Can LLMs Evaluate What They Cannot Annotate? Revisiting LLM Reliability in Hate Speech Detection

Paloma Piot[1], David Otero[1], Patricia Martín-Rodilla[2], Javier Parapar[1]
[1]IRLab, CITIC Research Centre, Universidade da Coruña, Spain
{paloma.piot, david.otero.freijeiro, javier.parapar}@udc.es
[2]IEGPS CSIC, Santiago de Compostela, Spain
p.m.rodilla@iegps.csic.es

Abstract

Hate speech spreads widely online and harms both individuals and communities, making automatic detection essential for large-scale moderation. However, accurately detecting hate speech remains a difficult task. Part of the challenge lies in subjectivity: what one person flags as hate speech, another may see as benign. Traditional annotation agreement metrics, such as Cohen's κ, oversimplify this disagreement, treating it as an error rather than meaningful diversity. Meanwhile, Large Language Models (LLMs) promise scalable annotation, but prior studies demonstrate that they cannot fully replace human judgement, especially in subjective tasks. In this work, we reexamine LLM reliability using a subjectivity-aware framework, cross-Replication Reliability (xRR), revealing that even under fairer lens, LLMs still diverge from humans. Yet this limitation opens an opportunity: we find that LLM-generated annotations can reliably reflect performance trends across classification models, correlating with human evaluations. We test this by examining whether LLM-generated annotations preserve the relative ordering of model performance derived from human evaluation (i.e. whether models ranked as more reliable by human annotators preserve the same order when evaluated with LLM-generated labels). Our results show that, although LLMs differ from humans at the instance level, they reproduce similar ranking and classification patterns, suggesting their potential as proxy evaluators. While not a substitute for human annotators, they might serve as a scalable proxy for evaluation in subjective NLP tasks.

Keywords: hate speech, IAA, xRR, LLMs, system evaluation

1. Introduction

Hate speech, defined as "*language characterised by offensive, derogatory, humiliating, or insulting discourse (Founta et al., 2018) that promotes violence, discrimination, or hostility towards individuals or groups (Davidson et al., 2017) based on attributes such as race, religion, ethnicity, or gender (ElSherief et al., 2018a,b; Das et al., 2023)*", continues to pose a major challenge across online platforms, undermining both individual safety and social cohesion (González-Bailón and Lelkes, 2022). Recent studies estimate that nearly one in three young people have experienced some form of cyberbullying (Kansok-Dusche et al., 2022), while around half of Black or African American adults report having faced racial harassment online (Anti-Defamation League, 2024). This growing prevalence of hate speech in social platforms underscores the urgent need for reliable automatic detection methods.

Despite notable progress in hate speech detection (Glavaš et al., 2020; Wang and Chang, 2022; Plaza-del arco et al., 2023; Roy et al., 2023; Piot and Parapar, 2025), the quality and consistency of human-annotated data remain central challenges (Toraman et al., 2022; Ljubešić et al., 2022; Barbarestani et al., 2024). The task is inherently subjective (Aroyo et al., 2019; Kanclerz et al., 2022; Hettiachchi et al., 2023), as judgements often depend on cultural norms, speaker intent, and contextual interpretation (Röttger et al., 2022; Davani et al., 2023; Dehghan et al., 2025). This subjectivity is further amplified by linguistic factors, as hate speech is often conveyed implicitly through sarcasm, metaphor, or coded language and depends on discourse-level context and target inference. As a result, annotation disagreement is not always an error but may reflect legitimate diversity in interpretation (Kralj Novak et al., 2022; Mostafazadeh Davani et al., 2022; Barbarestani et al., 2024).

Conventional inter-annotator agreement (IAA) measures, such as Cohen's κ, assume a single objective ground truth and penalise all disagreement equally (Cohen, 1960), making them improper for tasks involving subjectivity and ambiguity such as hate speech detection (Marchal et al., 2022; Plank, 2022; Bassi et al., 2025).

To address this limitation, subjectivity-aware metrics have been proposed. Among them, the xRR (cross-Replication Reliability) framework (Wong et al., 2021) has emerged as a promising alternative, as it evaluates how consistently annotators reproduce each other's labelling behaviour rather than enforcing a single gold label. This shift better captures pluralism in human judgements, offering a more realistic picture of annotation quality in socially grounded tasks (Dutta et al., 2023; Prabhakaran et al., 2024).

Meanwhile, LLMs have been explored as poten-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4358–4370
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tial automatic annotators, offering scalability and cost-efficiency (Gilardi et al., 2023; Törnberg, 2024). However, current evidence shows that LLMs cannot yet replace human annotators, particularly in tasks that require social or contextual interpretation (Mohta et al., 2023; Tseng et al., 2025; Baumann et al., 2025). Despite these advances, most prior studies assessing LLM annotations rely on traditional agreement metrics (Matter et al., 2024; Giorgi et al., 2025), which may undervalue their performance under subjective conditions. This raises an important motivational question: does evaluating LLM-based annotations with subjectivity-aware metrics alter the conclusions we may draw about their reliability as annotators?

Building on this motivation, we first investigate the role of LLMs as annotators in multilingual hate speech detection (English and Spanish). Our findings show that, even when adopting xRR to account for subjectivity, LLMs still fail to capture the full spectrum of human annotation patterns. Nonetheless, a complementary role emerges: rather than replacing human annotators, LLMs may serve as system evaluators. We conduct a second experiment to test whether LLM-generated annotations can reproduce the relative ranking of hate speech detection models established by human annotators. This framework reveals that LLMs, despite instance-level divergence, can reliably approximate classification models' performance trends, offering a scalable path for comparative evaluation without requiring full human annotation.

Our study is structured around three research questions that build sequentially on one another:

- ***RQ1****: How do traditional agreement metrics capture annotation quality in hate speech detection, both for human and LLM-generated labels?*
- ***RQ2****: If using subjectivity-aware metrics, do LLMs demonstrate greater reliability as annotators, or do they still diverge from human annotation patterns?*
- ***RQ3****: Even if LLMs are not reliable annotators, can their annotation patterns be leveraged for evaluation, preserving the relative ordering of hate speech detection models observed under human judgements?*

We structure our study around two complementary experiments: (i) an agreement analysis, addressing ***RQ1*** and ***RQ2***, where we compare traditional and subjectivity-aware metrics to assess annotation reliability across human and LLM judgements; and (ii) a ranking correlation experiment, addressing ***RQ3***, where we test whether LLM-based annotations maintain the relative ordering of model performance derived from human evaluation (for example, whether models ranked as more reliable by human annotators preserve the same ordering when evaluated with LLM-generated labels). Through these questions, we aim to deepen the understanding of how metric selection choice shapes annotation assessment and to clarify the emerging role of LLMs as evaluators in socially grounded tasks.

2. Related Work

Hate speech detection. Hate speech detection remains a challenging task due to its inherently subjective nature, where individual perceptions, socio-cultural context, and linguistic nuance often lead to divergent annotations (Fortuna and Nunes, 2018; Kanclerz et al., 2022). Beyond annotator-specific factors, disagreement is also driven by intrinsic properties of language itself. Hate speech is frequently expressed implicitly, through sarcasm, irony, metaphor, or coded language, rather than explicit slurs, making its interpretation highly dependent on pragmatics and shared background knowledge (Röttger et al., 2022; Davani et al., 2023). Furthermore, discourse-level phenomena such as context dependency, target ambiguity, and the need to infer speaker intent complicate annotation, as isolated instances may not provide sufficient information to determine whether content is harmful.

IAA in this domain tends to be low, reflecting not only annotator variability but also these linguistic and contextual ambiguities (Toraman et al., 2022; Ljubešić et al., 2022; Barbarestani et al., 2024). Prior research has emphasised that such disagreement is not merely noise but a signal of underlying subjectivity and interpretative complexity in annotation (Basile, 2020; Leonardelli et al., 2021; Guellil et al., 2024). This challenge has prompted renewed attention to how annotation quality is measured, especially in socially grounded tasks where disagreement among annotators is common.

IAA in subjective tasks. Traditional metrics such as Cohen's κ (Cohen, 1960) and Fleiss's κ (Shrout and Fleiss, 1979) are commonly used to assess annotation consistency. However, they often underestimate reliability in socially grounded tasks, as they penalise legitimate subjective variation (Pavlick and Kwiatkowski, 2019). Recent work has proposed ways to better capture human uncertainty, including soft-label modelling (Uma et al., 2022) and multi-annotator models, where each annotator's judgements are treated as separate subtasks while sharing a common learned representation (Mostafazadeh Davani et al., 2022). Yet, less attention has been paid to how agreement itself is measured. Metrics such as xRR (Wong et al., 2021) explicitly address subjectivity in annotation, capturing consistent patterns of disagreement rather than treating them as noise or error. For instance, in subjective annotation tasks, Dutta et al. (2023) demonstrate that xRR reveals community-specific perspectives in toxic comment annotation, while Prabhakaran et al. (2024) leverage xRR to analyse inter-group disagreement patterns, highlighting how perspectives vary across demographic groups.

LLMs as annotators. LLMs have been investigated as tools to assist or replace human annotators, providing a scalable way to generate labels for large datasets (Gilardi et al., 2023; Törnberg, 2024). Studies show that while LLMs can approximate human judgements in certain tasks, their reliability is generally lower than humans, especially in tasks that involve social nuance or context-dependent interpretation (Mohta et al., 2023; Tseng et al., 2025; Baumann et al., 2025). Recent work has therefore focused on understanding where LLMs succeed or fail in mimicking human annotations, leaving open questions about alternative uses.

Given this, our study is motivated by the gap in understanding LLM performance under subjectivity-aware metrics and its implications for evaluation. We focus on where LLMs diverge from humans and how they might serve as model evaluators in subjective tasks like hate speech.

3. Experimental Setup

To address our three research questions, we design a two-stage experimental study grounded in a unified setup. All experiments rely on the same hate speech datasets, language models, and LLM-generated annotations. This section details the datasets, the selected models, the prompting and annotation procedure, and the metrics used to assess IAA.

3.1. Data

For this study, we build upon the MetaHate (Piot et al., 2024) and MetaHateES (Piot et al., 2025) collections, two large-scale meta-collections that consolidate the major publicly available hate speech datasets in English and Spanish, respectively. Both resources were designed to standardise and unify heterogeneous hate speech resources. MetaHate integrates 36 publicly available datasets in English covering diverse social media platforms, domains and annotation schemes, while MetaHateES includes 10 different sources, focusing on European Spanish hate speech. We focus on these languages as high-resource languages, enabling controlled cross-lingual comparison (Fortuna and Nunes, 2018; Poletto et al., 2020; Vidgen and Derczynski, 2020), and providing robust baselines for future extension to low-resource varieties.

Given the focus of this work on annotation subjectivity and inter-annotator reliability, we selected from them only those datasets that contained individual-level annotation that could enable a fine-grained comparison between human and LLM judgements. Following this criterion, we initially identified two datasets from MetaHate: HateXplain (Mathew et al., 2020) and Measuring Hate Speech (MHS) (Kennedy et al., 2020; Sachdeva et al., 2022); and two from MetaHateES: DETESTS (Schmeisser-Nieto et al., 2024) and EXIST (Rodríguez-Sánchez et al., 2022). Table 1 summarises these datasets.

Dataset	Lang.	# annotations	# posts
EXIST	ES	6	4209
DETESTS	ES	3	9906
HateXplain	EN	3	$20\,109$
MHS	EN	1-815	$39\,565$

Table 1: Overview of the *potential* datasets that can be used in our experiments. “Lang.” denotes the dataset language (ES = Spanish, EN = English). “# annotations” indicates the number of independent judgements available per post, and “# posts” reports the number of unique, non-duplicated instances.

The task under study is binary hate speech detection, where each post is classified as either `hate` or `non-hate`. Most selected datasets follow this formulation directly. However, some require label harmonisation to ensure cross-dataset consistency. HateXplain was originally annotated with three labels: “normal”, “offensive” and “hate speech”. Following prior work (Piot et al., 2024), we merged “offensive” category into `no-hate`, as offensive content does not fit our hate speech definition. The MHS dataset includes an additional label, “unclear”; we removed all instances with at least one “unclear” annotation to maintain consistency with the binary setup.

To ensure a fair comparison across datasets, we restricted our analysis to instances with exactly three human annotators. Therefore, the EXIST dataset was excluded, and we only retained MHS instances meeting this criterion. After unifying label spaces, we derived a gold label for each post based on the majority vote across annotators and recorded whether the gold label was unanimous or not. Table 2 summarises the final datasets and the number of posts used in our experiments.

Dataset	Lang.	# annot.	# posts	# unanimous	% hate
DETESTS	ES	3	9906	8589	26.30
HateXplain	EN	3	20 109	13 910	29.49
MHS	EN	3	9133	6428	25.64

Table 2: Overview of the datasets *included* in our experiments. "Lang." denotes the dataset language (ES = Spanish, EN = English). "# annot." indicates the number of independent judgements used per post, "# posts" reports the number of unique, non-duplicated instances, and "# unanimous" reports the number of posts with unanimous labels.

3.2. Models

To generate LLM-based annotations for our experiments, we selected three state-of-the-art instruction-tuned models, all running in 4-bit quantised versions to optimise computational efficiency without compromising performance. The choice of three models mirrors the number of human annotators per dataset, ensuring a comparable annotation setup. Below are the details of the LLMs we employed:

- `Llama 3.1` (`Llama-3.1-8B-Instruct`) (Grattafiori et al., 2024): Developed by Meta AI, this model is part of the Llama 3 series, which includes models ranging from 8B to 405B parameters. The 8B variant is optimised for instruction-following tasks and has demonstrated strong performance across multilingual and reasoning benchmarks.

- `Nemo` (`Mistral-Nemo-Instruct-2407`) (Mistral AI team, 2025): This model comes from a collaboration between Mistral AI and NVIDIA. It is an instruction-tuned model based on the Mistral architecture, and it is designed to excel in multilingual tasks and has been fine-tuned to improve reasoning and coding accuracy.

- `DeepSeek` (`DeepSeek-R1-Distill-Llama-8B`) (DeepSeek-AI, 2025): This model is a distilled version of the `Llama-3.1-8B-Instruct` model, fine-tuned using samples from DeepSeek-R1. The distillation process aims to retain the performance of the original model while reducing computational requirements. It is particularly noted for its efficiency and effectiveness in reasoning tasks.

3.3. LLM judgements

To obtain automatic annotations, we queried each selected LLM using a prompt template. The models were asked to assess whether a given social media message contained hate speech and respond strictly with a binary label as: `True` (`hate`) or `False` (`no-hate`). No explicit definition or examples of hate speech were provided; the models generated labels based solely on their internal understanding of the task.

Prompts. All models were queried with the same two-part prompt to keep the setup consistent. The system prompt introduced the task as a moderation scenario, and the base instruction told the model to return a binary decision. Tables 3 and 4 provide the exact wording.

You are a content moderation system. Your task is to detect hate speech in social media messages.

Table 3: System prompt.

Respond only with 'True' if the message contains hate speech, or 'False' if it does not. Do not explain, justify, or add anything else. Respond with exactly one word: True or False.

Table 4: Base instruction.

Inference setup. During inference, all models were run under a unified configuration to ensure comparability. We set the maximum sequence length to 4096 tokens to accommodate longer social media posts without truncation. The generation was constrained to a single output token ($max\_tokens = 1$), since models were required to produce only one-word answers (`True` or `False`), that, for our selected models, correspond to a single token. To guarantee deterministic outputs, we disabled sampling by fixing $do\_sample = False$, and used a very low $temperature$ (0.01) together with $top\_p = 0.1$ and $top\_k = 5$. Finally, to prevent invalid generations, we restricted the output space using a custom `LogitsProcessor` from `transformers`[1] library that filtered all tokens except `True` and `False`. Each LLM was executed in 4-bit quantised inference mode using `unsloth`[2] framework. The resulting predictions constitute the *LLM judgements*[3].

LLM pseudo-raters. As we did with the human annotations, we also combined individual model

[1] https://huggingface.co/docs/transformers/

[2] https://unsloth.ai/

[3] Our code is available at https://github.com/palomapiot/hate-eval-agreement/

predictions. We derived a majority label, based on the majority vote across LLMs, and also recorded whether this majority label was unanimous or not (LLM consensus).

3.4. Metrics

Agreement metrics. To assess annotation quality and reliability, we compute both traditional and subjectivity-aware IAA measures. As standard metrics, we include **Cohen's** κ (Cohen, 1960), **Fleiss's** κ (Shrout and Fleiss, 1979) and **Krippendorff's** α (Hayes and Krippendorff, 2007).

Cohen's κ measures the degree of pairwise agreement between two annotators while correcting for agreement expected by chance. It is widely used in annotation studies to quantify consistency between raters and to distinguish systematic alignment from random coincidence. κ values typically range from 0 (no agreement beyond chance) to 1 (perfect agreement), with intermediate ranges (e.g., 0.2–0.4) often interpreted as fair or moderate agreement. Since our datasets were annotated by three annotators, we compute Cohen's κ for all annotator pairs and report the average, which also enables compatibility with the xRR framework (Wong et al., 2021) (see below), as xRR builds upon pairwise κ to quantify inter-group agreement. To complement pairwise analyses, we additionally compute Fleiss' κ which generalises Cohen's κ to multiple raters, and Krippendorff's α, a flexible measure that accommodates multiple rates, missing data and different measurement levels.

These metrics all correct for chance agreement, assuming that all disagreements carry equal weight and that it exists a single, objective ground truth. These assumptions are often problematic in socially grounded tasks such as hate speech detection, where annotator background, cultural framing, or interpretation of intent can legitimately diverge. To address these limitations, we also employ **xRR** (cross-Replication Reliability) framework (Wong et al., 2021), which extends traditional reliability analysis by accounting for systematic patterns of disagreement. xRR models the extent to which one group of annotators can reproduce the distribution of labels from another, offering a more nuanced and interpretable estimate of replicability in subjective settings. We omit metrics like simple percent agreement or Scott's π, as they either ignore chance or assume annotator interchangeability, making them unsuitable for subjective, categorical annotations.

4. Experiment 1: Agreement Analysis

The first experiment focuses on annotation agreement, addressing **RQ1** and **RQ2**. Our goal is to assess how well LLM-generated annotations align with human judgements and whether subjectivity-aware metrics provide a more nuanced view than traditional measures.

Traditional metrics. We begin by assessing annotation reliability using Cohen's κ (Cohen, 1960), a statistic that measures inter-rater reliability, assessing the agreement between two raters on categorical items while accounting for agreement that could occur by chance. To do so, we compared (i) pairwise κ between human annotators and LLMs: (ii) each rater vs. the consensus of the other raters within the group (*leave-one-out* κ), (iii) each human vs. LLM majority and each LLM vs. human majority (cross-group κ), and (iv) group-level κ based on the majority and consensus labels. Figure 1 reports (i) and Table 5 reports the rest of the κ scores.

	Rater	DETESTS	HateXplain	MHS
Leave-one-out: *Humans*	H1	0.840	0.470	0.487
	H2	0.822	0.471	0.478
	H3	0.852	0.462	0.478
	Mean	0.838	0.468	0.481
Cross-group: *LLM vs. humans leave-one-out*	Llama vs. H1+H2	0.052	0.027	0.019
	Llama vs. H1+H3	0.054	0.022	0.010
	Llama vs. H2+H3	0.051	0.027	0.025
	Nemo vs. H1+H2	**0.300**	**0.246**	**0.223**
	Nemo vs. H1+H3	**0.293**	**0.243**	**0.223**
	Nemo vs. H2+H3	**0.304**	**0.244**	**0.217**
	DeepSeek vs. H1+H2	0.043	0.001	−0.019
	DeepSeek vs. H1+H3	0.045	−0.006	−0.018
	DeepSeek vs. H2+H3	0.045	0.000	−0.018
Cross-group: *LLM vs. humans maj.*	Llama	0.048	0.029	0.023
	Nemo	**0.309**	**0.343**	**0.296**
	DeepSeek	0.042	0.005	−0.017
	Mean	0.133	0.126	0.101
Cross-group: *Human vs. LLMs maj.*	H1	0.060	0.025	0.015
	H2	0.059	0.031	0.023
	H3	0.061	0.024	0.018
	Mean	0.060	0.026	0.019
Group-level	Majority	0.066	0.029	0.020
	Consensus	0.036	0.001	0.006

Table 5: Cohen's κ agreement scores across human annotators and LLMs for all datasets. The table reports: (ii) leave-one-out κ (each rater vs. the consensus of others within their group), (iii) cross-group κ (each human vs. LLM majority and each LLM vs. human majority), and (iv) group-level κ based on majority and consensus labels. *H* represents "human".

Human annotators exhibit high pairwise agreement across datasets, with mean inter-human κ values ranging from 0.50 to 0.78. The *leave-one-out* analysis further confirms this reliability: each human annotator aligns closely with the consensus of the others, yielding κ values between 0.46 and 0.85. Among datasets, DETESTS achieves an almost perfect agreement ($\kappa \approx 0.84$), while HateXplain and MHS remain at moderate agreement ($\kappa \approx 0.47$). These results suggest that human annotators are relatively stable, even when evaluated "*one versus the rest*" scenario, highlighting a robust signal in the underlying judgement process.

In contrast, LLMs exhibit lower agreement both among themselves and with human annotators.

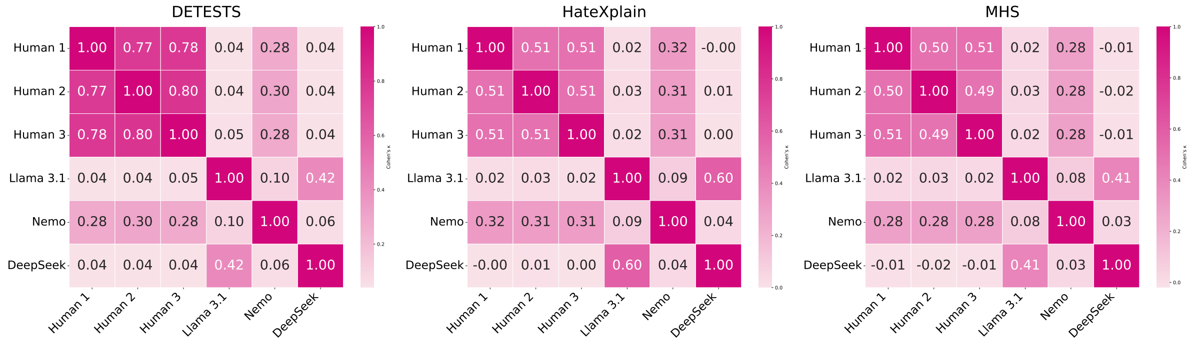

Figure 1: Pairwise Cohen's κ agreement between human annotators and LLMs across the three datasets. Cells show κ scores between pairs of raters.

Overall, pairwise inter-LLM κ is generally *slight* ($\kappa \approx 0.17 - 0.24$ on average). When comparing LLMs with both the human majority and the *leave-one-out* human evaluations, a clear pattern emerges: `Nemo` demonstrates the most consistent alignment with humans, achieving κ values near the *fair* agreement range (0.22-0.34) in pairwise comparisons. These scores are comparable to those reported for crowdworkers or lightly instructed annotators in prior subjective annotation studies (Modha et al., 2019; Assenmacher et al., 2025). This suggests that `Nemo` is able to capture meaningful aspects of human annotation patterns, despite not being fine-tuned for the task. By contrast, `Llama 3.1` and `DeepSeek` show negligible agreement with the human majority and *leave-one-out* human annotations (below 0.06), with `DeepSeek` even exhibiting negative values. This suggests that these models are unreliable for hate speech annotation.

Cross-group comparisons illustrate the persistent gap between human and LLM annotations. When each human is compared to the LLM majority, κ values remain below 0.06 on average, whereas individual LLMs compared to the human majority achieve slightly higher agreement (up to 0.34 for `Nemo`). Among the models, `Nemo` consistently shows the strongest alignment with human annotations, further confirming its comparatively higher sensitivity to human labelling patterns. These findings suggest that while aggregated LLM predictions (majority or consensus) slightly approach human alignment, they still fall far short of human reliability. The asymmetry between "human vs. LLM majority" and "LLM vs. human majority" underscores the uneven way LLMs capture human labelling patterns.

At the group level, κ scores based on majority or full consensus are minimal (< 0.07), confirming that aggregated pseudo-raters do not reach the reliability of human annotators. Overall, our analysis demonstrates that while LLMs can approximate certain trends in human annotations, the gap in agreement remains substantial, emphasising the need for careful interpretation of LLM-generated labels in subjective, socially grounded tasks such as hate speech detection.

Subjectivity-aware metrics. To better capture the reliability of annotations in subjective settings like hate speech detection, we complement traditional agreement measures with xRR. This framework begins with a general definition of Cohen's κ that is extended to cross-kappa (κ_x)–designed to measure annotation agreement between replications in a chance-corrected manner–and then is used to define $normalized\ \kappa_x$, that measures *similarity* between two replications (Wong et al., 2021). One of the utilities of this metric is to assess whether collected crowdsourced data closely mirror the target (i.e. data with expert annotations). In our study, the crowdsourced data are our LLMs judgements, and our target is the human-annotated data. According to the authors of the metric, a high $normalized\ \kappa_x$ can assure us that the crowdsourced annotators (i.e. LLMs) are functioning as an extension of the trusted annotators. This metric approximates the true correlation between two experiments' item-level mean scores.

Table 6 reports Fleiss' κ, Krippendorff's α and the average pairwise Cohen's κ for both human and LLM annotators across datasets, alongside the $normalized\ \kappa_x$ between Cohen's pairs. As expected, human annotators display strong internal agreement, with $\kappa, \alpha = 0.78$ for DETESTS and moderate values around 0.5 for HATEXPLAIN and MHS. In contrast, the LLMs exhibit slight to fair agreement ($\kappa \approx 0.17 - 0.24$), indicating substantial divergence from humans.

Although LLMs exhibit only slight-to-fair agreement, the $normalized\ \kappa_x$ values reveal a fair to moderate degree of similarity between the human and LLMs annotations ($0.35 \leq normalized\ \kappa_x \leq 0.41$). This metric does not evaluate pairwise agreement within a group but instead compares two independent replications of an annotation experiment (in our case, the human and the LLM groups). It quantifies how similarly the two groups behave on aver-

	DETESTS	HATEXPLAIN	MHS
Human Fleiss's κ	0.784	0.510	0.501
LLMs Fleiss's κ	0.045	-0.032	-0.010
Human Krippendorffs's α	0.784	0.510	0.501
LLMs Krippendorff's α	0.045	-0.032	-0.010
Human mean Cohen's κ	0.784	0.510	0.501
LLMs mean Cohen's κ	0.195	0.244	0.174
$normalized\ \kappa_x$	0.357	0.412	0.411

Table 6: Inter-rater reliability for human annotators and LLMs, computed as using Fleiss' κ, Krippendorff's α and mean pairwise Cohen's κ, with $normalized\ \kappa_x$ indicating the similarity between LLM and human annotation distributions. Metric values are reported up to three decimal; while they appear identical for humans, differences become apparent at the fifth or sixth decimal.

age across items, correcting for chance and for differences in the marginal label distributions. Hence, a moderate $normalized\ \kappa_x$ indicates that, although the LLMs disagree among themselves on individual items, the overall pattern of their aggregated annotations moderately mirrors the general tendencies observed in the human judgements. While traditional IAA metrics portray LLMs as largely unreliable, subjectivity-aware $normalized\ \kappa_x$ shows a more positive picture: although LLMs still fall short of human-level reliability, they partially capture human annotation patterns.

Error analysis. We conducted a qualitative error analysis on the HATEXPLAIN dataset, which was selected because it provides fine-grained, target-level annotations. This granularity allows us to identify which specific groups are systematically missed by LLMs. To capture overall trends, we computed per-target error rates, comparing both individual model predictions and the majority-vote aggregation against the human majority labels.

First, we examine error distributions, where LLMs show a clear asymmetry. Both, `Llama 3.1` and `DeepSeek` miss a large number of hate instances (2794 and 2896 false negatives, respectively) while generating comparatively few false positives (691 and 447), indicating a conservative detection bias. In contrast, `Nemo` attains the lowest false negative rate (620) but substantially over-predicts hate, with 7175 false positives. When considering the majority vote of all LLMs, the combined predictions yield 2819 false negatives and 594 false positives.

Figure 2 shows the the per-target analysis, highlighting systematic weaknesses of LLMs. Rare or nuanced targets such as *Asexual* (100% missed), *Non Religious* (55.2%), *Minority* (51.1%) and *Jewish* (35.6%) are particularly challenging, reflecting under-detection for low-frequency groups (e.g. target *Asexual* only has 5 posts in the dataset) or

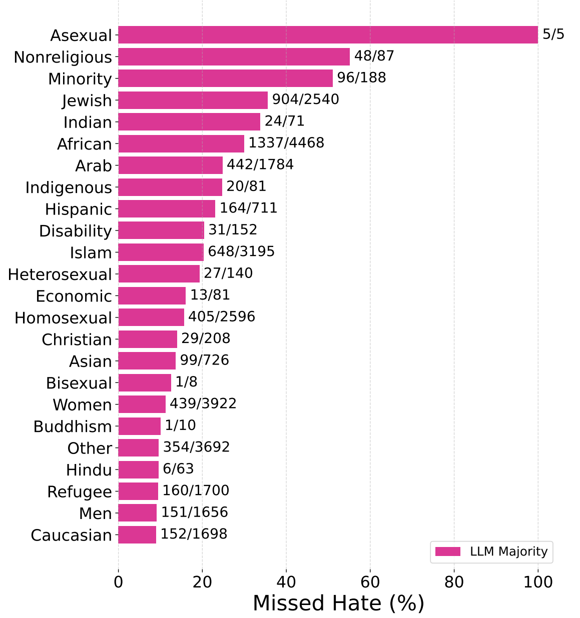

Figure 2: Per-target Missed Hate (%) for LLM majority.

contextually subtle hate. Similarly, racial and ethnic categories like *African* (29.9%) and *Arab* (24.8%) also exhibit substantial false negatives. In contrast, more frequent or explicitly signalled targets such as *Women* (11.2%), *Men* (9.1%), *Refugee* (9.4%) and *Caucasian* (9.0%) show lower miss rates, indicating stronger detection of gender and migration-related hate. Some religious categories, such as *Islam* (20.3%) and *Christian* (13.9%), are captured moderately; however, performance varies substantially across other faith-related labels, including *Non Religious* (55.2%) and *Jewish* (35.6%).

Across models, `Llama 3.1` and `DeepSeek` display near identical patterns, while `Nemo` shows lower miss rates across all targets, but often at the cost of over-predicting hate, as reflected in its high false positive count. Overall, these findings indicate that current LLMs effectively capture hate speech targeting frequent and explicitly represented gender categories (e.g., *Women* and *Men*), but struggle with less frequent or more nuanced gender and sexuality-related identities (e.g., *Asexual*), as well as other minority groups.

5. Experiment 2: Ranking Correlation

Section 4 showed that LLMs are not yet reliable as annotators at the instance level. However, their partial alignment with human labelling behaviour motivates exploring an alternative role: their potential as system evaluators. In this second experiment we focus on **RQ3**, which asks whether LLM-generated labels can be leveraged to assess hate speech detection classifiers by preserving

	DETESTS		HateXplain		MHS	
Annotator	τ	F1 Diff.	τ	F1 Diff.	τ	F1 Diff.
LLM consensus	0.898	0.381	0.757	0.386	0.617	0.376
LLM majority	0.910	0.353	0.825	0.348	0.743	0.347
`Llama 3.1`	0.931	0.357	0.846	0.338	0.788	0.351
`Nemo`	$\mathbf{0.952}$	$\mathbf{0.041}$	$\mathbf{0.964}$	$\mathbf{-0.074}$	$\mathbf{0.949}$	$\mathbf{-0.006}$
`DeepSeek`	0.893	0.342	0.419	0.370	-0.747	0.351
Human 1	0.997	-0.005	0.998	0.011	0.995	0.005
Human 2	0.997	-0.020	0.999	0.012	0.995	0.004
Human 3	0.998	-0.009	0.999	0.006	0.997	0.003

Table 7: Ranking correlation (Kendall's τ) and mean F1 difference for human vs. LLM annotation. For each dataset, results for the best LLM are **in bold**.

the relative ordering of model performance observed under human judgements. Unlike instance-level agreement, the primary concern for evaluation is whether the labels induce stable, human-consistent rankings across classifiers. For example, if we have three different classifiers, and evaluation under human annotations yields the ordering `Model A` > `Model B` > `Model C`, and evaluation under LLM-based annotations yields the same ordering, then we would conclude that LLM-based annotations can be considered valid for evaluation.

Experimental setup. To investigate this question, we simulate a set of classifiers using a controlled degradation procedure. We start from an "oracle system" that perfectly reproduces the human majority labels ($F1 = 1.0$), and generate synthetic classifiers by randomly flipping a proportion $p\%$ of predictions. We simulate degradation levels in the range $p \in [0, 0.99]$ with a step size of 0.01, resulting in 100 uniformly spaced degradation points. Each synthetic classifier is then evaluated against both human and LLM labels, producing paired performance scores. For each degradation level, label flipping is performed deterministically using a fixed seed derived from p to ensure reproducibility. From these scores, we derive rankings of classifiers under human and LLM evaluation, which we compare using Kendall's τ (Kendall, 1938). Kendall's τ is a rank correlation coefficient that measures the similarity between two orderings: it ranges from -1 (inverse ordering) to 1 (identical order), with 0 indicating no association. To complement this ranking-based measure, we also report the mean absolute difference in $F1_{BINARY}$ scores across models to quantify potential distortions in absolute performance. Algorithm 1 summarises our experiment. With this setup we can assess whether LLM-generated annotations can reproduce human-derived performance orderings, even when per-instance agreement is limited, providing a practical test of their utility as evaluators.

Algorithm 1: Correlation-based Evaluation of LLM Annotations

Input: Human labels H, LLM labels L, degradation levels $P = \{p_1, \dots, p_k\}$
Output: Kendall's τ correlation between rankings, AverageF1Diff
foreach $p \in P$ **do**
  Generate synthetic predictions D_p by flipping a proportion $p \in [0, 0.99]$ of labels in H (gold labels H themselves remain unchanged);
  Compute $F1_H(p) \leftarrow F1(H, D_p)$;
  Compute $F1_L(p) \leftarrow F1(L, D_p)$;
  Store $(p, F1_H(p), F1_L(p))$ in results;
Build rankings $Rank_H$ and $Rank_L$ from results ;
Compute Kendall's τ between $Rank_H$ and $Rank_L$;
Compute $AverageF1Diff \leftarrow \frac{1}{|P|} \sum_{p \in P} F1_H(p) - F1_L(p)$;
return $(\tau, AverageF1Diff)$;

Results. Table 7 presents Kendall's τ between rankings derived from human and LLM labels, alongside the mean absolute difference in $F1_{BINARY}$ scores. Across datasets, most LLM variants achieve high rank correlation ($\tau \approx 0.84$-0.96), indicating that despite low instance-level agreement (see Table 5), they largely preserve the relative order of classifiers. The strongest performer is `Nemo`, consistently achieving $\tau \geq 0.95$ with minimal $F1_{BINARY}$ deviation from human-based evaluation (e.g. 0.041 on DETESTS). In contrast, `DeepSeek` fails on MHS, ($\tau < 0$), producing an inverted ranking relative to human judgements. For `Llama 3.1` and `DeepSeek`, mean $F1_{BINARY}$ differences remain substantial, showing that high rank correlation does not ensure accurate absolute scores. Overall, these results indicate that preserving relative model rankings is a less stringent requirement unlike instance agreement and can be achieved by some LLMs even when κ is low. While LLMs offer coarse evaluations of relative model performance, their reliability is model and dataset dependent, and absolute score estimates remain biased.

6. Conclusions

This study examined the reliability of LLMs as annotators for hate speech detection through the lens of subjectivity-aware metrics. We compared traditional agreement metrics, such as Cohen's κ, with the cross-Replication Reliability framework, which accounts for systematic disagreement patterns across annotator groups.

Our findings show that while LLMs achieve only slight to fair agreement under conventional metrics, subjectivity-aware evaluation paints a more optimistic picture: their judgements partially mirror human annotation tendencies, though they remain below human-level reliability. However, this improvement is not uniform across models: among the evaluated systems, `Nemo` achieves substantially higher agreement (approaching fair κ), whereas `Llama 3.1` and `DeepSeek` remain in the slight agreement range. This indicates that our conclusions are partially influenced by model selection and that not all LLMs exhibit the same degree of alignment with human annotators. These results suggest that LLMs should not yet be treated as full substitutes for human annotators in socially grounded tasks. Furthermore, aggregation via majority voting does not necessarily improve alignment: in our setting, the majority vote largely reflects the behaviour of the two less aligned models, effectively diluting the comparatively stronger performance of `Nemo`.

However, when repurposed as evaluators rather than annotators, LLMs can reliably reproduce relative model performance. In this correlation experiment, stronger models achieve near-human ordering (Kendall's $\tau \geq 0.95$) with minimal mean F1 distortion, whereas weaker or misaligned models may produce inverted orderings (negative τ). This suggests that preserving relative model rankings is a weaker requirement than annotator-level agreement: carefully selected LLM labels can support model comparison and evaluation audits when human labels are scarce, provided that correlation is verified on the target dataset.

7. Ethics

Hate speech detection involves the processing of sensitive social media data and potentially harmful language. All datasets used in this study were publicly released and have been previously employed for research purposes. We strictly adhered to the terms of use of each dataset and ensured that no personally identifiable information was used or shared. While our experiments involved LLMs generating annotations on hate-related content, we took care to avoid exposing harmful text outside controlled experimental settings. We acknowledge the societal risks of misclassification and the potential biases embedded in both human and machine judgements. Our work aims to promote transparency and responsible evaluation for sensitive tasks like hate speech detection.

8. Limitations

Our analysis is limited to binary hate speech detection and to datasets in English and Spanish, which may restrict the generalisability of our findings to other languages or more fine-grained hate categories. The LLMs evaluated represent only a subset of available architectures, and performance may vary with different models or prompt formulations. We intentionally adopted simple prompts to isolate model behaviour, leaving prompt variation outside the scope of this work. Finally, we constrained comparisons to instances with exactly three human annotators to ensure parity with the number of LLM raters, which may have reduced overall data coverage.

Acknowledgments

The authors thank the funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The authors thank the financial support supplied by the grant PID2022-137061OB-C21 funded by MI-CIU/AEI/10.13039/501100011033 and by "ERDF/EU". The authors also thank the funding supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditations ED431G 2023/01 and ED431C 2025/49) and the European Regional Development Fund, which acknowledges the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

References

Anti-Defamation League. 2024. Online hate and harassment: The american experience 2024. Accessed: 03/01/2024.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Dennis Assenmacher, Paloma Piot, Katarina Laken, David Jurgens, and Claudia Wagner. 2025. Beyond the explicit: A bilingual dataset for dehumanization detection in social media.

Baran Barbarestani, Isa Maks, and Piek T.J.M. Vossen. 2024. Content moderation in online platforms: A study of annotation methods for inappropriate language. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 96–104, Torino, Italia. ELRA and ICCL.

Valerio Basile. 2020. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *AIxIA 2020 – Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, page 441–453, Berlin, Heidelberg. Springer-Verlag.

Davide Bassi, Dimitar Iliyanov Dimitrov, Bernardo D'Auria, Firoj Alam, Maram Hasanain, Christian Moro, Luisa Orrù, Gian Piero Turchi, Preslav Nakov, and Giovanni Da San Martino. 2025. Annotating the annotators: Analysis, insights and modelling from an annotation campaign on persuasion techniques detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17918–17929, Vienna, Austria. Association for Computational Linguistics.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. *Proceedings of the ICWSM 2023*, 17:1014–1023.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the ICWSM 2017*, 11(1):512–515.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Somaiyeh Dehghan, Mehmet Umut Sen, and Berrin Yanikoglu. 2025. Dealing with annotator disagreement in hate speech classification.

Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. 2023. Modeling subjectivity (by mimicking annotator annotation) in toxic comment identification across diverse communities.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the ICWSM 2018*, 12(1).

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018b. Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the ICWSM 2018*, 12(1).

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, ..., and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the ICWSM 2018*, 12(1).

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):653–670.

Goran Glavaš, Vanja Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sandra González-Bailón and Yphtach Lelkes. 2022. Do social media undermine social cohesion? a critical review. *Social Issues and Policy Review*, 17(1):155–180.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, ..., and Zhiyu Ma. 2024. The llama 3 herd of models.

Imane Guellil, Yousra Houichi, Sara Chennoufi, Mohamed Boubred, Anfal Yousra Boucetta, and Faical Azouaou. 2024. The annotators agree to not agree on the fine-grained annotation of hate-speech against women in Algerian dialect comments. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 133–139, Torino, Italia. ELRA and ICCL.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1):38–50.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. 2022. A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*, 24(4):2598–2615.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application.

Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695, Cham. Springer International Publishing.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikola Ljubešić, Igor Mozetič, and Petra Kralj Novak. 2022. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, 29(6):1481–1494.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI 2020*.

Daniel Matter, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer. 2024. Investigating the increase of violent speech in incel communities with human-guided gpt-4 prompt iteration. *Frontiers in Social Psychology*, 2.

Mistral AI team. 2025. Mistral NeMo | Mistral AI. `https://mistral.ai/news/mistral-nemo`. Accessed: 13/08/2025.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, and Daksh Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. *Working Notes of FIRE 2019*, 2517.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Paloma Piot, José Ramom Pichel Campos, and Javier Parapar. 2025. Bridging gaps in hate speech detection: Meta-collections and benchmarks for low-resource iberian languages.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):2025–2039.

Paloma Piot and Javier Parapar. 2025. *Towards Efficient and Explainable Hate Speech Detection via Model Distillation*, page 376–392. Springer Nature Switzerland.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, page 229–240.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

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

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

Wolfgang S. Schmeisser-Nieto, Pol Pastells, Simona Frenda, Alejandro Ariza-Casabona, Mireia Farrús, Paolo Rosso, and Mariona Tulé. 2024. Overview of detests-dis at iberlef 2024: Detection and classification of racial stereotypes in spanish - learning with disagreement. *Procesamiento del Lenguaje Natural*, 73:323–333.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. Evaluating large language models as expert annotators. In *Second Conference on Language Modeling*.

Petter Törnberg. 2024. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 0(0):08944393241286471.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065. Association for Computational Linguistics.