samples, while the proposed approach labelled 8022.

Table 4 presents the *Individual Match per class* for the Hate-MHS dataset. The LLM-as-a-judge approach outperforms the prompt-only baseline for all identity groups except for *gender* and *origin*. Most notably, the Judge approach shows a clear advantage in the *age* (23.32% vs. 18.58%) and *disability* (65.55% vs. 62.42%) categories. Conversely, the prompt-only approach retains a slight edge in the *gender* and *origin* categories. These results suggest that the Judge framework is more effective, providing a more balanced performance across the entire spectrum of identity groups.

Table 4: Individual match per class results using the LLM-as-a-judge (Judge) and prompt-only approaches with the Hate-MHS dataset.

	Age		Disability		Gender		Origin		Race		Religion		Sexuality	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Judge	59	**23.32**	314	**65.55**	2864	82.04	992	79.23	2460	**71.99**	758	**79.87**	1910	**80.80**
Prompt-only	47	18.58	299	62.42	2948	**84.45**	1029	**82.19**	2398	70.18	750	79.03	1896	80.20

Note: The best results comparing the Judge and prompt-only approaches are highlighted in bold.

Finally, per-class Cohen's Kappa scores for both the LLM-as-a-judge and prompt-only approaches are shown in Table 5. Comparisons between both approaches were performed on the full dataset (n = 8089), assigning an empty prediction to any example where the LLM-as-a-judge approach did not produce a label. The scores show that the proposed approach outperformed the prompt-only strategy in most cases, except for *gender* and *sexuality*, where the prompt-only approach achieved slightly higher agreement. The mean

per-class Cohen's Kappa is 0.60 for the LLM-as-a-judge approach and 0.59 for the prompt-only strategy. To assess statistical significance, a paired bootstrap procedure with 2000 samples was applied, computing the mean per-class Cohen's Kappa for each strategy in each bootstrap sample and generating a distribution of the differences (LLM-as-a-judge-prompt-only). The difference in Cohen's Kappa is 0.011 with a 95% confidence interval of 0.0024 to 0.019 (p = 0.01). This indicates a small but statistically significant improvement for the LLM-as-a-judge approach, reflecting its overall better performance.

Table 5: Cohen's Kappa scores using the LLM-as-a-judge (Judge) and prompt-only approaches with the Hate-MHS dataset.

Class	Hate-MHS	
	Judge	**Prompt-Only**
Age	**0.31**	0.27
Disability	**0.67**	0.63
Gender	0.64	**0.65**
Origin	**0.53**	0.51
Race	**0.70**	0.69
Religion	0.74	0.74
Sexuality	0.63	**0.64**
Mean	**0.60**	0.59

Note: The best results comparing the Judge and prompt-only approaches are highlighted in bold.

5.1 Label-Reduction Strategy Comparison

Following the procedure detailed in Section 4.5.4, three distinct label-reduction strategies are applied to the raw multi-label dataset to derive multi-class labels. Table 6 summarizes the evaluation results comparing the LLM-as-a-judge approach against the prompt-only baseline across these three datasets. It should be noted that, for these unified datasets, *Full Match*, *Match*, and *Individual Match* are consolidated into a single *Matches* row due to their identical values; the same applies to mismatches. This is because each instance in the multi-class setting is associated with exactly one identity group, meaning that both Y and $\hat{Y}$ are singleton sets.

Table 6: Evaluation results using the LLM-as-a-judge (Judge) and prompt-only approaches with the three multi-class dataset variants.

Dataset	Strategy		Matches	Mismatches	Total Samples
Minority-MC-Hate-MHS	**Judge**	**n (%)**	5093 (**62.96**)	2929 (**36.21**)	8022
	Prompt-only	**n (%)**	5085 (62.86)	3004 (37.14)	8089
Majority-MC-Hate-MHS	**Judge**	**n (%)**	5650 (69.85)	2372 (**29.32**)	8022
	Prompt-only	**n (%)**	5677 (**70.18**)	2412 (29.82)	8089
Random-MC-Hate-MHS*	**Judge**	**Avg. n (%) [CI]**	4733.22 (**58.51**) [4726.12–4740.32]	3288.78 (**40.66**) [3281.68–3295.88]	8022
	Prompt-only	**Avg. n (%) [CI]**	4719.74 (58.35) [4712.41–4727.07]	3369.26 (41.65) [3361.93–3376.59]	8089

Note: The best results comparing the Judge and prompt-only approaches are highlighted in bold. *Values for Random-MC-Hate-MHS respresent the average results and 95% confidence interval over 100 runs.

The results indicate that the choice of label-reduction strategy influences the matching rates. In the Minority-MC-Hate-MHS dataset, the Judge approach slightly outperformed the prompt-only baseline (62.96% vs. 62.86%). A similar trend is observed in the Random-MC-Hate-MHS dataset, where the Judge approach achieved a higher average matching rate of 58.51% compared to 58.35% for the prompt-only baseline. Conversely, in the Majority-MC-Hate-MHS dataset, the prompt-only baseline demonstrated a slight advantage (70.18% vs. 69.85%).

A more granular analysis of individual classes (Table 7) reveals that the Judge approach generally yields superior performance. Regardless of the label-reduction strategy, the LLM-as-a-judge approach outperforms the baseline for four out of seven identity groups (*age*, *disability*, *race*, and *religion*). Notably, the advantages of the LLM-as-a-judge approach are most pronounced in the Minority-MC-Hate-MHS dataset, where it achieves higher match rates for underrepresented groups (*age*, *disability*, and *religion*) while maintaining competitive performance for more represented identity groups (e.g., *origin* or *race*).

Table 7: Individual match per class results using the LLM-as-a-judge (Judge) and prompt-only approaches with the three multi-class dataset variants.

Class/Approach	Age	Disability	Gender	Origin	Race	Religion	Sexuality
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Minority-MC-Hate-MHS							
Judge	59 **(23.3%)**	299 **(67.2%)**	1096 (58.1%)	561 (64.2%)	1250 **(66.3%)**	697 **(81.1%)**	1131 (60.0%)
Prompt-only	47 (18.6%)	287 (64.5%)	1124 **(59.6%)**	582 **(66.6%)**	1219 (64.7%)	689 (80.1%)	1137**(60.3%)**
Total per class	253	445	1886	874	1885	860	1886
Majority-MC-Hate-MHS							
Judge	5 **(13.2%)**	105 **(62.5%)**	2861 (82.0%)	153 **(57.5%)**	1637 **(63.6%)**	105 **(32.2%)**	784 (64.0%)
Prompt-only	4 (10.5%)	101 (60.1%)	2948 **(84.5%)**	151 (56.8%)	1586 (61.6%)	98 (30.1%)	789 **(64.4%)**
Total per class	38	168	3491	266	2575	326	1225
Random-MC-Hate-MHS*							
	n (%) [CI]	n (%) [CI]	n (%) [CI]	n (%) [CI]	n (%) [CI]	n (%) [CI]	n (%) [CI]
Judge	20.6 **(17.8%)** [20–21]	163.6 **(56.2%)** [162–164]	1526 (61.8%) [1521–1530]	342.9 (52.2%) [340–345]	1355 **(58.2%)** [1352–1359]	283.6 **(54.4%)** [281–285]	1040 **(63.6%)** [1037–1043]
Prompt-only	16.6 (14.3%) [15–17]	156.5 (53.7%) [155–157]	1586 **(64.2%)** [1582–1591]	342.9 (52.1%) [340–345]	1306 (56.1%) [1301–1310]	275.9 (52.9%) [273–278]	1035 (63.3%) [1032–1038]
Total per class	115.95	291.22	2470.51	657.47	2329.22	521.53	1636.10

Note: The best results comparing the Judge and prompt-only approaches are highlighted in bold. The best results comparing the different label-reduction strategies are underlined. *Values for Random-MC-Hate-MHS represent the average individual match and 95% confidence interval over 100 runs.

Furthermore, Table 7 provides insights into why the highest match rate is associated with the majority-class assignment. This superiority is driven by significant class imbalance; for example, the *gender* category in the Majority-MC-Hate-MHS contains 3491 samples, vastly outweighing the *age* class (38 samples). Such a distribution often leads to inflated accuracy due to majority-class bias. This is further evidenced by the Cohen's Kappa scores in Table 8; despite the high raw match rates, the majority strategy yields a mean Kappa of 0.48, which is notably lower than the 0.54 achieved by the minority strategy. This indicates that the agreement in the majority dataset is less robust and more susceptible to the effects of the skewed class distribution.

Table 8: Cohen's Kappa scores using the LLM-as-a-judge (Judge) and prompt-only approaches with the three multi-class variants.

Class/Approach	Age	Disability	Gender	Origin	Race	Religion	Sexuality	Mean
Minority-MC-Hate-MHS								
Judge	**0.31**	**0.67**	0.51	**0.43**	**0.62**	**0.72**	0.53	**0.54**
Prompt-only	0.27	0.63	**0.53**	0.41	0.61	0.71	**0.54**	0.53
Majority-MC-Hate-MHS								
Judge	0.19	0.65	0.64	0.23	**0.61**	**0.40**	0.63	0.48
Prompt-only	0.19	0.65	**0.65**	0.22	0.60	0.38	**0.64**	0.48
Random-MC-Hate-MHS*								
Judge	**0.22 [0.21–0.23]**	**0.55 [0.54–0.55]**	0.47 [0.47–0.47]	**0.32 [0.31–0.32]**	**0.54 [0.54–0.54]**	**0.48 [0.47–0.48]**	0.51 [0.51–0.51]	**0.44 [0.44–0.44]**
Prompt-only	0.20 [0.19–0.21]	0.52 [0.52–0.53]	**0.48 [0.48–0.48]**	0.30 [0.29–0.30]	0.53 [0.53–0.53]	0.46 [0.46–0.47]	**0.52 [0.52–0.52]**	0.43 [0.42–0.43]

Note: The best results comparing the Judge and prompt-only approaches are highlighted in bold. The best results comparing the different label-reduction strategies are underlined. *Values for Random-MC-Hate-MHS represent the average Cohen's Kappa (κ) and 95% confidence interval over 100 runs.

Regarding random-class assignment, the results serve as a neutral baseline, with performance that generally falls between the minority and majority strategies. While the matching rates for the Random-MC-Hate-MHS dataset are relatively stable, they do not exhibit the same level of class-specific robustness as the minority strategy. As shown in Table 8, the mean Cohen's Kappa for the random variant (0.44) is the lowest of the three strategies. This suggests that, while random assignment avoids the extreme bias of the majority strategy, it does not intentionally optimise for the identification of underrepresented groups.

Therefore, to provide a definitive answer to RQ3, and considering the balanced class distribution presented in Table 7 and the superior agreement metrics in Table 8, the minority-class assignment is the optimal strategy, as it yields the highest statistical reliability ($\kappa = 0.54$) while ensuring a fair evaluation across all identity groups.

Despite the slightly higher agreement observed in most classes, and contrary to the results obtained in the multi-label setting, the paired bootstrap analysis (2000 resamples) for the multi-class setting yielded a mean difference (LLM-as-a-judge-prompt-only) of 0.010, with a 95% confidence interval of −0.001 to 0.021 ($p = 0.11$). As the confidence interval includes zero, the p-value is greater than 0.05. Consequently, the difference is not statistically significant, thereby suggesting comparable performance between both strategies after applying the label-reduction technique.

The results obtained enable a comprehensive answer to our initial research questions. Regarding RQ1, it has been demonstrated that the LLM-as-a-judge approach can be effectively integrated into a synthetic annotation architecture, as this integration has increased annotation performance. In response to our second research question (RQ2), the proposed LLM-based judging approach achieves moderate agreement with human annotators, improving Cohen Kappa scores by 1% compared to prompt-only approaches. Nevertheless, these results were obtained using the initial LLM-as-a-judge configuration (Llama3.1-70B and Gemma3-27B). In Section 5.2, an ablation study on model selection is performed to determine the optimal configuration.

5.2 Ablation Study: Model Selection for Classifier and Judge

To determine the optimal configuration of LLMs for the proposed pipeline, we conducted an ablation study, evaluating different combinations of models acting as the initial classifier and the subsequent judge. Following the previous findings, the minority-class assignment was used as the label-reduction strategy for all configurations. The results are summarised in Tables 9–11.

Table 9: Matches and mismatches evaluation results for the ablation study on the Minority-MC-Hate-MHS.

Models (Classifier/Judge)	Matches		Mismatches		Total Labelled Samples	% of MHS Samples Labelled
	n	%	n	%		
Llama3-70B/Gemma3-27B	5093	62.96%	2929	36.21%	8022	**99.17%**
Gemma3-27B/Llama3-70B	4908	60.67%	2960	36.59%	7868	97.27%
Qwen3.5-27B/Gemma4-31B	5366	66.34%	1568	**19.38%**	6934	85.72%
Gemma4-31B/Qwen3.5-27B	5850	**72.32%**	1890	23.37%	7740	95.69%
Qwen3.5-27B/Llama3-70B	5703	70.50%	2094	25.89%	7797	96.39%
Llama3-70B/Qwen3.5-27B	5196	64.24%	2503	30.94%	7699	95.18%
HATE-MHS Samples					8089	100.00%

Note: The best results comparing different models are highlighted in bold.

Table 10: Individual matches per class results for the ablation study on the Minority-MC-Hate-MHS.

Models	Age		Disability		Gender		Origin		Race		Religion		Sexuality	
	IM	%	IM	%	IM	%	IM	%	IM	%	IM	%	IM	%
Llama3/Gemma3	59	23.32	299	67.19	1096	58.11	561	64.19	1250	66.31	**697**	**81.05**	1131	59.97
Gemma3/Llama3	**76**	**30.04**	283	63.60	1018	53.98	539	61.67	1247	66.15	682	79.30	1063	56.36
Qwen3.5/Gemma4	58	22.92	**331**	**74.38**	1211	64.21	554	63.39	1343	71.25	615	71.51	1254	66.49
Gemma4/Qwen3.5	71	28.06	329	73.93	**1425**	**75.56**	554	63.39	**1493**	**79.20**	645	75.00	**1333**	**70.68**
Qwen3.5/Llama3	62	24.51	329	73.93	1367	72.48	546	62.47	1425	75.60	661	76.86	1313	69.62
Llama3/Qwen3.5	54	21.34	302	67.87	1108	58.75	**578**	**66.13**	1304	69.18	662	76.98	1188	62.99
n per class	253		445		1886		874		1885		860		1886	

Note: The best results comparing different models are highlighted in bold.

Table 11: Cohen's Kappa scores for the ablation study on the Minority-MC-Hate-MHS. Class.

	Llama3/Gemma3	Gemma3/Llama3	Qwen3.5/Gemma4	Gemma4/Qwen3.5	Qwen3.5/Llama3	Llama3/Qwen3.5
Age	0.31	0.29	0.31	**0.34**	0.31	0.30
Disability	0.67	0.56	0.68	**0.69**	0.65	0.68
Gender	0.51	0.47	0.60	**0.67**	0.66	0.51
Origin	0.43	0.49	0.60	**0.62**	0.56	0.54
Race	0.62	0.63	0.69	**0.72**	0.70	0.65
Religion	0.72	0.71	0.73	**0.74**	0.73	0.73
Sexuality	0.53	0.50	0.70	**0.71**	0.70	0.56
Mean	0.54	0.52	0.62	**0.64**	0.62	0.56

Note: The best results comparing different models are highlighted in bold.

Overall performance across different model pairs is detailed in Table 9. The combination of Gemma4-31B (Classifier) and Qwen3.5-27B (Judge) achieved the highest overall agreement, reaching a *Match* rate

of 72.32% and labelling 7740 samples (95.69% of the dataset). While the Qwen3.5-27B/Llama3-70B pair also showed strong results (70.50% *Match*), the Gemma4-31B/Qwen3.5-27B combination demonstrated a superior ability to align predictions with the ground truth.

The impact of these combinations on specific identity groups is further explored through the *Individual Match per class* metric (Table 10). The Gemma4-31B/Qwen3.5-27B configuration again emerged as the top performer, achieving the highest match rates for *gender* (75.56%), *race* (79.20%), and *sexuality* (70.68%). Interestingly, while this pair dominated the most represented classes, the Gemma3-27B/Llama3-70B combination proved more effective for the *age* category (30.04%), suggesting that different model architectures may possess different sensitivities to specific identity markers.

To validate the reliability of these agreements, we analysed the per-class Cohen's Kappa scores (Table 11). The results confirm a significant performance leap when employing the Qwen3.5 and Gemma4 models. Both the Qwen3.5-27B/Gemma4-31B and Gemma4-31B/Qwen3.5-27B configurations achieved the highest mean Kappa scores ($\kappa = 0.62$ and $\kappa = 0.64$, respectively), substantially outperforming combinations involving Llama3-70B and Gemma3-27B (which ranged from 0.52 to 0.54). Specifically, the Gemma4-31B/Qwen3.5-27B pair showed particularly strong agreement in *sexuality* (0.71), *race* (0.72), and *religion* (0.74), reinforcing the robustness of this combination. Nevertheless, the Qwen3.5-27B/Llama3.1-70B combination also achieved a strong Kappa score ($\kappa = 0.62$).

In conclusion, the results of the ablation study provide address RQ4, demonstrating that the choice of models significantly influences the pipeline's effectiveness. The synergy between Gemma4-31B as the classifier and Qwen3.5-27B as the judge offers the best balance between coverage and accuracy, consistently yielding the highest match rates and the strongest statistical agreement across most identity groups.

Building upon these findings, we continue evaluating whether the proposed LLM-as-a-judge approach provides a tangible advantage over the prompt-only approach. To ensure a fair and rigorous comparison, we employed both Gemma4-31B and Qwen3.5-27B as sole classifiers for the prompt-only strategy, as these two models constituted the top-performing combination in the ablation study. The results of these configurations are detailed in Tables 12–14.

As shown in Table 12, the prompt-only approach yielded overall match rates of 61.27% for Gemma4-31B and 69.69% for Qwen3.5-27B. Although Qwen3.5-27B performed better as a standalone classifier, both models were significantly outperformed by the Gemma4-31B/Qwen3.5-27B LLM-as-a-judge configuration. This configuration demonstrated superior performance compared to the prompt-only baselines in terms of both match (72.32%) and mismatch rates (27.32%)

Table 12: Evaluation results for the prompt-only approach (baseline) using Gemma4-31B and Qwen3.5-27B on the Minority-MC-Hate-MHS dataset.

Metric	Prompt-Only (Gemma4-31B)		Prompt-Only (Qwen3.5-27B)	
	n	%	n	%
Matches	4956	61.27	5637	**69.69**
Mismatches	3133	38.73	2452	**30.31**
Total Samples	8089	100.00	8089	100.00

Note: The best results comparing both models are highlighted in bold.

Table 13: Individual match rates per class results for Gemma4-31B and Qwen3.5-27B using the prompt-only approach (baseline) on the Minority-MC-Hate-MHS dataset.

	Age		Disability		Gender		Origin		Race		Religion		Sexuality	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Prompt-only (Gemma4)	59	23.32	327	73.48	1039	55.09	512	58.58	1167	61.91	589	68.49	1263	66.97
Prompt-only (Qwen3.5)	60	**23.72**	332	**74.61**	1319	**69.93**	546	**62.47**	1371	**72.73**	682	**79.30**	1327	**70.36**
Total per class	253		445		1886		874		1885		860		1886	

Note: The best results comparing both models are highlighted in bold.

Table 14: Cohen's Kappa scores for Gemma4-31B and Qwen3.5-27B obtained using the prompt-only approach (baseline) on the Minority-MC-Hate-MHS dataset.

	Age	Disability	Gender	Origin	Race	Religion	Sexuality	Mean
Prompt-only (Gemma4)	**0.30**	**0.67**	0.52	**0.59**	0.62	0.71	**0.69**	0.59
Prompt-only (Qwen3.5)	0.29	0.61	**0.64**	0.53	**0.68**	**0.72**	0.67	0.59

Note: The best results comparing both models are highlighted in bold.

An analysis of individual identity groups (Table 13) reveals that Qwen3.5-27B consistently achieved higher match rates than Gemma4-31B across all categories, with the most pronounced differences observed in *gender* (69.93% vs. 55.09%), *race* (72.73% vs. 61.91%), and *religion* (79.30% vs. 68.49%). However, both models showed similar difficulties in the *age* category, with match rates of 23.32% (Gemma4-31B) and 23.72% (Qwen3.5-27B). Regarding the differences relative to the Gemma4-31B/Qwen3.5-27B-judge setting, the LLM-as-a-judge approach outperformed standalone (baseline) Qwen3.5-27B in all categories except *disability* and *religion*, where Qwen3.5-27B achieved slightly better results (74.61% vs. 73.93% and 79.30% vs. 75.00%, respectively).

Regarding statistical agreement, both prompt-only configurations yielded an identical mean Cohen's Kappa score of 0.59 (Table 14), which was lower than the 0.64 achieved by the Gemma4-31B/Qwen3.5-27B LLM-as-a-judge configuration. In this case, the LLM-as-a-judge approach outperformed the prompt-only baseline in each category. To assess the significance of these differences, paired bootstrap analyses (2000 resamples) were conducted. The LLM-as-a-judge framework showed a statistically significant improvement over the Gemma4-only approach, with a mean difference of 0.057 (95% CI $[0.0466, 0.0674]$, $p < 0.001$), and over the Qwen3.5-only approach, with a mean difference of 0.052 (95% CI $[0.0390, 0.0642]$, $p < 0.001$).

These results consistently demonstrate that the synergy between a classifier and a judge provides a statistically significant improvement in labelling reliability across all identity groups. This confirms that the proposed LLM-as-a-judge strategy effectively mitigates the limitations of single-model classification, justifying the additional computational cost of the pipeline. As a result, these outcomes strengthen the answer to RQ2, with our approach achieving a substantial strength of agreement with human annotators and improving Cohen's Kappa scores by 5% when compared to prompt-only approaches.

5.3 Application on SHSS Dataset

Although the MHS corpus was needed to evaluate the classification performance of the synthetic annotation strategies, one of the main contributions of this paper is the conversion of the SHSS dataset into a multi-class one. Hence, the proposed methodology has been applied to SHSS using the best-performing setting (minority-class assignment, Gemma4-31B as the classifier, and Qwen3.5-27B as the judge). From the initial 7265 hateful samples, the proposed methodology has classified a total of 6325 unique samples. Table 15 reports the distribution of labels in the multi-label dataset prior to unification. Notably, the total count of texts exceeds 6325, as individual observations may be associated with multiple categories. Therefore, the reported percentages represent the proportion of texts in which each label appears individually, rather than a mutually exclusive distribution across classes. Please note that a '*none*' class has been added, since 122 examples were not associated with any class and this classification was accepted by the judge.

Table 15: Class distribution of the multi-label version of the SHSS dataset.

Gender		Origin		Race		Disability		Sexuality		Religion		None		Age	
n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
3041	48.07	1929	30.50	696	11.00	595	9.41	559	8.84	249	3.94	122	1.93	121	1.91
Total samples														6325	

The multi-label distribution is visualized in Fig. 7a. To further examine the inter-label relationships, a co-occurrence heatmap is provided in Fig. 7b, illustrating the frequency with which pairs of labels appear within the same instances.

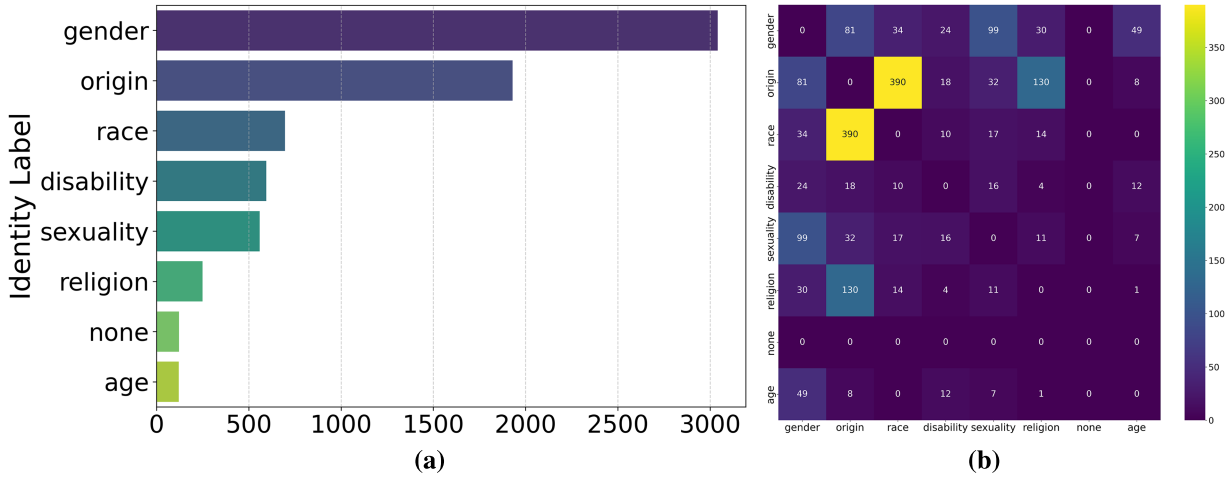

Figure 7: Analysis of the multi-label version of the SHSS dataset. (**a**) shows the class distribution, while (**b**) presents the relationships between labels via a co-occurrence heatmap.

After obtaining the multi-label dataset, the minority-class assignment was used as the label reduction strategy, as it has been shown to provide the best results. Table 16 and Fig. 8 present the class distribution of the multi-class variant after applying the minority-class assignment unification strategy.

Table 16: Class distribution of the multi-class version of the SHSS dataset.

Gender		Origin		Race		Disability		Sexuality		Religion		None		Age	
n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
2724	43.07	1351	21.36	525	8.30	681	10.77	553	8.74	248	3.92	122	1.93	121	1.91
Total samples														6325	

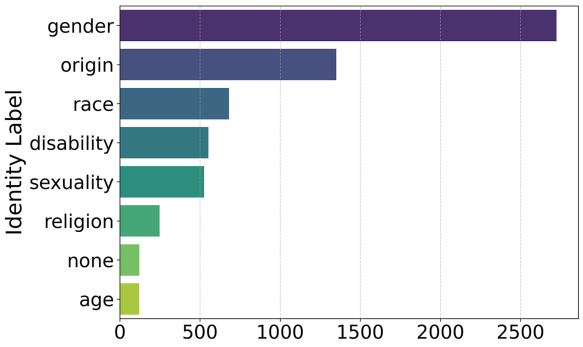

Figure 8: Class distribution of the unified multi-class version of the SHSS dataset.

5.3.1 Evaluation of a SHSS Subset on Human-Annotated Data

To validate the annotation performance on the SHSS dataset, a manual review of 100 instances was conducted, as no multi-class annotated data were available. These 100 samples were extracted in a stratified manner from the multi-class version of the SHSS, and each sample was manually annotated into one of the following categories: gender, origin, race, disability, sexuality, religion, age, and none. The latter category was added since, during annotation, some hateful samples did not target any of the identity groups in question (See Qualitative Error Analysis section below for more details). In addition, the multi-label synthetic annotations were also preserved to conduct a deeper analysis.

The evaluation compared the performance of the pipeline before and after the minority-label unification process. As shown in Table 17, before unifying labels, higher match and individual match rates were obtained compared to the post-unification phase (81% vs.70%). However, a lower rate for the full match metric was found in the multi-label data (58% vs. 70%).

Per-class individual match metrics are outlined in Table 18. As expected, the label unification process led to a slight decrease in performance, particularly for the most highly represented classes, such as origin.

To further assess the reliability of the synthetic labels, the annotator agreement between human and synthetic labels was measured using per-class Cohen's Kappa scores (see Table 19). The results indicate an overall substantial strength of agreement, with mean Kappa scores greater than 0.61. Specifically, *disability*, *age*, and *sexuality* exhibited the highest levels of agreement.

Table 17: Evaluation results for the Gemma4-31B/Qwen3.5-27B LLM-as-a-judge approach on a 100-sample subset of the SHSS dataset.

Metric	SHSS			
	Pre-Unification		Post-Unification	
	n	%	n	%
Matches				
Full Match	58	58%	70	70%
Match	81	81%	70	70%
Individual Match	81	81%	70	70%
Mismatches				
Full Mismatch	19	19%	30	30%
Individual Mismatch	19	19%	30	30%
Total Samples	100			

Table 18: Individual Match results for the manually annotated subset of the SHSS dataset before (Pre-Unif.) and after (Post-Unif.) the unification strategy.

	Age		Disability		Gender		Origin		Race		Religion		Sexuality		None	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Pre-Unif.	9	100	12	100	13	100	20	83.33	7	100	6	100	11	84.62	3	33.33
Post-Unif.	9	100	11	91.67	13	100	11	45.83	6	85.71	6	100	11	84.62	3	33.33
Total per class	9		12		20		24		7		6		13		9	

Table 19: Cohen's Kappa scores obtained using the Gemma4-31B/Qwen3.5-27B LLM-as-a-judge synthetic annotation on the SHSS multi-label (Pre-Unification) and multi-class (Post-Unification) manually annotated datasets.

Class	SHSS	
	Pre-Unification	Post-Unification
Age	0.84	0.84
Disability	0.95	0.91
Gender	0.52	0.72
Origin	0.73	0.54
Race	0.63	0.60
Religion	0.60	0.60
Sexuality	0.82	0.82
None	0.20	0.20
Mean	0.66	0.65

The similarity of these results to those obtained for Minority-MC-Hate-MHS (0.65 vs. 0.64) provides a clear answer to RQ5, confirming that the strategy validated on the English dataset generalizes effectively to the Spanish SHSS dataset, exhibiting consistent performance across both languages. Therefore, the proposed

strategy establishes a robust baseline for synthetic labelling in the absence of extensive human-annotated gold standards.

Qualitative Error Analysis

To provide deeper insights into the behaviour of the proposed architecture and examine the nature of the misclassifications, the manually annotated subset of the SHSS was also used to conduct a qualitative error analysis. This analysis includes incorrectly labelled samples, as well as samples that the system failed to label.

Regarding misclassifications, Fig. 9 displays the error heatmap, which reveals that the most frequent misclassifications occur between semantically related categories. Specifically, it is observed that there is a recurring confusion between *origin*, *race*, and *religion* (see first three samples of Table 20 for examples). In these cases, the non-unified labels always contain the human label. However, after the label reduction strategy, the category that is selected does not match human preferences. Therefore, these failure cases are more attributable to the transversal nature of these hate speech samples that comprise multiple target identity groups than to the classification performance. Furthermore, a significant number of targets were incorrectly labelled as *none*, particularly in the *gender* category (see the last two samples of Table 20 for examples). This suggests that the model struggles, in some cases, to identify some derogatory expressions based on gender.

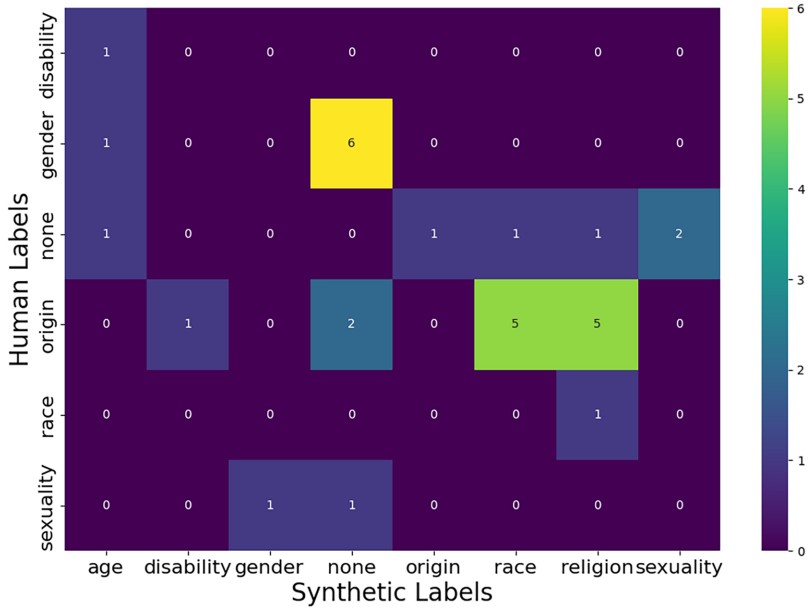

Figure 9: Misclassification heatmap of the manually annotated SHSS dataset after label unification.

Table 20: Examples of misclassification in the labelled SHSS subset. Offensive expressions have been masked using angle brackets to omit potentially sensitive or harmful content, ensuring neutrality and avoiding unintended ideological or political implications. A '-' on the predicted label column indicates that no label was assigned by the synthetic annotation system.

Text (in Spanish)	Human	Non-Unified	Unified
5 años llevan con esto. De las más de 200 manadas de <*derogatory expression based on origin*> y demás escoria desde esto, callados como buenas <*political reference*> que son.	Origin	Origin, race	Race

(Continued)

Table 20 (continued)

Text (in Spanish)	Human	Non-Unified	Unified
El gorrilla (aparcachoches) que es de toda la vida los multan y ahora viene un gorrilla <*origin reference*> y no pasa nada	Origin	Origin, race	Race
Prefiero una mujer <*race reference*> como protagonista y heroína. . .que a una <*race reference*> o <*religious reference*> que no inspiran no jalan no venden no llaman	Race	Race, religion	Religion
Cuando veais a los pobres <*origin reference*>, acordaron que ademas de sangraros gracias a las ayudas y subvenciones que nuestros mayores no reciben, se llevan el dinero a traves de <*religious reference*> y locutorios...	Origin	Origin, religion	Religion
<*gender-based derogatory expression*>, que es lo que te hace falta, <*gender-based insult*>	Gender	–	–
Dice Andres que te calles maldita <*derogatory expression based on gender*>	Gender	–	–

Concerning samples that the judge-based approach failed to label, these can be categorised into two separate groups: (1) samples that were discarded due to disagreement loops, where the classifier and the judge failed to reach a consensus, and (2) samples in which the generation output was impossible to parse.

Qualitative analysis of the first category reveals that these loops frequently occurred in instances of 'out-of-scope' hate speech, specifically, attacks based on political affiliation, which is not one of the defined identity groups (see Table 21). In these cases, the classifier correctly identified that the content was hateful but appropriately assigned no label based on the provided category list. However, the judge often flagged these empty classifications as failures, arguing that the omission of a label ignored the clear hateful intent of the text. This creates a paradox where the classifier adheres to the categorical constraints while the judge penalises the lack of a label despite the absence of a valid target. This conservative behaviour, while reducing the total number of labelled samples, prevents the introduction of forced or noisy labels into the final dataset, ensuring that only instances with a clear, defined identity target are preserved.

Table 21: Examples of instances discarded due to disagreement loops extracted from the SHSS dataset annotated using the Gemma4-31B/Qwen3.5-27B setting. Offensive expressions have been masked using angle brackets to omit potentially sensitive or harmful content, ensuring neutrality and avoiding unintended ideological or political implications.

Text (In Spanish)	Predicted Label	Justification (Classification Model)	Feedback (Judge Model)
@USER Molecular? Pero si los de <*political reference*> son gente re ignorante, re cavernicolas, para qué ocupan terminologías q no entienden? XD	–	The text targets individuals based on their political affiliation. However, political affiliation is not included in the provided list of defined hate speech categories (race, religion, origin, gender, sexuality, age, disability). Therefore, none of the defined categories apply to this text.	The input provided does not contain a proposed classification to evaluate. The instructions explicitly require assessing the accuracy of a specific one or two-category classification against the text. Since no classification was supplied in the 'Clasificación' field, it is impossible to determine if the labels accurately reflect the hate speech targets present in the text.

(Continued)

Table 21 (continued)

Text (In Spanish)	Predicted Label	Justification (Classification Model)	Feedback (Judge Model)
@USER @USER Es imposible escuchar eso!!! Para ellos recursos = riqueza. Por ello son un grupo de monos ignorantes destructores de sociedades. La evidencia empírica lo demuestra #NoAl<*political reference*>	race	The text uses the dehumanizing term 'monos' (monkeys), which is a historically rooted racial slur used to degrade individuals based on perceived race or ethnicity. Even though the author includes a political hashtag (#No<*political reference*>), the specific choice of imagery ('monos') targets the perceived race or ethnicity of the group being described as 'ignorant' and 'destructors of societies', aligning it with the defined 'race' category.	The proposed classification 'race' is incorrect because the text targets a group based on their political ideology ('<*political reference*>') rather than their race, ethnicity, or skin color. The dehumanizing language ('monos') is directed at individuals holding specific political beliefs. Since political affiliation is not one of the defined hate speech categories provided (race, religion, origin, gender, sexuality, age, disability), and the text does not contain elements targeting any of the listed protected groups, the classification fails to accurately reflect the content according to the given definitions.
🚨🚨 Los okupas causan graves disturbios en Santiago y <*political reference*> les respalda LINK <*generic insult*>	–	The provided text contains derogatory language targeting individuals based on their housing status ('okupas') and political affiliation ('<*political reference*>'). However, none of the defined hate speech categories (race, religion, origin, gender, sexuality, age, disability) encompass political beliefs or social/housing status. According to the strict definitions provided, the text does not fit into any of the specified categories.	The proposed classification is empty, which fails to address the clear hateful intent present in the text. The text uses derogatory language ('<*generic insult*') and targets specific groups ('okupas' and '<*political reference*>'). Although the defined categories do not explicitly include political affiliation or housing status, an empty classification incorrectly implies that no hate speech is present or that no target exists. In the context of evaluating hate speech, failing to identify the targeted hostility results in an inaccurate assessment.

Note: A '–' in the predicted label column indicates that no label was assigned by the synthetic annotation system. The emoji represents an emergency light. On social media, it is generally used to announce something.

The second category of unlabelled samples, parsing errors, exhibited a strong correlation with political content. As illustrated in Table 22, several instances containing political rhetoric resulted in invalid JSON outputs, as the models generated new lines until the maximum context length was reached, failing to correctly close the JSON structure. This pattern suggests that when samples do not fit within the provided identity group taxonomy, the generation process becomes unstable, leading to malformed responses. Therefore, this failure should not be considered a systemic flaw, but rather a consequence of attempting to classify general hate speech into a restricted set of identity-based categories.

Table 22: Examples of non-classified instances in the SHSS dataset due to parsing errors.

Text (In Spanish)	Error
Las malditas <*political reference*> no saben perder porque todo lo obtienen por la fuerza o con la fuerza, o con corrupción, o con violencia, o con la propaganda del terror...l@USER maldit@USER nunca han obtenido nada con Dignidad, Decencia, Integridad, Transparencia. LINK	Invalid JSON when parsing {\n "response": {\n "classification": \n \n \n. . . \n Invalid JSON: EOF while parsing a value.

(Continued)

Table 22 (continued)

Text (In Spanish)	Error
Auto de delincuentes encontrado sector la florida intensos operativos para dar con delincuentes que asesinaron a la mujer policía que realizaba su trabajo hacia la comunidad basta de esto en <*country*> sigan apoyando políticos populistas y terminaremos peor LINK	Invalid JSON when parsing {\n "response": {\n "classification": \n \n \n. . . \n Invalid JSON: EOF while parsing a value.
Lo único q quiero es cambio de gobierno wn para q saquen a la etos ineptos que han manejado como el hoyo la pandemia LINK	Invalid JSON when parsing {\n "response": {\n "classification": \n \n \n. . . \n Invalid JSON: EOF while parsing a value.

Finally, instances where both humans and synthetic annotations agreed on the absence of any of the established identity groups were analysed. While these samples can be considered correct, they were not supposed to occur; however, during manual labelling we encountered samples that could not be classified into any of the identity groups. As shown in Table 23, these instances also contain hate speech related to politics.

Table 23: Examples of instances in the SHSS subset where both human and synthetic annotations agreed on the absence of any of the established identity groups. Offensive expressions have been masked using angle brackets to omit potentially sensitive or harmful content, ensuring neutrality and avoiding unintended ideological or political implications.

Text (In Spanish)
Y a mí qué me importa lo que tú "notes", <*political reference*>?
<*political reference*> siendo <*political reference*>
Lástima que, veinte años después, haya tenido que aparecer esa mugre <*political reference*> adoradora de <*political reference*> para poner en cuestión su figura

5.4 Cost Analysis

The computational cost of the synthetic annotation approaches was evaluated in terms of total execution time and average processing time per sample for the different model combinations (see Table 24).

The prompt-only baseline demonstrated the lowest computational overhead, with a total execution time of 1 h 29 m and an average of 0.66 s per sample. In contrast, the LLM-as-a-judge strategies naturally increased the processing time due to the sequential nature of the classifier-judge interactions and the possibility of multiple iterations per sample. Among the tested combinations, the combination of Gemma3-27B (classifier) and Llama3.1-70B (judge) was the most efficient iterative configuration, requiring 2 h 40 m in total (1.19 s/sample). The highest computational demand was observed in the Qwen3.5-27B/Gemma4-31B configuration, which took 5 h 8 m (2.28 s/sample). On average, the iterative approach increased the time per sample by a factor of approximately 2× to 3.5× compared to the prompt-only strategy.

Table 24: Computational cost of the synthetic annotation strategies.

Classifier	Judge	Time (hh:mm)	Avg. per Sample (s/Sample)
Gemma3-27B	Llama3.1-70B	02:40	1.19
Qwen3.5-27B	Gemma4-31B	05:08	2.28
Gemma4-31B	Qwen3.5-27B	04:02	1.80
Qwen3.5-27B	Llama3.1-70B	03:06	1.38
Llama3.1-70B	Qwen3.5-27B	03:29	1.55
Prompt-only		01:29	0.66

6 Discussion

The present study evaluated the efficacy of an LLM-as-a-judge synthetic annotation approach useful for translating binary-annotated datasets into multi-class ones. Our results demonstrate that integrating a judging mechanism into the annotation pipeline significantly enhances the reliability and accuracy of labels compared to traditional prompt-only approaches, achieving a substantial strength of agreement with human annotators. As a result, the proposed strategy proves to be a versatile tool for translating binary labelled datasets into multi-class ones.

6.1 Comparative Analysis: LLM-as-a-Judge vs. Prompt-Only

In contrast to current state-of-the-art approaches that rely on prompt-only strategies for generating synthetic labels [19,60–62], our research proposes a novel LLM-as-a-judge methodology for synthetic annotation. By leveraging a conversational loop between the classifier and judge models, our approach has shown to be more aligned with human annotations than prompt-only strategies.

Furthermore, our annotation performance assessment differs from prevalent strategies in the literature. While authors such as Kazemi et al. [61], Nahum et al. [62], and Horych et al. [60] evaluate the utility of synthetic labels through the performance of downstream classification models, this approach does not directly measure the quality of the annotations themselves. In those settings, final performance depends not only on label correctness, but also on the generalisability and biases of the downstream model. Consequently, our work evaluates synthetic labels directly against ground truth annotations, isolating annotation performance from downstream modelling effects.

Additional evaluation techniques were identified in the literature, although they were not applicable to the present study for different reasons. For instance, Parfenova et al. [19] employed human annotators to evaluate synthetic labelling performance. While this is a valid and reliable strategy, its scalability is significantly limited.

6.2 Multi-Label Performance

While the objective was to provide a multi-class classification, both the LLM-as-a-judge and prompt-only approaches were allowed to assign up to two labels to a given instance, yielding multi-label raw outputs that were further processed. These multi-label outputs, which serve as intermediate results, were also evaluated using Hate-MHS as the ground truth. The results obtained in the multi-label setting demonstrate that the proposed LLM-as-a-judge framework provides a measurable improvement over the prompt-only baseline in multi-label hate speech annotation. Specifically, for more than 94% of samples, at least one correct synthetic label was generated by the LLM-as-a-judge approach, compared to approximately 94% obtained by the prompt-only approach (values indicated by the *Match* metric). Furthermore, a higher percentage of

individual matches (77.24% vs. 76.75%) was also observed for the proposed approach. Accordingly, a smaller number of *mismatches* was found when applying the LLM-as-a-judge technique. It is important to note, however, that the LLM-as-a-judge approach labelled a smaller total number of samples (n = 8022) compared to the prompt-only baseline (n = 8089). This difference suggests that the judge-based system may be more selective or restrictive in its assignments, rejecting incorrect classifications. However, percentage rates were computed considering the full dataset (n = 8089), thereby penalising non-classified samples. Consequently, these results demonstrate that the proposed approach provides a modest improvement in the number of classifications where at least one label is aligned with human annotators (*Matches*). Moreover, this method was also shown to identify, in relative terms, a greater number of identity groups aligned with human annotators than the prompt-only approach (*Individual Matches*).

Additionally, it is important to highlight that *full match rates* are inherently penalised in this context. Although the generation of synthetic labels is constrained to a maximum of two identity groups, ground truth labels from Hate-MHS can comprise a greater number of groups. Consequently, this configuration imposes a structural ceiling for this metric.

A granular analysis of identity groups reveals that the advantage of the LLM-as-a-judge approach, with the initial configuration (Llama3.1-70B/Gemma3-27B), is selective rather than general. While it demonstrated a better performance for most identity groups (five out of seven), a clear bias was evident in the *gender* and *origin* classes, where the prompt-only approach consistently outperformed it. This discrepancy is particularly significant given that *gender* and *origin* are among the most critical categories in hate speech classification. The lower performance of the judge-based system on these two classes may be linked to a plausible bias from $\mathcal{G}_{judge}$ towards other identity groups.

The inter-annotator agreement, measured by Cohen's Kappa, further clarifies these dynamics. The most aligned class with human annotations, regardless of the annotation approach, is religion. This tendency may be attributed to the clear independence between religion and other identity groups, a situation that is not observed between other groups, such as origin or race, which frequently overlap. In general, the LLM-as-a-judge approach offers slightly better results, achieving a significant 1% increase on average compared to the prompt-only baseline and demonstrating a moderate strength of agreement with human annotators. However, as these results were obtained using the initial model configuration (Llama3.1-70B/Gemma3-27B), they serve as a baseline for the subsequent ablation study (Section 6.4), which explores whether different model combinations can further amplify the advantages of the judge-based framework.

6.3 Impact of Label Reduction Strategies

The transition from multi-label to multi-class employing the label reduction strategies noticeably degrades performance. As expected, the move to a stricter single-label framework led to a decrease in *Match* and *Individual Match* metrics and a corresponding increase in mismatches. In the multi-label setting (Hate-MHS), a *match* is recorded if at least one synthetic label appears in the ground truth; in the multi-class settings (minority, majority, and random) only a *full match* is possible. Analogously, while each sample in the multi-label setting could contribute up to two *individual matches*, in multi-class settings only one is possible. Therefore, although the number of *Full Matches* increases after label reduction, the number of *Matches* and *Individual Matches* decreases substantially.

An analogous effect was observed for mismatch metrics. In the multi-class setting, *Full Mismatch* and *Individual Mismatch* are equivalent, since each instance contains only one identity group. Consequently, any misclassification simultaneously counts as both *Full* and *Individual Mismatch*. This led to a marked increase in the number of mismatches under the multi-class formulation. In this scenario, the LLM-as-a-judge approach yielded slightly better results than the prompt-only strategy, although the advantage was

modest. Nevertheless, such a significant increase occurred that it surpasses *Individual Mismatches* observed with multi-label data. This indicates that, regardless of the annotation approach, the application of the label-reduction strategy results in a higher overall mismatch rate.

Consequently, the application of whatever label-reduction strategy will have a detrimental effect on performance. In our case, we evaluated three distinct reduction strategies: minority-class assignment (Minority-MC-Hate-MHS), majority-class assignment (Majority-MC-Hate-MHS), and random-class assignment (Random-MC-Hate-MHS).

The majority-class assignment yielded the highest overall match rates (approximately 70% for both the LLM-as-a-judge and Prompt-only approaches). However, this strategy introduced a significant imbalance in the resulting dataset, overwhelmingly favoring the most frequent identity groups (e.g., *race* and *gender*) and substantially reducing the representation of minority groups such as *age* and *religion*. This suggests that while majority-class assignment maximizes immediate match metrics, it does so by reinforcing the dominant classes, which could lead to biased downstream models.

Conversely, the random-class assignment resulted in the lowest overall performance, with match rates dropping to approximately 58% and a mean Cohen's Kappa of approximately 0.44. The lower match rates and agreement scores indicate that random selection is an unreliable strategy for label unification. However, these results represent a neutral baseline that provides context for the other unification strategies.

The minority-class assignment (Minority-MC-Hate-MHS), while showing a slightly lower overall match rate (62.96% for the Judge) compared to the majority approach, provided the most balanced distribution across identity groups. Most importantly, it achieved the highest mean Cohen's Kappa score ($\kappa = 0.54$ for the Judge), indicating a superior strength of agreement with human annotators when considering the dataset as a whole. In this setting, the LLM-as-a-judge approach outperformed the prompt-only baseline in five out of seven classes (*age*, *disability*, *origin*, *race*, and *religion*), demonstrating that the proposed synthetic annotation strategy is particularly effective at identifying minority identity groups that are often overlooked by simpler prompting strategies.

These results justify the adoption of the minority-class assignment in our final pipeline. By prioritising the least represented classes, we not only maintain a moderate strength of agreement with the human ground truth but also mitigate class imbalance, ensuring a more equitable distribution of identity groups for future model training.

Finally, it is important to understand behavioural consequences resulting from the application of label reduction strategies. For example, when using the initial configuration (Llama3.1-70B/Gemma3-27B) the gain achieved by the LLM-as-a-judge approach is statistically significant in the multi-label context, whereas no such significance is observed in the multi-class setting. This difference is attributed to the label reduction strategy. In the multi-label setting, partial agreements—where some, but not all, labels are correctly predicted—facilitate the emergence of more nuanced differences between models. In contrast, the multi-class formulation collapses multiple co-occurring labels into a single class, thereby removing information about label combinations and interdependencies, while simplifying the prediction space. Nevertheless, the distribution of errors is altered because partial agreements are suppressed when shifting from multi-label to multi-class. As a result, performance differences that are detectable in the multi-label scenario may become attenuated or statistically indistinguishable after label reduction. This behaviour is consistently observed throghout the study, obtaining worsewith poorer results occurring whenever multi-class performance is compared to multi-label.

6.4 Ablation Study

To ensure that the observed improvements were not the result of a specific model's bias or a particular size configuration, and to investigate whether the modest gains could be further enhanced, we conducted an ablation study testing various combinations of LLM families (Llama, Gemma, and Qwen) and sizes (ranging from 27B to 70B).

The results (Tables 9–11) demonstrate that the iterative framework is robust and model-agnostic. Interestingly, the combination of Gemma4-31B (classifier) and Qwen3.5-27B (judge) achieved the highest match rate (72.32%) and the highest mean Cohen's Kappa score ($\kappa = 0.64$), outperforming the initial Llama3-70B/Gemma3-27B setup by 10 percentage points ($\kappa = 0.54$). Moreover, combinations employing Qwen3.5-27B as the classifier with either Llama3-70B or Gemma4 as the judge also provided remarkable results, achieving the second-best Kappa scores ($\kappa = 0.62$). While the Qwen3.5-27B/Llama3.1-70B configuration performed similarly to the Gemma4-31B/Qwen3.5-27B setting in terms of matches (70.50% vs. 72.32%), the Qwen3.5-27B/Gemma4-31B pair obtained lower results (66.34% vs.72.32%), likely due to a noticeable decrease in the number of labelled samples.

Regarding size asymmetry, the results suggest that it is preferable to use similar-sized models or to employ the largest model as the judge. For example, when using Gemma3-27B as the judge with Llama3-70B as the classifier, there were almost no rejections (99.17% of samples were labelled), yet this resulted in the lowest match percentage. Similarly, one of the worst match rates was obtained when Qwen3.5-27B served as the judge for Llama3.1-70B. However, this is not attributable to poor judging performance by Qwen3.5-27B, as the overall best results were achieved when Qwen3.5-27B acted as the judge for Gemma4-31B. These findings suggest that Gemma4-31B exhibits superior classification performance compared to the other models, but is highly restrictive when acting as a judge. Conversely, Qwen3.5-27B demonstrates superior judging performance alongside valuable classification abilities, while Llama3.1-70B shows modest classification performance but a high-quality judging criterion. Moreover, these results suggest that the 'judging ability' of a model is a capability distinct from its 'classification power'.

The selection of the optimal model combination yielded definitive results regarding the effectiveness of the LLM-as-a-judge approach compared to the prompt-only technique. Using the optimal configurations, the combination of the Gemma4-31B/Qwen3.5-27B significantly outperformed the results achieved by these two models as standalone classifiers within the prompt-only approach. This superior performance was consistent across metrics, achieving better results in both matches and Cohen's Kappa scores. Consequently, this provides empirical justification for the claim that the interaction between a classifier and a judge effectively filters noise and refines the final labels.

These findings are further supported by the evaluation performed on the manually annotated SHSS subsample. The results obtained for this subsample strengthen the validity of the proposed evaluation approach, as indicated by the proximity of the Cohen's Kappa scores (0.65 vs. 0.64). Moreover, such a level of agreement denotes the exceptional cross-lingual capabilities of state-of-the-art LLMs and the generalisability of the proposed approach. Finally, qualitative analysis of the classification performance sheds light on the internal behaviour of the architecture, revealing that errors might be largely attributable to the nuanced nature of hate speech and the presence of political targets outside the defined identity groups taxonomy. Therefore, the judge-based approach acts as a conservative filter, effectively preventing the introduction of noisy labels by rejecting ambiguous or out-of-scope instances; although this leads to a reduction in the total number of labelled samples, it significantly enhances the overall reliability and purity of the final multi-class dataset.

6.5 Limitations and Generalizability

The improved classification performance and robustness of the LLM-as-a-judge approach come at the expense of a measurable increase in computational cost. Utilising two LLMs instead of one inherently increases computational demands, and the iterative evaluation process between $\mathcal{G}_{classifier}$ and $\mathcal{G}_{judge}$ further extends processing time. As shown in our results, this approach increases the processing time per sample by approximately 2 to 3.5 times compared to the prompt-only baseline. This overhead is primarily driven by the sequential dependency between the classifier and the judge, as well as the potential for multiple iterations (up to 10) to resolve mismatches.

However, we argue that this increase in latency is negligible within the context of synthetic dataset annotation for high-stakes tasks, such as hate speech detection. Unlike real-time inference applications, where user interaction is crucial, synthetic annotation is an offline, one-time process. The total execution time of the most demanding configuration was only 5 h 8 m for the entire dataset, a duration that is entirely acceptable given the gains in label quality, as demonstrated by a superior agreement with human annotators (Mean Kappa κ = 0.64). Furthermore, the ablation study reveals that this cost can be further optimised by selecting specific model pairs. For instance, the Qwen3.5 (classifier) and Llama3.1 (judge) configuration offers a highly competitive balance between speed (1.38 s per sample) and performance.

The number of classified samples is another nuanced yet critical consideration. The judge-based approach is inherently more selective, resulting in a smaller number of labelled samples in the final dataset. However, as shown in the qualitative analysis, the primary reason for the non-classified instances is the presence of political targets outside the defined identity groups taxonomy. Conversely, the prompt-only methodology labelled a higher number of samples, as it lacks the judging step and prevents any possibility of disagreement. The same pattern was observed when labelling the SHSS, where the LLM-as-a-judge technique labelled 6325 out of 7265 hateful samples.

Despite these limitations, a novel strategy for converting binary-labelled datasets into multi-class ones has been developed. This technique has proven to outperform prompt-only approaches and, while it has been applied within the hate speech domain, it possesses the potential for broader applicability and could be extended to any field requiring multi-class labelled datasets. Given any binary-labelled dataset and its corresponding class definitions, the prompts can be updated with the new definitions, thereby adapting the proposed strategy to the target domain. Moreover, the cross-lingual efficacy of the proposed approach has been demonstrated between English and Spanish, further broadening its generalisation capabilities of the proposed approach.

7 Conclusion

The proliferation of hate speech in recent years underscores the need for effective mitigation strategies within online environments. The present study addresses this need by proposing a novel methodology for synthetically converting binary hate speech datasets, which are considerably more common, into multi-class datasets, the utility of which has been highlighted by the research community [56,57]. By leveraging an LLM as a classifier in conjunction with a separate LLM functioning as a judge, our approach has successfully transformed binary classification datasets into more nuanced, multi-class categorisations.

Our strategy has exhibited superior performance and higher fidelity relative to human labels than prompt-only classifications, yielding a 5% increase in the Kappa score against the prompt-only baseline. Notably, the methodology demonstrated robust cross-lingual capabilities, achieving results for the Spanish subset similar to those obtained for English data. These results provide a positive answer to our research questions, showing that the integration of the LLM-as-a-judge paradigm is not only feasible but also

significantly beneficial. Furthermore, our findings reveal that a larger classifier does not automatically guarantee better results; instead, we observed that optimal performance is achieved when using similar-sized models or when the model used as the judge exhibits high judging performance, regardless of whether it is smaller than the classifier. Specifically, the significant synergy between Gemma4-31B (classifier) and Qwen3.5-27B (judge) outperformed larger configurations, suggesting that 'judging ability' is a distinct capability from 'classification power', and that different model architectures may possess different sensitivities to specific identity markers.

While these improvements come at the expense of computational efficiency, given the offline nature of dataset annotation and the importance of high-quality annotations, the observed computational costs are justifiable. Moreover, although the LLM-as-a-judge approach labels fewer samples, it prevents the inclusion of noisy data. Nevertheless, the use of this methodology should not be universal. Applying the proposed approach should be carefully considered in scenarios where accuracy does not outweigh temporal, financial, or computational demands. In those cases, researchers should be aware of the sub-optimal nature of relying on prompt-only techniques for synthetic labelling.

Furthermore, it is important to note that the proposed methodology possesses broader applicability and holds promise for enhancing classifications in other binary problem domains requiring more granular analytical resolution. Future research could build upon this by, for example, using the proposed strategy to refine the CWE categorisation in code vulnerability detection datasets, or to disaggregate the typology of spam messages. In addition, future work could investigate different LLM-as-a-judge strategies to optimise either classification accuracy or processing speed, explore other prompt-only strategies such as Chain of Thought [86], or expand the target identity groups to reduce the number of non-classified samples.

In conclusion, the methodology presented in this study constitutes a significant advancement in multi-class text classification, offering a viable pathway for future machine learning research to generate large-scale, granular, multi-class datasets for model training. Through this strategy, binary-labelled datasets can be transformed into multi-class labelled ones, providing significantly higher alignment with human annotators than prompt-only approaches and addressing the scarcity of multi-class datasets. As a result, the proposed technique demonstrably mitigates the challenges associated with data scarcity in multi-class classification tasks.

Acknowledgement: The authors want to thank "VII Convocatoria de Ayudas a la Investigación" from RTVE and the support provided by the INTELIA (Interaction Technologies and Artificial Intelligence) lab.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Antonio Moreno-Cediel; methodology, Antonio Moreno-Cediel, Antonio Garcia-Cabot and Eva Garcia-Lopez; software, Antonio Moreno-Cediel; validation, Antonio Moreno-Cediel, Antonio Garcia-Cabot and Eva Garcia-Lopez; formal analysis, Eva Garcia-Lopez; investigation, Antonio Moreno-Cediel, Antonio Garcia-Cabot and Eva Garcia-Lopez; data curation, Antonio Moreno-Cediel; writing—original draft preparation, Antonio Moreno-Cediel, Antonio Garcia-Cabot and Eva Garcia-Lopez; supervision, Antonio Garcia-Cabot and Eva Garcia-Lopez; funding acquisition, Antonio Garcia-Cabot. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data and the code that support the findings of this study are available from the Corresponding Author, Eva Garcia-Lopez, upon reasonable request.

Ethics Approval: not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Prompt Templates

[System]
You are a highly skilled and meticulous expert in hate speech detection and classification. Your task is to analyze the text provided and categorize it according to the defined hate speech categories, providing a clear and concise justification of your classification.

Defined Hate Speech Categories:
- race: speech targeting individuals or groups based on perceived race or ethnicity, including but not limited to: Black or African American, Latino or non-white Hispanic, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic white.
- religion: speech targeting individuals or groups based on religious beliefs or affiliation, including but not limited to: Jews, Christians, Buddhists, Hindus, Mormons, Atheists, Muslims.
- origin: speech targeting individuals or groups based on national origin, citizenship status or immigration status, including references to specific countries, immigrants, migrant workers, or undocumented persons.
- gender: speech targeting individuals or groups based on gender identity (Women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)).
- sexuality: speech targeting individuals or groups based on sexual orientation (Bisexual, gay, lesbian, heterosexual).
- age: speech targeting individuals or groups based on age (children, teenagers, adults, middle-aged, seniors).
- disability: speech targeting individuals or groups based on disability status, including physical, learning, cognitive, mental health, or sensory problems.

Instructions:
1. Classification Limit: You must classify the text into one category, and only if absolutely necessary into two categories. Prioritize the most direct target of the hatred if multiple categories apply.
2. Objective Analysis: Base your classification solely on the content of the text and the definitions provided above. Avoid making assumptions.
3. Justification is crucial: Your justification must specifically explain how the text aligns with the chosen category(ies).
4. Error Awareness: Your performance will be evaluated by an expert. Be prepared to revise your classification, reflect on your error and modify the assigned category.

Output Format (JSON):
```
{
   "classification": "<hate speech category(ies)>",
   "justification": "<justification>"
}
```
[User]
<Text>

Figure A1: $\mathcal{G}_{classifier}$ prompt.

[System]

You are a highly skilled evaluator specialized in hate speech classification. Your task is to assess the accuracy of a given text classification against the provided category definitions. You will be presented with a text and a proposed classification (one or two categories). Your job is to determine if the proposed classification is correct.

Defined Hate Speech Categories:

- race: speech targeting individuals or groups based on perceived race or ethnicity, including but not limited to: Black or African American, Latino or non-white Hispanic, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic white.
- religion: speech targeting individuals or groups based on religious beliefs or affiliation, including but not limited to: Jews, Christians, Buddhists, Hindus, Mormons, Atheists, Muslims.
- origin: speech targeting individuals or groups based on national origin, citizenship status or immigration status, including references to specific countries, immigrants, migrant workers, or undocumented persons.
- gender: speech targeting individuals or groups based on gender identity (Women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)).
- sexuality: speech targeting individuals or groups based on sexual orientation (Bisexual, gay, lesbian, heterosexual).
- age: speech targeting individuals or groups based on age (children, teenagers, adults, middle-aged, seniors).
- disability: speech targeting individuals or groups based on disability status, including physical, learning, cognitive, mental health, or sensory problems.

Instructions:

1. Classification Limit: A valid classification must contain one or two categories. Classifications with more than two categories should be immediately rejected.
2. Objective Evaluation: Base your evaluation solely on the definitions provided and the content of the text. Do not introduce external knowledge or assumptions.
3. Evaluation Output: Assign either "accept" or "reject" to the "evaluation" field in the JSON output.
   - "accept": If the proposed classification accurately reflects the content of the text based on the provided definitions.
   - "reject": If the proposed classification is inaccurate or does not fully capture the hateful intent of the text based on the provided definitions.
4. Justification is crucial: Provide a clear and concise justification for your evaluation. Your justification should explain why the classification is accepted or rejected. Do not suggest what the correct classification should be; simply explain why the given classification is valid or invalid.
5. Prioritization: If multiple categories could potentially apply, and the proposed classification only includes one or two, evaluate whether those choices represent the most representative targets of the hate speech. If multiple categories could potentially apply, and the proposed classification only includes one or two, evaluate whether those choices represent the *most* salient targets of the hate speech.

Output Format (JSON):

```
{
  "evaluation": "<reject | accept>",
  "justification": "<justification>"
}
```

[User]

<Classification Outcome from $\mathcal{G}_{classifier}$>

Figure A2: $\mathcal{G}_{judge}$ prompt.

[System]
You are a highly skilled and meticulous expert in hate speech detection and classification. Your task is to analyze the text provided and categorize it according to the defined hate speech categories, providing a clear and concise justification of your classification.

Defined Hate Speech Categories:
- race: speech targeting individuals or groups based on perceived race or ethnicity, including but not limited to: Black or African American, Latino or non-white Hispanic, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic white.
- religion: speech targeting individuals or groups based on religious beliefs or affiliation, including but not limited to: Jews, Christians, Buddhists, Hindus, Mormons, Atheists, Muslims.
- origin: speech targeting individuals or groups based on national origin, citizenship status or immigration status, including references to specific countries, immigrants, migrant workers, or undocumented persons.
- gender: speech targeting individuals or groups based on gender identity (Women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)).
- sexuality: speech targeting individuals or groups based on sexual orientation (Bisexual, gay, lesbian, heterosexual).
- age: speech targeting individuals or groups based on age (children, teenagers, adults, middle-aged, seniors).
- disability: speech targeting individuals or groups based on disability status, including physical, learning, cognitive, mental health, or sensory problems.

Instructions:
1. Classification Limit: You must classify the text into one category, and only if absolutely necessary into two categories. Prioritize the most direct target of the hatred if multiple categories apply.
2. Objective Analysis: Base your classification solely on the content of the text and the definitions provided above. Avoid making assumptions.
3. Justification is crucial: Your justification must specifically explain how the text aligns with the chosen category(ies).

Output Format (JSON):
```
{
   "classification": "<hate speech category(ies)>",
   "justification": "<justification>"
}
```
[User]
<text>

Figure A3: Prompt-only approach prompt.

References

1. Matsuda MJ. Public response to racist speech: considering the Victim's story. Mich Law Rev. 1989;87(8):2320–81.
2. Brown A. What is hate speech? Part 1: the myth of hate. Law Philos. 2017;36(4):419–68. doi:10.1007/s10982-017-9297-1.
3. Council of Europe. Recommendation No. R (97) 20 of the committee of ministers to member states on 'Hate Speech'. In: Committee of Ministers on 30 October 1997 at the 607th Meeting of the Ministers' Deputies. Strasbourg Cedex, France: Council of Europe; 1779. p. 106–8.
4. European Union Council. Council framework decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. Off J Eur Union. 2008;328:55–8.
5. United Nations. United Nations strategy and plan of action on hate speech: detailed guidance on implementation for United Nations field presences. New York, NY, USA: United Nations; 2020. 52 p.
6. Chetty N, Alathur S. Hate speech review in the context of online social networks. Aggress Violent Behav. 2018;40(8):108–18. doi:10.1016/j.avb.2018.05.003.
7. Chanda S, Dhaka A, Pal S. Towards safer online spaces: deep learning for hate speech detection in code-mixed social media conversations. In: Proceedings of the Companion Publication of the 16th ACM Web Science Conference; 2024 May 21–24; Stuttgart, Germany. New York, NY, USA: Association for Computing Machinery; 2024. p. 103–9. doi:10.1145/3630744.3663610.
8. Das M, Mathew B, Saha P, Goyal P, Mukherjee A. Hate speech in online social media. SIGWEB Newsl. 2020;2020:1–8. doi:10.1145/3427478.3427482.
9. Maarouf A, Pröllochs N, Feuerriegel S. The virality of hate speech on social media. Proc ACM Hum Comput Interact. 2024;8(CSCW1):1–22. doi:10.1145/3641025.
10. Tontodimamma A, Nissi E, Sarra A, Fontanella L. Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics. 2021;126(1):157–79. doi:10.1007/s11192-020-03737-6.
11. Paz MA, Montero-Díaz J, Moreno-Delgado A. Hate speech: a systematized review. Sage Open. 2020;10(4):2158244020973022. doi:10.1177/2158244020973022.
12. Castaño-Pulgarín SA, Suárez-Betancur N, Vega LMT, López HMH. Internet, social media and online hate speech. Systematic Review Aggress Violent Behav. 2021;58(6):101608. doi:10.1016/j.avb.2021.101608.
13. Persily N, Tucker JA. Social media and democracy: the state of the field, prospects for reform. Cambridge, UK: Cambridge University Press; 2020. doi:10.1017/9781108890960.
14. Anjum Katarya R. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. Int J Inf Secur. 2024;23(1):577–608. doi:10.1007/s10207-023-00755-2.
15. Vaswani A. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
16. Baumann J, Röttger P, Urman A, Wendsjö A, Plaza-del-Arco FM, Gruber JB, et al. Large language model hacking: quantifying the hidden risks of using LLMs for text annotation. arXiv:2509.08825. 2025.
17. Alizadeh M, Kubli M, Samei Z, Dehghani S, Zahedivafa M, Bermeo JD, et al. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. J Comput Soc Sci. 2024;8(1):17. doi:10.1007/s42001-024-00345-9.
18. Shi L, Giunchiglia F, Luo R, Shi D, Song R, Diao X, et al. An empirical study of LLMs via in-context learning for stance classification. Inf Process Manag. 2026;63(1):104322. doi:10.1016/j.ipm.2025.104322.
19. Parfenova A, Marfurt A, Pfeffer J, Denzler A. Text annotation via inductive coding: comparing human experts to LLMs in qualitative data analysis. In: Chiruzzo L, Ritter A, Wang L, editors. Findings of the Association for Computational Linguistics: NAACL 2025. Kerrville, TX, USA: Association for Computational Linguistics; 2025. p. 6471–84.
20. Li D, Jiang B, Huang L, Beigi A, Zhao C, Tan Z, et al. From generation to judgment: opportunities and challenges of LLM-as-a-judge. arXiv:2411.16594. 2024.
21. Niu T, Joty S, Liu Y, Xiong C, Zhou Y, Yavuz S. JudgeRank: leveraging large language models for reasoning-intensive reranking. arXiv:2411.00142. 2024.

22. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning—based text classification: a comprehensive review. ACM Comput Surv. 2021;54(3):1–40. doi:10.1145/3439726.
23. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, et al. A survey on text classification: from traditional to deep learning. ACM Trans Intell Syst Technol. 2022;13(2):1–41. doi:10.1145/3495162.
24. Fields J, Chovanec K, Madiraju P. A survey of text classification with transformers: how wide? how large? how long? how accurate? how expensive? how safe? IEEE Access. 2024;12:6518–31. doi:10.1109/ACCESS.2024.3349952.
25. Scott S, Matwin S. Text classification using WordNet hypernyms. In: Usage of WordNet in Natural Language Processing Systems. Kerrville, TX, USA: Association for Computational Linguistics; 1998. p. 45–51.
26. McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification [Internet]. 1998 [cited 2026 May 29]. Available from: https://aaai.org/papers/041-ws98-05-007/.
27. Joachims T. A statistical learning learning model of text classification for support vector machines. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery; 2001. p. 128–36. doi:10.1145/383952.383974.
28. Soucy P, Mineau GW. A simple KNN algorithm for text categorization. In: Proceedings of the 2001 IEEE International Conference on Data Mining; 2001 Nov 29–Dec 2; San Jose, CA, USA. p. 647–8. doi:10.1109/ICDM.2001.989592.
29. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning. arXiv:1605.05101. 2016.
30. Elman JL. Finding structure in time. Cogn Sci. 1990;14(2):179–211. doi:10.1207/s15516709cog1402_1.
31. Kim Y. Convolutional neural networks for sentence classification. In: Moschitti A, Pang B, Daelemans W, editors. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Kerrville, TX, USA: Association for Computational Linguistics; 2014. p. 1746–51. doi:10.3115/v1/D14-1181.
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 (Long and Short Papers). Kerrville, TX, USA: Association for Computational Linguistics; 2019. p. 4171–86. doi:10.18653/v1/N19-1423.
33. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv:1908.10084. 2019.
34. Zhang J, Huang Y, Liu S, Gao Y, Hu X. Do BERT-like bidirectional models still perform better on text classification in the era of LLMs. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Findings of the Association for Computational Linguistics: EMNLP 2025. Kerrville, TX, USA: Association for Computational Linguistics; 2025. p. 18980–9. doi:10.18653/v1/2025.findings-emnlp.1033.
35. Matarazzo A, Torlone R. A survey on large language models with some insights on their capabilities and limitations. arXiv:2501.04040. 2025.
36. Zhang H, Xu B, Xiao S, Zhang C, Ji L. Zero- and few-shot Chinese cybersecurity event detection via meta-distillation learning. Inf Process Manag. 2026;63(1):104344. doi:10.1016/j.ipm.2025.104344.
37. Geetanjali Kumar M. Exploring hate speech detection: challenges, resources, current research and future directions. Multimed Tools Appl. 2025;84(31):38423–59. doi:10.1007/s11042-025-20716-2.
38. Malik JS, Qiao H, Pang G, van den Hengel A. Deep learning for hate speech detection: a comparative study. Int J Data Sci Anal. 2025;20(4):3053–68. doi:10.1007/s41060-024-00650-6.
39. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators; arXiv:2003.10555. 2020.
40. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv:1909.11942. 2019.
41. Abusaqer M, Saquer J, Shatnawi H. Efficient hate speech detection: evaluating 38 models from traditional methods to transformers. In: Proceedings of the 2025 ACM Southeast Conference; 2025 Apr 24–26; Cape Girardeau, MO, USA. New York, NY, USA: The Association for Computing Machinery (ACM); 2025. p. 203–14. doi:10.1145/3696673.3723061.

42. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.
43. Fetahi E, Susuri A, Hamiti M, Kastrati Z, Canhasi E, Misini A. Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. Soc Netw Anal Min. 2025;15(1):82. doi:10.1007/s13278-025-01497-w.
44. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Kerrville, TX, USA: Association for Computational Linguistics; 2020. p. 8440–51. doi:10.18653/v1/2020.acl-main.747.
45. More P, Gangurde P, Shinkar A, Mathur JN, Patil S, Borate V. Identifying political hate speech using transformer-based approach. In: Proceedings of the 2025 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI); 2025 Apr 23–25; Chennai, India. p. 1–6. doi:10.1109/RAEEUCCI63961.2025.11048250.
46. Angger Saputra R, Sibaroni Y. Multilabel hate speech classification in Indonesian political discourse on X using combined deep learning models with considering sentence length. J Ilmu Komputer Dan Informasi. 2025;18(1):113–25. doi:10.21609/jiki.v18i1.1440.
47. Chapagain S, Hamdi SM, Boubrahimi SF. Advancing hate speech detection with transformers: insights from the MetaHate. In: Karampelas P, Day MY, Ting IH, Alhajj R, editors. Advances in social networks analysis and mining. Cham, Switzerland: Springer Nature; 2025. p. 432–9. doi:10.1007/978-3-032-13509-4_32.
48. Albladi A, Islam M, Das A, Bigonah M, Zhang Z, Jamshidi F, et al. Hate speech detection using large language models: a comprehensive review. IEEE Access. 2025;13(6):20871–92. doi:10.1109/ACCESS.2025.3532397.
49. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv:2005.14165. 2020.
50. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, et al. SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In: May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM, editors. Proceedings of the 13th International Workshop on Semantic Evaluation. Kerrville, TX, USA: Association for Computational Linguistics; 2019. p. 54–63. doi:10.18653/v1/S19-2007.
51. Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, et al. A large labeled corpus for online harassment research. In: Proceedings of the 2017 ACM on Web Science Conference; 2017 Jun 25–28; Troy, NY, USA. New York, NY, USA: Association for Computing Machinery; 2017. p. 229–33. doi:10.1145/3091478.3091509.
52. Jahan MS, Oussalah M. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing. 2023;546:126232. doi:10.1016/j.neucom.2023.126232.
53. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. arXiv:1703.04009. 2017.
54. Kennedy B, Atari M, Davani AM, Yeh L, Omrani A, Kim Y, et al. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. Lang Resour Eval. 2022;56(1):79–108. doi:10.1007/s10579-021-09569-x.
55. Kennedy CJ, Bacon G, Sahn A, von Vacano C. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. arXiv:2009.10277. 2020. doi:10.48550/arXiv.2009.10277.
56. Vidgen B, Harris A, Nguyen D, Tromble R, Hale S, Margetts H. Challenges and frontiers in abusive content detection. In: Roberts ST, Tetreault J, Prabhakaran V, Waseem Z, editors. Proceedings of the Third Workshop on Abusive Language Online. Kerrville, TX, USA: Association for Computational Linguistics; 2019. p. 80–93. doi:10.18653/v1/W19-3509.
57. Vidgen B, Derczynski L. Directions in abusive language training data, a systematic review: garbage in, garbage out. PLoS One. 2020;15(12):e0243300. doi:10.1371/journal.pone.0243300.
58. Piot P, Martín-Rodilla P, Parapar J. MetaHate: a dataset for unifying efforts on hate speech detection. Proc Int AAAI Conf Web Soc Medium. 2024;18:2025–39. doi:10.1609/icwsm.v18i1.31445.
59. Wu Y, Wan J. A survey of text classification based on pre-trained language model. Neurocomputing. 2025;616(2):128921. doi:10.1016/j.neucom.2024.128921.

60. Horych T, Mandl C, Ruas T, Greiner-Petter A, Gipp B, Aizawa A, et al. The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection. In: Chiruzzo L, Ritter A, Wang L, editors. Findings of the Association for Computational Linguistics: NAACL 2025. Kerrville, TX, USA: Association for Computational Linguistics; 2025. p. 1370–86. doi:10.18653/v1/2025.findings-naacl.75.
61. Kazemi A, Natarajan Kalaivendan SB, Wagner J, Qadeer H, Verma K, Davis B. Synthetic vs. gold: the role of LLM generated labels and data in cyberbullying detection. In: Angelova G, Kunilovskaya M, Escribe M, Mitkov R, editors. Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing—Natural Language Processing in the Generative AI Era. Shoumen, Bulgaria: INCOMA Ltd.; 2025. p. 531–40.
62. Nahum O, Calderon N, Keller O, Szpektor I, Reichart R. Are LLMs better than reported detecting label errors and mitigating their effect on model performance. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Kerrville, TX, USA: Association for Computational Linguistics; 2025. p. 26782–809. doi:10.18653/v1/2025.emnlp-main.1360.
63. Piot P, Otero D, Martín-Rodilla P, Parapar J. Can LLMs evaluate what they cannot annotate? Revisiting LLM reliability in hate speech detection. arXiv:2512.09662. 2025.
64. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. New York, NY, USA: The Association for Computing Machinery (ACM); 2023. p. 46595–623. doi:10.5555/3666122.3668142.
65. Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, et al. A survey on LLM-as-a-judge. arXiv:2411.15594. 2024.
66. Tonneau M, Liu D, Fraiberger S, Schroeder R, Hale SA, Rottger P. From languages to geographies: towards evaluating cultural bias in hate speech datasets. In: Chung YL, Talat Z, Nozza D, Plaza-del-Arco FM, Rottger P, Mostafazadeh Davani A, et al., editors. Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024). Kerrville, TX, USA: Association for Computational Linguistics; 2024. p. 283–311. doi:10.18653/v1/2024.woah-1.23.
67. Pereira-Kohatsu JC, Quijano-Sánchez L, Liberatore F, Camacho-Collados M. Detecting and monitoring hate speech in twitter. Sensors. 2019;19(21):4654. doi:10.3390/s19214654.
68. Arango Monnar A, Perez J, Poblete B, Saldaña M, Proust V. Resources for multilingual hate speech detection. In: Narang K, Mostafazadeh Davani A, Mathias L, Vidgen B, Talat Z, editors. Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). Kerrville, TX, USA: Association for Computational Linguistics; 2022. p. 122–30.
69. Castillo-lópez G, Riabi A, Seddah D. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In: Scherrer Y, Jauhiainen T, Ljubešić N, Nakov P, Tiedemann J, Zampieri M, editors. Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023). Kerrville, TX, USA: Association for Computational Linguistics; 2023. p. 1–13. doi:10.18653/v1/2023.vardial-1.1.
70. Vásquez J, Andersen S, Bel-Enguix G, Gómez-Adorno H, Ojeda-Trueba SL. *HOMO*-MEX: a Mexican Spanish annotated corpus for LGBT+phobia detection on twitter. In: The 7th Workshop on Online Abuse and Harms (WOAH). Stroudsburg, PA, USA: ACL; 2023. p. 202–14. doi:10.18653/v1/2023.woah-1.20.
71. García-Díaz JA, Cánovas-García M, Colomo-Palacios R, Valencia-García R. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. Future Gener Comput Syst. 2021;114(2):506–18. doi:10.1016/j.future.2020.08.032.
72. Cervantes I. El español en el mundo anuario del instituto cervantes 2025. Madrid, Spain: Instituto Cervantes; 2025.
73. Hualde JI, Olarrea A, O'Rourke E. The handbook of hispanic linguistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2012.
74. Lee N, Jung C, Myung J, Jin J, Camacho-Collados J, Kim J, et al. Exploring cross-cultural differences in English hate speech annotations: from dataset construction to analysis. arXiv:2308.16705. 2023.
75. Lee N, Jung C, Oh A. Hate speech classifiers are culturally insensitive. In: Dev S, Prabhakaran V, Adelani DI, Hovy D, Benotti L, editors. Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). Kerrville, TX, USA: Association for Computational Linguistics; 2023. p. 35–46. doi:10.18653/v1/2023.c3nlp-1.5.

76. Sachdeva P, Barreto R, Bacon G, Sahn A, von Vacano C, Kennedy C. The measuring hate speech corpus: leveraging rasch measurement theory for data perspectivism. In: Abercrombie G, Basile V, Tonelli S, Rieser V, Uma A, editors. Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022. Paris, France: European Language Resources Association; 2022. p. 83–94.
77. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The Llama 3 herd of models. arXiv:2407.21783. 2024. doi:10.48550/arXiv.2407.21783.
78. Team G, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, et al. *Gemma* 3 technical report. arXiv:2503.19786. 2025.
79. Qwen Team. Qwen3.5: towards native multimodal agents [Internet]. 2026 [cited 2026 Jan 1]. Available from: https://qwen.ai/blog?id=qwen3.5.
80. Google Deepmind. Gemma 4: byte for byte, the most capable open models [Internet]. 2026 [cited 2026 Jan 1]. Available from: https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/.
81. Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu CH, et al. Efficient memory management for large language model serving with PagedAttention. arXiv:2309.06180. 2023.
82. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. doi:10.1177/001316446002000104.
83. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. doi:10.2307/2529310.
84. Nasution AH, Onan A. ChatGPT label: comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. IEEE Access. 2024;12:71876–900. doi:10.1109/ACCESS.2024.3402809.
85. Pangakis N, Wolken S. Knowledge distillation in automated annotation: supervised text classification with LLM-generated training labels. arXiv:2406.17633. 2024.
86. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903. 2022.