uc3m | Universidad Carlos III de Madrid

Master's Degree in Computational Social Sciences

Academic Year 2024/25

Master's Thesis

Beyond Agreement: An Analysis of Hate Speech Annotations by Open-Source Large Language Models

---

Irene García-Espantaleón Artal

Supervisor: Iñaki Úcar Marqués

Madrid, June 2025

Abstract

This thesis investigates the use of open-source large language models (LLMs) as annotators in the task of hate speech detection. Rather than treating human-labeled data as an infallible gold standard, the study explores disagreement, both among models and between models and humans, as a signal of ambiguity or complexity. Five open-source LLMs were used to classify nearly 15,000 texts, and their outputs were analyzed to identify patterns of consensus and divergence. Entropy-based metrics captured the degree of inter-model disagreement, while regression analyses examined how textual features influenced annotation variability. Results show that certain features, such as stylized language or vague hostility, systematically predict model disagreement and misalignment with human labels. These findings suggest that LLM consensus does not always indicate accuracy and underscore the importance of embracing label variation in subjective annotation tasks.

A GitHub repository containing the replication materials is available at https://github.com/igarciaespantaleon/beyond-agreement-masters-thesis.

Keywords

Large Language Models · open-source · automated text annotation · inter-rater agreement · label variation · hate speech detection

Table of Contents

List of Figures

List of Tables

Introduction

Large Language Models (LLMs) are increasingly being employed for text annotation across a wide range of tasks, offering advantages in terms of cost efficiency and consistency. Their performance is typically evaluated by benchmarking their annotations against human-labeled data, which is treated as the gold standard. However, this evaluation framework often assumes a single correct label for each instance, treating variation among human annotators as noise. Disagreements are commonly resolved through majority voting or discarded altogether, reducing human judgment to a binary reference point, with which the model either agrees (right) or disagrees (wrong). Yet, label variation should not be viewed merely as a source of error. On the contrary, disagreement among annotators can offer valuable insights into the complexity or inherent ambiguity of a given instance.

In this experiment, I investigate the application of five open-source LLMs to the task of hate speech detection. My objective is to gain insight into how these models label text and how their outputs vary across instances. Rather than treating inter-coder disagreement as failure, I interpret it as an indicator of linguistic, contextual or sociocultural complexity. Additionally, I compare the consensus outputs of the five models to human-provided annotations, with the aim of identifying systematic differences between automated and human labeling practices.

The remaining pages of this thesis are structured as follows. The next section reviews prior work on human annotation practices, outlining the role of disagreement in labeled datasets and the emergence of LLMs as text annotators. This is followed by a detailed explanation of the methodology, including dataset selection, model configuration, and experimental design. The results section presents findings on inter-model agreement and divergence from human annotations, supported by statistical modeling. The discussion section interprets these findings, focusing on how specific content characteristics affect model behavior, and considers what these patterns reveal about the interpretive limits of LLMs. The final sections address limitations, potential applications, and directions for future research, concluding with reflections on the broader role of LLMs in handling subjective classification tasks.

Background and related work

Text annotation and the assumption of ground truth

Coding complex concepts from text corpora is one of the core tasks in both quantitative and qualitative social science research (Zhang & Lin, 2024). The need for large volumes of labeled data is becoming more pressing for NLP tasks and AI development (Marchal et al., 2022; Plank, 2022). Traditionally, researchers have relied on human annotators, either domain experts or crowd workers, to carry out this work. However, manual annotation presents several challenges, one of which is the influence of human subjectivity (Li et al., 2023). In the text annotation workflow, disagreement over which label corresponds to each piece of text is not uncommon, since natural language expressions are often interpretable. Denton et al. use the term *annotator positionality* to describe "how annotator social identity shapes their understanding of the world" (2021, p. 2). Despite this, labeling work is usually based on the idea that there is one correct label for each item, a *ground-truth* which "can be ascertained by as few as three human data annotators" (Smart et al., 2024, p. 9). Any variation is thus regarded as an error, a source of noise that must be "smoothed out" (Zhang & de Marneffe, 2021).

This assumption has been increasingly recognized as a "convenient idealization" (Leonardelli et al., 2023): in most cases, reconciling different subjective views is simply not feasible. Smart et al. (2024) challenge this assumption that subjectivity in labeling can be resolved through consensus mechanisms such as majority voting. They argue that relying on a single "ground truth" is inappropriate in annotation tasks involving semantic nuance or cross-cultural sensitivity, as it obscures genuine variation in human interpretation. Instead, such variation should be acknowledged and examined as part of the phenomenon under study. This critique aligns with Törnberg's (2024) observation that text annotation is rarely a purely technical task. Rather, it requires interpretive work, the definition of conceptual boundaries, and negotiation of meaning, all of which introduce inevitable subjectivity. Ensuring consistency and transparency in how this subjectivity is handled is more realistic and honest than pretending it can be eliminated.

Other scholars have proposed a shift in how annotation data is conceptualized. Plank (2022) introduces the term *human label variation* to describe the legitimate diversity of interpretations that often arises in annotation tasks. Preserving this variation in the resulting labeled datasets allows models to learn not only from dominant interpretations,

but also from contested or uncertain ones. As a result, several researchers advocate for annotation strategies that retain label distributions, such as multi-label or probabilistic schemes, or for models that are trained to predict human disagreement (Marchal et al., 2022; Uma et al., 2021; Baan et al., 2022; Zhang & de Marneffe, 2021), challenging the notion that dataset consistency must come at the expense of representativity and human interpretive diversity.

LLMs as text annotators

Manual labeling is also constrained by the high cost and time investment required for recruiting and training annotators, and labeling inconsistency due to fatigue or lapses in attention. In this context, LLMs emerge as a compelling alternative. These models have demonstrated strong performance in annotation tasks without any manually labeled training data, using natural language prompts (Törnberg, 2024). Their ability to generalize across tasks stems from the vast and diverse datasets they are trained on, allowing them to develop sophisticated representations of language and context. Moreover, they are often perceived as more objective than humans, although it is important to acknowledge that LLMs are not free from bias; prior studies have demonstrated that their outputs can reflect and amplify societal stereotypes embedded in their training data (Belal et al., 2023).

LLMs have been evaluated as annotators on a wide range of tasks and topics, such as stance, relevance and topic classification, hate speech detection, political ideology identification, named entity recognition, etc. (Huang et al., 2023; Yu et al., 2023; Gilardi et al., 2023). In their experiments, many authors choose proprietary LLMs (namely, OpenAi's GPT models) based on their popularity, sophistication or performance. Other works compare ChatGPT's performance to lexicon-based unsupervised methods (Belal et al., 2023), or encoder-based transformer models from the BERT family (Kuzman et al., 2023), in diverse annotation tasks. However, concerns have been expressed regarding the lack of transparency, risk of data leakage, and challenges to reproducibility associated with closed-source models (Spirling, 2023; Törnberg, 2024). Some authors (Yu et al., 2023; Alizadeh et al., 2024) compare ChatGPT and open-source models and conclude that open models only achieve their closed counterpart's levels of accuracy through fine-tuning.

These experiments have yielded mixed findings: while some studies suggest that LLMs could serve as substitutes for human annotators, others have found that their performance aligns more closely with that of lower-quality crowd workers (Ostyakova, 2025). Wang et al. (2021) and Li et al. (2023) emphasize that while LLMs may perform well on relatively simple or low-stakes annotation tasks, their reliability decreases for more complex or high-stakes scenarios, where subjective nuance and potential consequences demand greater care. In light of this, some authors advocate for a task-sensitive division of labor between LLMs and human annotators.

Lastly, although several studies emphasize the promise of LLMs as ready-to-use tools for data annotation, often requiring only minimal prompt design and no extensive manual labeling (Li et al., 2023; Huang et al., 2023); others caution against assuming that these models are universally suited to all tasks. Instead, they argue for more context-aware approaches, highlighting the need to align model capabilities and prompt strategies with the specific demands of each annotation scenario (Weber & Reichardt, 2024).

Regardless of how many or which LLMs are compared, most papers focus on the individual performance of each model (or model-and-prompt combination). Nearly all studies use metrics such as accuracy or F1 scores to assess their outcomes (Ollion et al., 2024; Pavlovic & Poesio, 2024), comparing model-generated labels to human annotations, taken as the gold standard. Traditionally, the gold-label paradigm assumes binary correctness: model predictions are evaluated as either right or wrong (Baan et al., 2022), relying on the single-truth assumption discussed above.

Several studies have proposed alternative experimental designs to address the limitations of accuracy-based evaluation and reliance on gold labels. Conversely, these approaches focus on how well model outputs reflect *human disagreement patterns* or internal model uncertainty. For instance, Lee et al. (2023) prompt different LLMs with natural language inference (NLI) tasks and compare the probability distributions of the models' labels to distributions of human disagreement from multi-annotated datasets. Their findings show that LLMs frequently fail to align with the diversity of human judgments. Similarly, Pavlovic and Poesio (2024) instruct GPT to output class probability distributions in opinion labeling tasks, revealing that the model's outputs exhibit higher average entropy and often diverge from empirically observed human opinion distributions. Other researchers have proposed directly eliciting confidence scores from models, as a proxy for model uncertainty in settings where no gold-standard annotations exist (Li et al.,

2023). These approaches collectively reflect a growing interest in capturing the *degree* and *nature* of model uncertainty, rather than enforcing binary correctness against potentially oversimplified ground truths.

To the best of my knowledge, no prior research has been devoted to examining how the features of the labeled texts (both in form and content) may influence the degree of agreement among LLMs as annotators, or how those same features might account for systematic divergences between model and human annotations. Understanding these dynamics could offer valuable insight into LLM labeling behavior. This forms the basis for the strategy adopted in the present study.

Methodology

Task selection: hate speech detection

Hate speech detection was selected for this experiment because it involves a degree of subjectivity, aligning with the project's focus on disagreement as a potential indicator of ambiguity or complexity. As Lee et al. (2023) note, tracing disagreement in a task like hate speech detection is not only relevant, but necessary, as it is a signal of diverse interpretations and opinions that we care about capturing. Other works highlight that hate speech detection tasks typically only achieve a moderate inter-rater agreement for human labelers (Toliyat et al., 2025). Schmidt & Wiegand (2017) report that hate speech detection efforts are marked by the absence of a commonly accepted definition of the concept and often vague annotation guidelines, making this an annotation task where relying on human labels is even more questionable.

In addition, hate speech detection is a relatively common use case for LLMs, which suggests that most models have been exposed to similar examples during pretraining. Notwithstanding, it should be emphasized that social media often hosts *implicit* forms of hate speech, making this task even more nuanced and context-dependent (Huang et al., 2023). Finally, the availability of large-scale, high-quality public datasets in this domain made it a practical choice.

While I acknowledge that hate speech detection is inherently subjective and shaped by both the author's intent and the reader's perspective, it often takes place in public, highly interactive spaces like social media, where harmful content can influence beliefs and behaviors (Akhtar et al., 2021). Therefore, some degree of regulation is necessary to

ensure user safety and protection against harmful content. In this context, an objective layer, though imperfect, becomes central.

Dataset and feature engineering

The dataset used in this project is drawn from an experiment by Vidgen et al. (2021). It contains 41,144 short texts synthetically authored by human annotators to train hate speech detection models, inspired by real-world hate sites. Each text is labeled for hate type and target group, with an additional set of "perturbations" designed to challenge model performance. Although these synthetic examples reduce some of the natural ambiguity found in real-world data, they introduce controlled variation and difficulty, and ensure a more balanced representation of hate types and targets than typically found in organic corpora.

For this study, I selected a subset of 14,954 original texts, excluding perturbations and the first annotation round (which lacked target labels). Texts targeting elderly people (n = 23) were removed due to low representation, as were items labeled with type *support*, which expressed endorsement of Nazi ideology and were qualitatively distinct from the other categories. The original target taxonomy was aggregated into broader groups (*ethnicity, gender, LGBTQ, foreign, religion, disability,* and *class*) to ensure adequate sample sizes and clearer interpretation (see Appendix A for full mapping).

It is worth noting that the *type* variable is multiclass, with each hateful text assigned a single label, while the *target* variable is multilabel, as a given message may direct hate toward multiple social groups. These metadata fields were only recorded for texts labeled as *hateful*; in the case of non-hateful messages, the target and type labels are *none*.

The four hate types used in the dataset were defined by the original authors as follows:

- **Derogation**: Content that attacks, demeans, or insults a group.
- **Animosity**: Implicit or subtle abuse against a group.
- **Threatening**: Language that expresses or supports harm against a group or its members.
- **Dehumanization**: Content that treats people as less than human, often through animalistic or disease metaphors.

These categories are not equally distributed across the dataset. As shown in Figure 1, *derogation* is the most common type across all target groups, while *threatening* and *dehumanization* are much rarer. On the other hand, *ethnicity* is the most targeted group, whereas *disability* and *class* appear far less frequently.

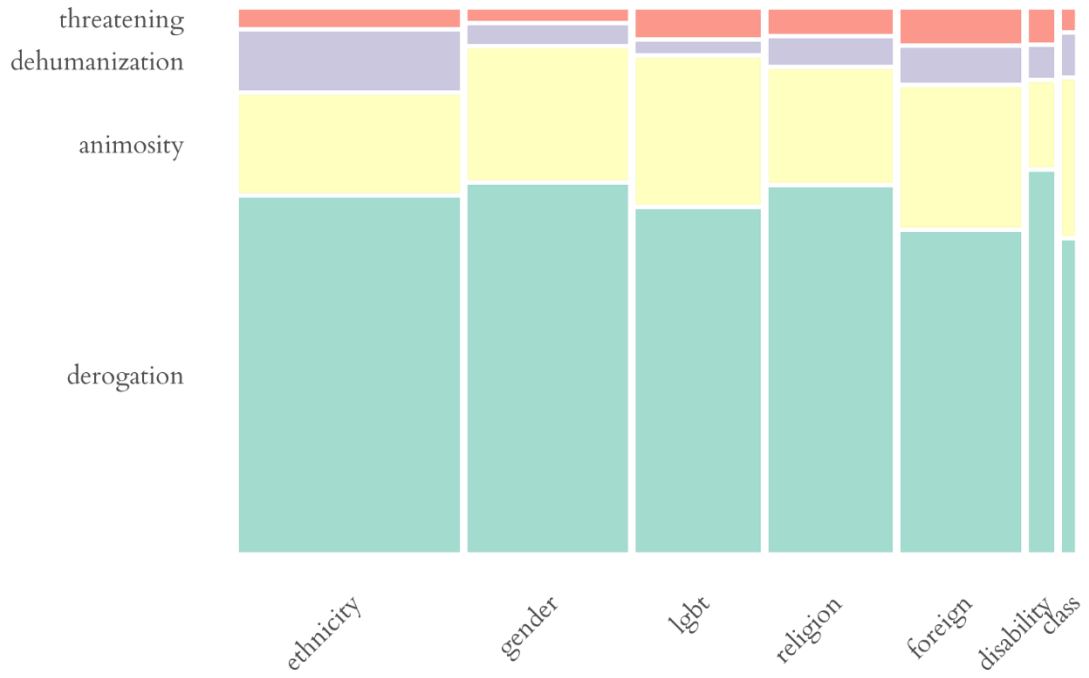

Figure 1. Hate speech typology by targeted group

Regarding feature creation, basic surface-level metrics were computed, including the number of words per text, introduced as a control variable to account for the potential influence of text length on model disagreement. A binary variable was created to flag the 310 messages containing three or more repeated letters (e.g., *trrraaaashhh*). Another binary feature was designed to detect leetspeak (n = 254), a type of writing that uses alphanumeric substitutions (e.g., *w0m3n*), intended to bypass content regulation on social media. Syntactic complexity was approximated by identifying the presence of subordinate clauses. Using dependency parsing via spaCy, the number of adverbial, complement, and clausal modifiers per text was counted. Finally, the variables of character and target count, although useful for data exploration, had to be dropped from regressions to avoid collinearity.

Model selection and setup

The use of open-sourced, self-hosted large language models in this study reflects the author's methodological and ethical preferences. As many scholars have pointed out, open-source systems promote a research culture rooted in public collaboration and epistemic openness (Spirling, 2023; Törnberg, 2024). This capacity for scrutiny, though not exploited to inspect model architecture in this project, supports a more transparent and reproducible analysis pipeline.

More pragmatically, this decision answered to the demands of the experimental setup. The size of the dataset would have made web-based interaction with proprietary LLMs highly inefficient, and API access to the models can be quite costly (Liesenfeld et al., 2023), especially considering that, in order to assess disagreement, my goal was to compare outputs across as many models as possible. In contrast, open-source LLMs can be self-hosted locally and are free of charge.

Furthermore, the experiment required flexibility to iterate across multiple prompting strategies and model types, which I was able to carry without concern for access or budget restrictions. In addition, using locally deployed models grants full control over parameter-setting and version. Closed-source systems are often updated without notice, with no access to the internal configurations that might affect annotation behavior (Spirling, 2023).

The iterative process mentioned began with a series of exploratory tests on toy data. Initially, I evaluated several zero-shot classification LLMs, but found their outputs to be generally less accurate than those of text generation models. Within this category, instruction-tuned and chat-oriented were preferred. Although these models were not specifically trained for hate speech classification, their instruction-following capabilities make them more responsive to annotation prompts and more consistent in their outputs. I conducted an initial comparative run of seventeen open-sourced LLMs, ranging from 1 to 12 billions of parameters, on a small subset of hate speech data. Due to hardware limitations, I restricted the final set to the five models that yielded the most accurate and consistent annotations:

- *mistralai/Mistral-7B-Instruct-v0.3*
- *deepseek-ai/deepseek-llm-7b-chat*

- *Qwen/Qwen3-8B*
- *01-ai/Yi-1.5-9B-Chat*
- *microsoft/Phi-3-mini-4k-instruct*

Seeking to ensure diversity among the final set, I avoided including more than one version of each family to maximize architectural and training differences, and enhance the interpretability of inter-model disagreement.

The following prompt format was adopted upon experimentation with various formulations:

```
"Classify the following text as EITHER hate speech OR not hate speech.\n\n",
"Text: ", txt, "\n\n",
"Complete with ONLY a label: 'hate speech' OR 'not hate speech'. DO NOT provide any explanation.\n\n",
"Answer: This text is classified as"
```

Despite the explicit instruction to return only the label, most models still generated verbose outputs. However, this format proved effective in making the expected label more easily extractable from the generated text, reducing the need for post-processing.

Although some authors (Törnberg, 2024) recommend tailoring prompts to each model, I opted for a common prompt. This helps preserve the validity of inter-model comparisons, since the aim is to isolate the influence of model-specific characteristics on classification behavior and minimize the effect of the prompt's phrasing.

Model outputs were generated using the default settings, which correspond to greedy decoding. This approach deterministically selects the highest-probability token at each step and was deemed appropriate for a binary classification task. This choice makes the generation process fully reproducible and consistent (Ziems et al., 2024) and is argued to enhance the overall quality of results (Alizadeh et al., 2024). Generation was limited to a maximum of 10 new tokens. These settings allow for the possibility of meaningful variation and intercoder disagreement while minimizing the risk of off-topic or hallucinated responses that could obscure the analysis.

It should be made clear that I chose to work with pretrained large language models without any fine-tuning on human-annotated datasets. This decision was intentional: fine-

tuning would have meant using human annotation as benchmark, which defeats the purpose of this study. I was not interested in aligning model predictions with any particular human-labeled dataset's criteria or my own, since this would have limited the opportunity to observe meaningful disagreement (while still understanding that these models are themselves trained by humans on human-generated data). For the same reason, I chose zero-shot instead of few-shot prompting.

All five models were used to annotate the subset of nearly 15,000 texts drawn from the original dataset. Finally, all annotations were run on a university-hosted server.

Strategy

The strategy of this study involves a series of classification and regression tasks aimed at understanding the factors associated with disagreement in hate speech annotations. I modeled two outcomes derived from the annotation data.

The first outcome is an ordinal variable, disagreement_level, which captures the degree of disagreement among LLMs for each text. This variable was computed based on the entropy of the five model predictions: an entropy score of zero was labeled as *agreement*, a value around 0.71 corresponded to *mild disagreement* (4-1 splits), and a score of 0.97, to *severe disagreement* (3–2 splits).

Secondly, I defined a binary outcome variable, llm_human_mismatch, to capture cases where the five LLMs reached full consensus on a label that disagreed with the majority human annotation. This is taken as a sign of how models systematically annotate differently from human labelers.

Each of these outcomes was modeled using regression techniques appropriate to the task: ordinal logistic regression for disagreement_level and standard logistic regression for llm_human_mismatch. In addition to these interpretable models, I also explored more complex approaches such as random forests and gradient boosting, better suited for capturing non-linearities. However, these models did not provide meaningful improvements in explanatory power or prediction quality. Since the goal of the experiment is to understand which variables are associated with disagreement and misalignment, not to optimize prediction, I chose to retain the simpler, interpretable models for the main analyses. I explored a range of interaction terms, but ultimately

excluded them due to limited effect and theoretical justification. Similarly, I tested quadratic terms for the continuous predictors, but neither improved model fit.

Results

Agreement among models

The five LLMs predicted varying proportions of hate speech for this dataset, from highest to lowest: 64.58% for Mistral-7B-Instruct-v0.3; 52.17% for Qwen3-8B; 51.98% for deepseek-llm-7b-chat; 50.88% for Yi-1.5-9B-Chat; and 41.85% for Phi-3-mini-4k-instruct.

Krippendorff's alpha for the five models was 0.594, indicating a moderate level of agreement, but notable given the difficulty of the task. Full agreement was observed in 58.6% of cases; the rest showed moderate (22.7%) or severe (18.6%) disagreement. Qwen3-8B was the model that most often agreed with the majority label (90.8% of times), followed by Yi-1.5-9B-Chat (90.3% of times); the most dissenting model was Phi-3-mini-4k-instruct (agreed with the majority label 84.82% of times).

The model ensemble showed similar disagreement patterns for texts labeled as "hate speech" and "not hate speech" by human annotators. However, the proportion of agreement, moderate and severe disagreement varied across the different text features measured. To investigate this, I used an ordinal logistic regression predicting *level of disagreement* (none, mild, severe). The initial model included all candidate predictors, but stepwise selection via AIC resulted in a more parsimonious specification, excluding class, foreign status, and subordinate clause count.

Several features emerged as significant predictors. Posts labeled as animosity had markedly increased odds of higher disagreement levels, while threatening and dehumanizing content were associated with reduced odds. Mentions of religion, ethnicity, LGBT and disability also predicted lower disagreement. In contrast, lexical markers such as leetspeak and repeated letters were linked to higher disagreement, as were mentions of gender. Text length had a small but significant negative effect. Coefficients are reported as odds ratios in Figure 2, alongside 95% confidence intervals. In this context, each odds ratio represents the multiplicative change in the odds of receiving a higher disagreement label (from agreement to mild or from mild to severe), associated with a one-unit increase in the corresponding predictor.

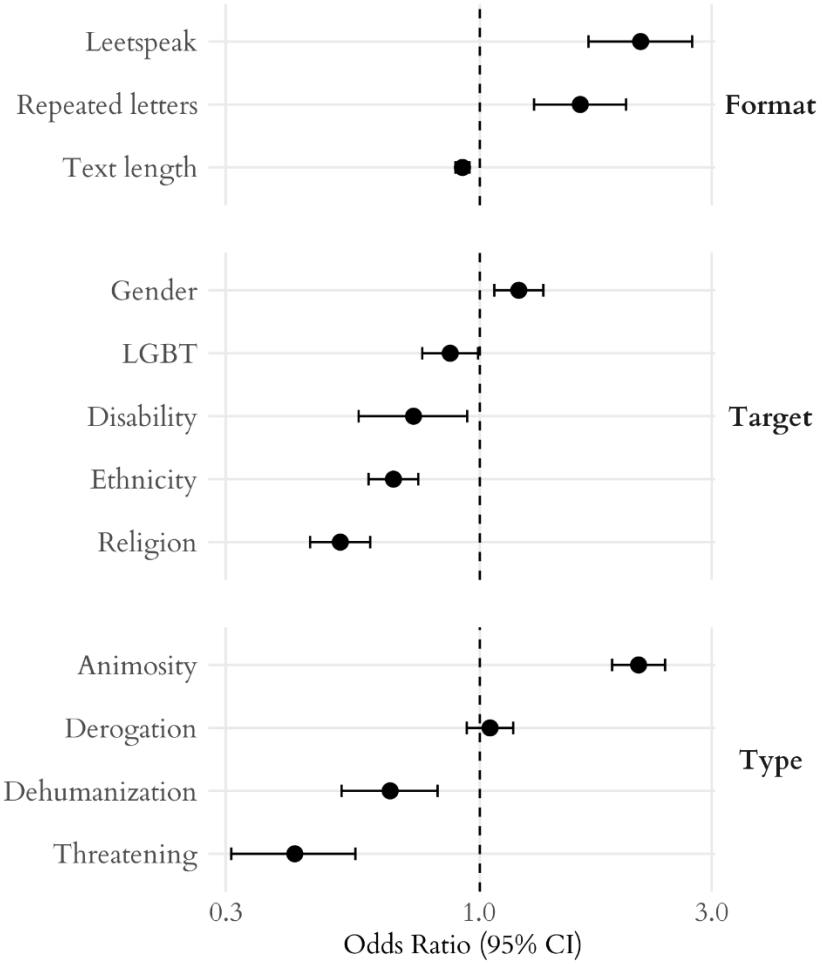

Figure 2. Ordinal logistic regression results for level of disagreement

Although the model's pseudo-R^2 was modest (McFadden's $R^2 = 0.018$), it improved on the null model, indicating that these features capture some meaningful variation in model disagreement.

To further evaluate model performance, predicted disagreement levels were compared against the actual labels using a confusion matrix. The results show that the model systematically predicts agreement: most 8,658 actual agreement cases were correctly classified, but it failed to identify any instances of mild disagreement, and most cases of severe disagreement were also misclassified as agreement. Only 81 texts with severe disagreement were correctly predicted. This skew suggests that the model tends to default to the majority class, potentially due to class imbalance or limited discriminative power.

		Actual		
		Agreement	**Mild**	**Severe**
Predicted	**Agreement**	8,658	3,294	2,699
	Mild	0	0	0
	Severe	81	98	81

Table 1. Confusion matrix for predictions of disagreement level

A binary logistic regression model was also trained to distinguish between agreement and any form of disagreement, revealing similar patterns; its coefficients and confusion matrix are provided in Appendix B.

LLM-human divergence

For the second part of this analysis, I drew a subset of the rows in which the five LLMs unanimously agreed on the label, be it “hate speech” or “not hate speech”. This was intended to capture a sample of “easy” cases for the models, examples where the classification was clearest, minimizing the ambiguity or complexity of the text.

For this subset of 8,739 texts, Krippendorff’s alpha between the model ensemble and the human annotations was 0.611, which represents a moderate level of agreement. Given that this subset includes only the most model-certain predictions, the alpha value suggests that model consensus does not necessarily imply human-model alignment, underscoring the complexity and diversity of this interpretative task. I noticed that, when models “agree to disagree” with human annotators, they're more likely to label as hate speech a text that humans considered not hate speech, than the other way around, which is consistent with previous research.

To assess which factors predict divergence between model and human annotations, I fitted a logistic regression using the same set of features as before. Several predictors were statistically significant: animosity texts had the highest odds of disagreement, while threatening content and references to religion, ethnicity, or class were linked to lower disagreement. Stylized features such as leetspeak and repeated letters also increased the likelihood of mismatch.

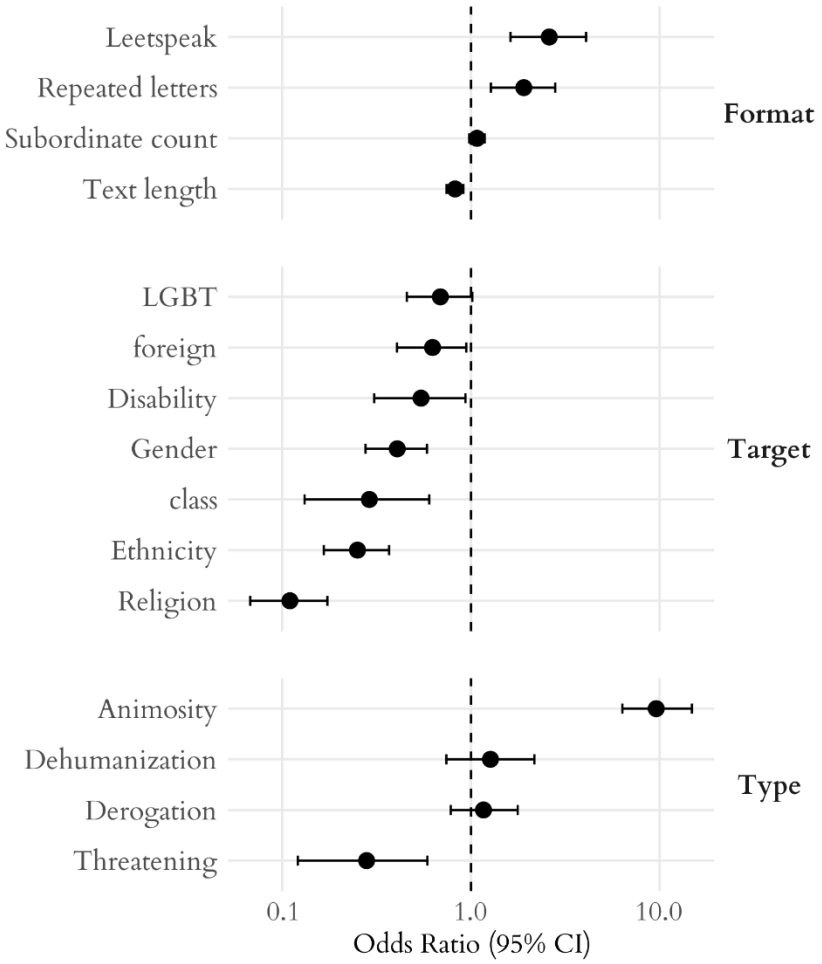

Figure 3. Logistic regression results for LLM-human mismatch

The logistic regression model predicting LLM-human mismatch achieved a McFadden's pseudo-R^2 of 0.10, indicating a modest but meaningful improvement over the null model. This suggests that the included features explain a nontrivial portion of the variation in disagreement between model and human annotations, especially compared to the lower value observed for the ordinal regression.

Again, a confusion matrix was examined to assess the predictive accuracy of the logistic regression model for *llm_human_mismatch*. The model correctly identified 6,897 cases of agreement between LLMs and human annotators, as well as 230 cases of mismatch. However, it also misclassified 1,467 mismatches as agreements and falsely flagged 145 agreement cases as mismatches. These results indicate a strong tendency to predict alignment with human labels, likely reflecting the majority distribution in the data. While the model captures some signals associated with mismatch, its performance remains

limited: most disagreement cases go undetected, and the overall ability to distinguish between alignment and divergence is modest.

		Actual	
		Agreement	**Mismatch**
Predicted	**Agreement**	6,897	1,467
	Mismatch	145	230

Table 2. Confusion matrix for predictions of LLM-human mismatch

Discussion

Together, these findings suggest that both the explicitness of the content and mention to certain target groups shape model interpretability and agreement. Threatening and dehumanizing language appears to elicit strong consensus across models and between models and human annotators, likely because such language relies on more overt and recognizable cues. In contrast, animosity, a more ambiguous and context-dependent label, consistently led to higher disagreement. These patterns reinforce existing research suggesting that explicit hate is easier to detect, while implicit or stylized expressions remain a challenge for automated systems (Huang et al., 2023).

Regarding target groups, references to religion and ethnicity were strongly associated with more consistent labeling in both model-model and model-human comparisons, suggesting these categories benefit from more defined social or linguistic boundaries. Other targets, such as disability and LGBT, also showed negative associations with disagreement, though their effects were weaker. Mentions of class and foreign status were only significant when predicting model-human disagreement, both with negative coefficients indicating higher alignment. Notably, content referencing gender was associated with higher disagreement among models, yet greater alignment with human annotations.

The negative effect of text length on disagreement suggests that increased context helps disambiguate meaning, making model predictions more consistent and more aligned with human judgments. In contrast, the presence of leetspeak and repeated letters was linked to higher disagreement in both settings, likely by distorting recognizable cues used for hate speech detection. Finally, subordinate clause count was not a significant predictor in

either model, indicating that syntactic complexity alone does not strongly influence disagreement patterns in this context.

Ultimately, although the models were not developed with predictive accuracy as their primary goal, the confusion matrices offer useful diagnostic information. In both the ordinal and binary settings, predictions were skewed toward agreement, indicating a systematic bias. This tendency limits their ability to detect more subtle forms of dissent, but it suggests that the selected features are better at capturing strong consensus than more ambiguous instances.

Limitations and future work

This experiment faced several limitations. The number of explanatory variables was limited and did not capture the complexity of each instance, failing to explain when or why models converge in their predictions. Although the study explicitly avoided using human annotations as a benchmark for evaluating model performance, some reliance on human-labeled data was unavoidable. The manually assigned labels for *type* and *target* were treated as fixed throughout the analysis. In particular, the multiclass *type* label may reintroduce assumptions of single-truth classification, and the *target* variable was only available for texts annotated as hateful, preventing a fuller understanding of disagreement in non-hateful instances that still reference identifiable groups. Furthermore, the aggregation of target groups could have been structured differently, potentially influencing observed patterns.

The selection of surface-level textual features was constrained by the scope of the project; richer or more context-aware features would likely improve explanatory power. Additionally, the study focused exclusively on English-language texts. This choice reflects both practical constraints and the need for direct interpretation of the texts by the researcher. However, it limits the generalizability of the findings, especially given documented performance gaps in LLMs when applied to other languages (Ollion et al., 2024).

The study examined only five open-source models due to hardware constraints. These were evaluated in their base form without fine-tuning, which may have affected performance: some of the observed variation could reflect model limitations or instability, rather than meaningful disagreement. Finally, comparisons were drawn between a specific subset of models and a specific group of annotators. Since the study challenges

the idea of a single ground truth, the generalizability of any observed alignment or misalignment between humans and models should be interpreted with caution.

Despite these limitations, the experiment offers insights into how pretrained LLMs handle subjectivity and ambiguity in classification tasks and opens several opportunities for further research.

Firstly, a deeper investigation into the architectural and training differences among models, including their parameters, weights, training data composition, and safety alignment mechanisms, could offer valuable insights into why different models annotate text the way they do. Future research could also examine the impact of prompt variation and model versioning, which might give way to different reasoning patterns. Additionally, exploring other annotation tasks, particularly multi-label classifications, may yield further insight. Unlike binary tasks such as hate speech detection, multi-label tasks allow for overlapping categories, offering a richer context in which to observe model uncertainty.

Furthermore, future studies could collect more informative features by eliciting additional annotations directly from the models (such as *type* and *target* used here, but also irony detection, topic identification, etc.). Other metrics, like Flesch-Kincaid grade levels, could be computed to account for text readability. The presence of slurs or explicit abusive language has also been shown to distinguish human annotations from LLMs': the latter are likely to "overfit on the use of slurs and pejorative terms, treating them as hateful irrespective of how they are used" (Vidgen et al., 2021, p. 1668). Expanding the scope of data collection to include multiple platforms could also be revealing, since tone, style and social norms might really differ from one social site to another.

Another important field involves comparing model disagreement with human annotator disagreement, where suitable multi-annotated datasets are available. While this study intentionally avoided using human labels as ground truth, aligning model variation with documented human disagreement could shed light on whether LLMs are approximating human interpretive diversity or simply introducing noise.

Lastly, future work should extend beyond English-language content. Exploring additional languages and cultural settings would allow for a more comprehensive understanding of how models perform across diverse political, economic, and sociocultural environments, bridging a gap in LLM literature that cannot be overlooked (Schmidt & Wiegand, 2017).

Conclusion

This study has examined how large language models behave as annotators in the task of hate speech detection, focusing not on performance metrics, but on patterns of disagreement, both among models and between models and human annotators. The results suggest that certain linguistic and semantic features systematically influence agreement, with more explicit forms and certain target groups of hate speech yielding greater consensus. Conversely, ambiguous and oddly formatted messages appear to challenge model coherence, revealing the limits of current LLMs in interpreting disguised or implicit forms of harmful speech.

The findings also demonstrate that model consensus does not necessarily imply alignment with human judgments, reinforcing the notion that annotation is not merely a technical task, but one defined by interpretive complexity. While LLMs offer powerful, scalable tools for data annotation, they do not eliminate the need for human oversight, particularly in tasks marked by subjectivity and sociocultural nuance. Despite advancements in model design and prompting strategies, human annotation remains essential to ground, interpret, and critically evaluate the judgments these systems make.

As suggested by prior research, rather than replacing human annotators, LLMs may best be used to augment them, supporting hybrid annotation pipelines. Future research should continue exploring complementary strategies, especially in domains where label variation is most insightful and clarity, fairness, and accountability are crucial.

References

Akhtar, S., Basile, V., & Patti, V. (2021). *Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection* (arXiv:2106.15896). arXiv. https://doi.org/10.48550/arXiv.2106.15896

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). *Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning* (arXiv:2307.02179). arXiv. https://doi.org/10.48550/arXiv.2307.02179

Baan, J., Aziz, W., Plank, B., & Fernández, R. (2022). *Stop Measuring Calibration When Humans Disagree* (arXiv:2210.16133). arXiv. https://doi.org/10.48550/arXiv.2210.16133

Belal, M., She, J., & Wong, S. (2023). *Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis* (arXiv:2306.17177). arXiv. https://doi.org/10.48550/arXiv.2306.17177

Denton, R., Díaz, M., Kivlichan, I., Prabhakaran, V., & Rosen, R. (2021). *Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation* (arXiv:2112.04554). arXiv. https://doi.org/10.48550/arXiv.2112.04554

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *Companion Proceedings of the ACM Web Conference 2023*, 294–297. https://doi.org/10.1145/3543873.3587368

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification* (arXiv:2303.03953). arXiv. https://doi.org/10.48550/arXiv.2303.03953

Lee, N., An, N., & Thorne, J. (2023). Can Large Language Models Capture Dissenting Human Voices? *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4569–4585. https://doi.org/10.18653/v1/2023.emnlp-main.278

Leonardelli, E., Abercrombie, G., Almanea, D., Basile, V., Fornaciari, T., Plank, B., Rieser, V., Uma, A., & Poesio, M. (2023). SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2304–2318. https://doi.org/10.18653/v1/2023.semeval-1.314

Li, M., Shi, T., Ziems, C., Kan, M.-Y., Chen, N., Liu, Z., & Yang, D. (2023). CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1487–1505. https://doi.org/10.18653/v1/2023.emnlp-main.92

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). *Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators*. 1–6. https://doi.org/10.1145/3571884.3604316

Marchal, M., Scholman, M., Yung, F., & Demberg, V. (2022). *Establishing annotation quality in multi-label annotations*. *Proceedings of the 29th International Conference on Computational Linguistics*, 3659–3668. International Committee on Computational Linguistics. Available at: https://aclanthology.org/2022.coling-1.322/

Ollion, É., Shen, R., Macanovic, A., & Chatelain, A. (2024). The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, *6*(1), 4–5. https://doi.org/10.1038/s42256-023-00783-6

Ostyakova, L., Mikhailova, A., & Konovalov, V. (2025). Redefining Annotation Practices: Leveraging Large Language Models for Discourse Annotation. In *Analysis of Images, Social Networks and Texts: 12th International Conference, AIST 2024*, 131–147. https://doi.org/10.1007/978-3-031-88036-0_7

Pavlovic, M., & Poesio, M. (2024). *The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.01299

Plank, B. (2022). The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Smart, A., Wang, D., Monk, E., Díaz, M., Kasirzadeh, A., Liemt, E. V., & Schmer-Galunder, S. (2024). *Discipline and Label: A WEIRD Genealogy and Social Theory of Data Annotation* (arXiv:2402.06811). arXiv. https://doi.org/10.48550/arXiv.2402.06811

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, *616*(7957), 413. https://doi.org/10.1038/d41586-023-01295-4

Toliyat, A., Etemadpour, R., & Filatova, E. (2025). Inter-Annotator Agreement and Its Reflection in LLMs and Responsible AI. *The International FLAIRS Conference Proceedings*. https://doi.org/10.32473/flairs.38.1.139049

Törnberg, P. (2024). *Best Practices for Text Annotation with Large Language Models* (arXiv:2402.05129). arXiv. https://doi.org/10.48550/arXiv.2402.05129

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, *72*, 1385–1470. https://doi.org/10.1613/jair.1.12752

Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2021). Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. *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)*, 1667–1682. https://doi.org/10.18653/v1/2021.acl-long.132

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). Want To Reduce Labeling Cost? GPT-3 Can Help. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205. https://doi.org/10.18653/v1/2021.findings-emnlp.354

Weber, M., & Reichardt, M. (2024). *Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2401.00284

Yu, H., Yang, Z., Pelrine, K., Godbout, J. F., & Rabbany, R. (2023). *Open, Closed, or Small Language Models for Text Classification?* (arXiv:2308.10092). arXiv. https://doi.org/10.48550/arXiv.2308.10092

Zhang, X. F., & De Marneffe, M. C. (2021). Identifying inherent disagreement in natural language inference. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4908–4915. https://doi.org/10.18653/v1/2021.naacl-main.390

Zhang, Y., & Lin, H. (2024). *The Risks of Using Large Language Models for Text Annotation in Social Science Research*. https://doi.org/10.31235/osf.io/79qu8

APPENDIX A

Table A1. Mapping of original target labels to aggregated categories

Aggregated category	Original labels
Gender	wom, indig.wom, bla.wom, asi.wom, gay.wom, mus.wom
Ethnicity	mixed.race, ethnic.minority, indig, indig.wom, non.white, bla, bla.wom, bla.man, african, asi, asi.wom, asi.man, arab, hispanic, trav
Foreign	asi.east, asi.south, asi.chin, asi.pak, eastern.europe, russian, pol, other.national, immig, asylum, ref, for
LGBT	trans, gendermin, bis, gay, gay.man, gay.wom, lgbtq
Religion	jew, mus, mus.wom
Disability	dis
Class	wc

APPENDIX B

Figure B1. Logistic regression results for disagreement among LLMs

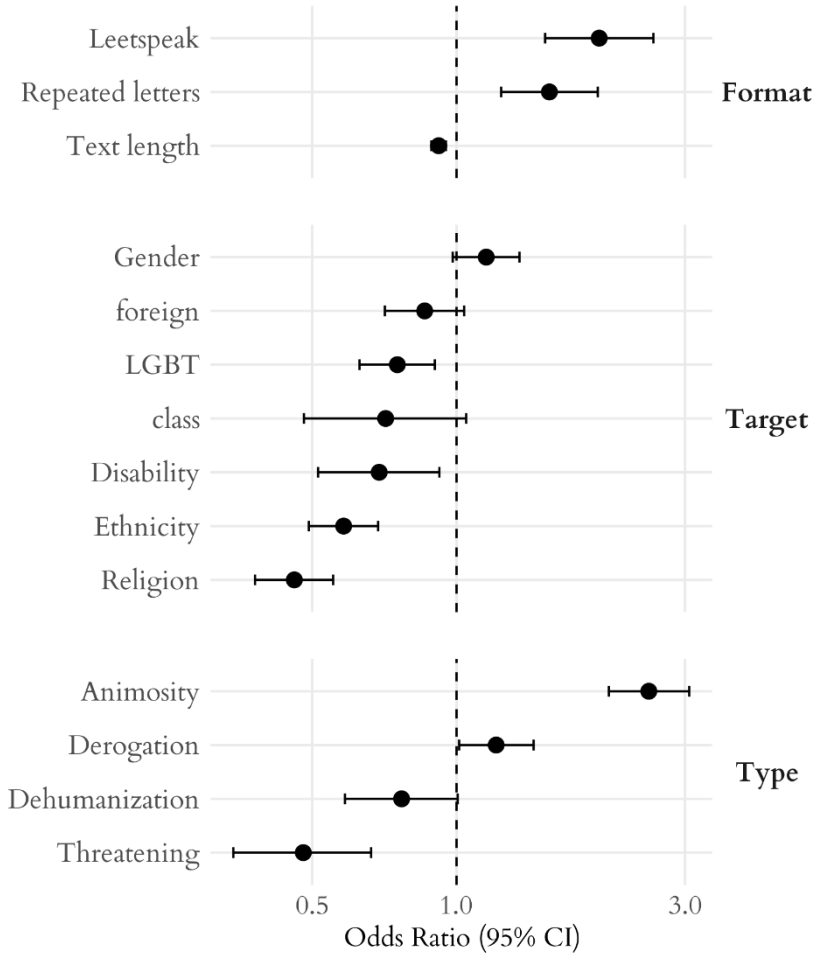

Table B1. Confusion matrix for predictions of disagreement among LLMs

		Actual	
		Agreement	**Disagreement**
Predicted	**Agreement**	7,860	5,080
	Disagreement	879	1,092

APPENDIX C

DECLARACIÓN DE USO DE IA GENERATIVA EN EL TRABAJO DE FIN DE MÁSTER

He usado IA Generativa en este trabajo

Marca lo que corresponda:

SI	NO

Si has marcado SI, completa las siguientes 3 partes de este documento:

Parte 1: reflexión sobre comportamiento ético y responsable

Ten presente que el uso de IA Generativa conlleva unos riesgos y puede generar una serie de consecuencias que afecten a la integridad moral de tu actuación con ella. Por eso, te pedimos que contestes con honestidad a las siguientes preguntas (*Marca lo que corresponda*):

Pregunta		
1. En mi interacción con herramientas de IA Generativa he remitido **datos de carácter sensible** con la debida autorización de los interesados.		
SÍ, he usado estos datos con autorización	NO, he usado estos datos sin autorización	**NO, no he usado datos de carácter sensible**
2. En mi interacción con herramientas de IA Generativa he remitido **materiales protegidos por derechos de autor** con la debida autorización de los interesados.		
SÍ, he usado estos materiales con autorización	NO, he usado estos materiales sin autorización	NO, no he usado materiales protegidos
3. En mi interacción con herramientas de IA Generativa he remitido **datos de carácter personal** con la debida autorización de los interesados.		
SÍ, he usado estos datos con autorización	NO, he usado estos datos sin autorización	**NO, no he usado datos de carácter personal**

4. Mi utilización de la herramienta de IA Generativa ha **respetado sus términos de uso**, así como los principios éticos esenciales, no orientándola de manera maliciosa a obtener un resultado inapropiado para el trabajo presentado, es decir, que produzca una impresión o conocimiento contrario a la realidad de los resultados obtenidos, que suplante mi propio trabajo o que pueda resultar en un perjuicio para las personas.

SI	NO

Si **NO** has contado con la autorización de los interesados en alguna de las preguntas 1, 2 ó 3, explica brevemente el motivo *(por ejemplo, "los materiales estaban protegidos pero permitían su uso para este fin" o "los términos de uso, que se pueden encontrar en esta dirección (…), impiden el uso que he hecho, pero era imprescindible dada la naturaleza del trabajo".*

Parte 2: declaración de uso técnico

Utiliza el siguiente modelo de declaración tantas veces como sea necesario, a fin de reflejar todos los tipos de iteración que has tenido con herramientas de IA Generativa. Incluye un ejemplo por cada tipo de uso realizado donde se indique: *[Añade un ejemplo].*

Documentación y redacción

- *Revisión o reescritura de párrafos redactados previamente*

Declaro haber hecho uso del sistema de IA Generativa **ChatGPT 4o** para solicitar la reducción del número de palabras en párrafos escritos por mí. Por ejemplo, he usado el prompt "can you give suggestions on how to shorten this section? do not implement them directly, just point them out". También he solicitado la traducción de textos redactados en español, así como la revisión de textos o expresiones redactados en inglés. Ejemplos de prompts para este fin son: "does the first part of the first sentence in this paragraph make sense in english?", o "can you think of synonyms for imbued?".

- *Búsqueda de información o respuesta a preguntas concretas*

Declaro haber hecho uso del sistema de IA Generativa **ChatGPT 4o** para responder a dudas concretas de manera más ágil o estructurada que mediante búsquedas en Google. Por ejemplo, he usado el prompt "from how many authors do I start referencing with et al.".

Desarrollar contenido específico

Se ha hecho uso de IA Generativa como herramienta de soporte para el desarrollo del contenido específico del TFM, incluyendo:

- *Asistencia en el desarrollo de líneas de código (programación)*

Declaro haber hecho uso del sistema de IA Generativa **ChatGPT 4o** como apoyo para la escritura de código en R. Por ejemplo, utilizando el prompt “I'm interested in computing krippendorff's alpha for the models and human labels (the label of models is the column "agreement_type", and for humans it's "label"). How should I prepare the data to do that?”.

- *Procesos de optimización*

Declaro haber hecho uso del sistema de IA Generativa **ChatGPT 4o** de manera frecuente para resolver errores de código y otras dudas relacionadas.

- *Tratamiento de datos: recogida, análisis, cruce de datos…*

Declaro haber hecho uso de modelos abiertos de generación de texto (en concreto, **mistralai/Mistral-7B-Instruct-v0.3**, **deepseek-ai/deepseek-llm-7b-chat**, **Qwen/Qwen3-8B**, **01-ai/Yi-1.5-9B-Chat**, **microsoft/Phi-3-mini-4k-instruct**) para etiquetar textos, analizando posteriormente estas etiquetas. El prompt utilizado en todos los casos fue “Classify the following text as EITHER hate speech OR not hate speech. Complete with ONLY a label: 'hate speech' OR 'not hate speech'. DO NOT provide any explanation”.

- *Inspiración de ideas en el proceso creativo*

Declaro haber hecho uso del sistema de IA Generativa **ChatGPT 4o** para orientarme sobre algunas decisiones que debía tomar, así como los siguientes pasos a seguir, en distintas etapas del proceso de elaboración del TFM. Por ejemplo, he utilizado el prompt “Do you think these three models are all worth keeping?” para recibir recomendaciones sobre las distintas alternativas que me planteaba esta decisión.

Parte 3: reflexión sobre utilidad

Por favor, aporta una valoración personal (formato libre) sobre las fortalezas y debilidades que has identificado en el uso de herramientas de IA Generativa en el desarrollo de tu trabajo. Menciona si te ha servido en el proceso de aprendizaje, o en el desarrollo o en la extracción de conclusiones de tu trabajo.

El uso de la IA generativa ha permitido la realización de este trabajo, en la medida en que varios modelos de generación de texto fueron empleados para etiquetar datos, y sus resultados fueron utilizados para explorar relaciones y extraer conclusiones sobre el propio proceso. Este enfoque me ha animado a informarme y reflexionar sobre los beneficios y las limitaciones de los modelos open-source frente a los cerrados.

Además de esto, la IA generativa me ha servido como herramienta para optimizar el desarrollo de líneas de código, un ámbito en el que apenas tengo un año de experiencia, así como para resolver todas mis dudas sobre éste y otros aspectos de mi TFM, de una manera dinámica y personalizada. También me ha asistido en la redacción del trabajo en un idioma del que no soy hablante nativa. En todo momento he evitado compartir datos de carácter sensible o personal en estas interacciones, y he tratado de mantener presente que la IA puede cometer errores y alucinaciones.

En general, encuentro que la IA generativa me ha resultado muy útil como asistente a lo largo de este proceso, si bien es cierto que en ocasiones he buscado apoyarme demasiado en estas herramientas, sobre todo en la toma de decisiones, lo cual ha acabado resultando ineficiente.