From Label Agreement to Downstream Performance: Evaluating LLM-Generated Training Data

Dario Mammana
dario.mammana.4448@student.uu.se
Uppsala University
Uppsala, Sweden

Petra Kralj Novak*
novakpe@ceu.edu
Central European University
Vienna, Austria

Abstract

Large language models offer a scalable and inexpensive alternative to human annotation, but agreement with a small human-labeled prompt-development sample may not adequately reflect the quality of LLM-generated labels as training data. We investigate this problem in multilingual hate-speech classification using an established diamond-standard dataset of English, Italian, and Slovenian social-media texts with four ordinal classes. We use Claude Haiku 4.5 to generate an alternative set of training annotations, fine-tune equivalent XLM-R classifiers on the human and Claude labels, and evaluate them against held-out human annotations. The classifier trained on Claude labels performs worse across languages and metrics, with the largest differences in less frequent classes, suggesting that systematic annotation differences can propagate into downstream classifier behavior. Our results show that, for the low-cost model examined here, high agreement on a prompt-development sample does not guarantee equivalent downstream classifier performance, highlighting the need to evaluate low-cost LLM annotations beyond prompt-level agreement.

Keywords

LLM annotation, annotation agreement, downstream evaluation, multilingual hate-speech detection

1 Introduction

Constructing labeled training datasets is one of the most costly and time-consuming stages of supervised machine learning. Human annotation is especially demanding for subjective tasks, where multiple judgments may be required for each item and disagreement cannot always be resolved by identifying a single correct label. Large language models (LLMs) offer a potentially inexpensive and scalable alternative to employing human annotators. Recent studies have shown that LLMs can perform some text-annotation tasks with accuracy comparable to that of crowd workers and at a substantially lower cost [5, 12]. This raises a practical question: can human annotations be replaced by LLM-generated labels when constructing training data for downstream classifiers?

Existing evaluations frequently address this question by measuring agreement between LLM-generated and human-provided labels. This agreement is informative, but it does not directly measure the usefulness of the resulting training data. Agreement describes the similarity between annotation sources, but it does not show whether the observed differences materially affect a classifier trained on those annotations. High overall agreement may also obscure systematic differences in less frequent classes. The quality of LLM-generated training data should therefore also be evaluated through its effect on the downstream model trained from those labels.

Hate-speech detection provides a particularly demanding setting for this evaluation. Definitions of hate speech and related forms of inappropriate content vary across annotation schemes, platforms, and social contexts [1, 4, 9]. Individual judgments may depend on context and annotator-specific severity thresholds, making disagreement an inherent property of the data rather than merely annotation error [2, 7]. Agreement tends to characterize clearer cases, while different labels for the same item can indicate ambiguity or proximity to category boundaries. Retaining both judgments preserves this disagreement in the training data rather than resolving it to a single label. Multilingual settings introduce an additional dimension, as annotation patterns and LLM behavior may vary across languages.

In this paper, we evaluate the downstream performance trade-offs of replacing human annotations with a low-cost, zero-shot LLM annotation workflow for multilingual hate-speech classification. We use a previously established diamond-standard dataset and experimental setup [6], covering English, Italian, and Slovenian social-media texts labeled on an ordinal scale distinguishing appropriate, inappropriate, offensive, and violent content. We generate an alternative set of training annotations using Claude Haiku 4.5, fine-tune equivalent XLM-R classifiers on the human- and LLM-annotated data, and evaluate both against held-out human annotations. We approximate a realistic annotation workflow in which a small human-labeled sample is available for prompt development, but obtaining human labels for the complete training set is impractical. We evaluate the resulting LLM annotations at two levels. First, we examine agreement and class-level differences between human and LLM annotations of the training data. Second, we measure how the annotation source affects downstream classification performance. This distinction is central to our study: similarity to human annotations and usefulness as training supervision are related, but they are not equivalent criteria.

Our results provide preliminary evidence that the classifier trained on human annotations performs better than the classifier trained on LLM-generated labels in the examined setting. The performance difference varies across languages and classes and is most pronounced for the less frequent and more severe hate-speech categories. Rather than drawing general conclusions across all LLM families, our findings provide a benchmark for a low-cost model and demonstrate why prompt-level agreement alone is insufficient to guarantee downstream utility.

*Also with Jožef Stefan Institute, Ljubljana, Slovenia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IS 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
https://doi.org/10.70314/is.2026.sikdd.43

2 Related Work

Recent studies have investigated LLMs as inexpensive and scalable text annotators. Gilardi et al. [5] report that ChatGPT outperformed crowd workers on several annotation tasks, while Ollion et al. [12] find that its suitability depends on the task and intended use. Results are less encouraging for subjective and sensitive classification problems, including hate speech [11]. Human–LLM collaborative annotation has therefore also been explored as an alternative to complete replacement [14].

A separate line of research treats annotator disagreement as meaningful information rather than annotation error. This perspective is particularly relevant to hate-speech classification, where judgments depend on context and individual severity thresholds. Kralj Novak et al. [7] retained two annotations per item when training monolingual classifiers. Grigor et al. [6] subsequently pooled the English, Italian, and Slovenian diamond-standard data to train and evaluate multilingual classifiers. The present study uses this disagreement-aware dataset and modeling setup to compare human and LLM annotation as alternative sources of training labels.

3 Methods

We simulate a realistic annotation workflow in which a small human-labeled sample is used for prompt development, while the remaining training data are annotated by the LLM. Human and Claude annotations are then used to fine-tune otherwise identical XLM-R classifiers, which are evaluated against the same held-out human-annotated test set. Krippendorff's ordinal alpha is used throughout to compare agreement between human and Claude annotations, between model predictions, and between predictions and human test annotations (Figure 1).

3.1 Datasets with Human Annotations

The dataset, introduced by Kralj Novak et al. [7], comprises three language-specific subsets of roughly 53,000 unique items each: YouTube comments in English and Italian and Twitter posts in Slovenian. Each item was labeled independently by two human annotators following the IMSyPP annotation guidelines[1] using four ordinal categories: 0–*appropriate*, 1–*inappropriate*, 2–*offensive*, and 3–*violent*. Inter-annotator agreement, measured using Krippendorff's ordinal alpha, is reported in Figure 1.

By design, the class distributions differ substantially between the training and test data. The training data were selected using an optimized sampling-before-labeling strategy [13] designed to increase the representation of minority hate-speech classes. The test data, by contrast, constitute a temporal holdout collected independently during a later period and closely reflect the natural class distribution [7].

3.2 Prompt Engineering and LLM Annotations

We use Anthropic's `claude-haiku-4-5-20251001` to generate an LLM-annotated counterpart to the complete human-annotated training set. Both training sets contain the same multilingual texts and differ only in their annotation source.

The prompt-development sample contains 150 human-annotated items per language, selected from the training set, and includes items on which the annotators disagree. Each ordinal category occurs at least 15 times across the two human annotations. Inter-annotator agreement in the sample is approximately the same as in the complete training set.

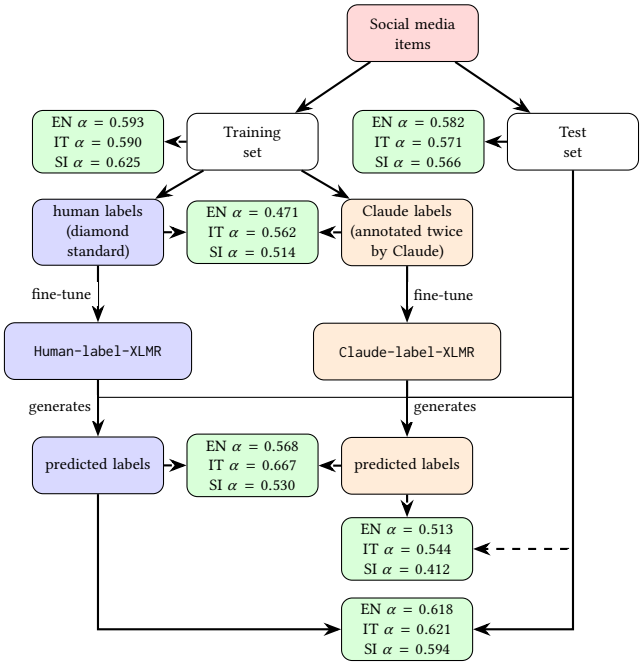

Figure 1: Pipeline overview. Green cells report Krippendorff's ordinal α between connected annotation sources or predictions. They summarize agreement across the experiment: between human annotators, human and Claude annotations, the two classifiers, and each classifier and human test annotations.

The LLM annotation procedure mirrors the diamond-standard structure of the human data by producing two labels for every item. This design is motivated by evidence that annotator disagreement provides useful supervision: agreement identifies relatively clear cases, while different labels indicate uncertainty near class boundaries [6]. Using two Claude labels therefore allows us to examine whether LLM annotations can provide both human-like judgments and comparable diversity in the training signal.

We iteratively refined the prompts on the English prompt-development sample, aiming to balance agreement with human judgments, measured by macro-F1, and diversity between the two Claude annotations, measured by Krippendorff's ordinal alpha (Table 1). Starting from the human annotation instructions, we first varied the temperature, but even the maximum temperature produced near-perfect within-Claude agreement ($\alpha = 0.99$). We therefore introduced strict and lenient annotator personas to induce variation in annotation thresholds while keeping the guidelines fixed. In total, we evaluated five configurations varying temperature, persona definitions, and task instructions. Across all configurations, the most prominent misclassification pattern involved the two most frequent classes, *appropriate* and *offensive*. The prompts and confusion matrices for all configurations, together with the code and experimental setup, are available in the project's GitHub repository.[2]

[1] Annotation guidelines: http://imsypp.ijs.si/wp-content/uploads/2021/12/IMSyPP_D2.2_multilingual-dataset.pdf

[2] Project GitHub repository: https://github.com/dariomammana/llm_labels_for_hate_speech

Table 1: Prompt development results across languages. α denotes Krippendorff's ordinal alpha; macro-F1 is averaged across all four classes.

Language	LLM α	Human α	Human-LLM α	Macro-F1
EN	0.878	0.593	0.615	0.595
IT	0.887	0.590	0.693	0.691
SI	0.912	0.625	0.618	0.533

Table 2: Label distribution of the two annotation regimes.

Source	Appropriate	Inappropriate	Offensive	Violent
Human	189,674 (59.27%)	10,960 (3.42%)	112,289 (35.09%)	7,111 (2.22%)
Claude	193,765 (60.56%)	28,138 (8.80%)	92,011 (28.76%)	6,017 (1.88%)

The prompts for the two strongest persona-based configurations were then translated into Italian and Slovenian and evaluated on the corresponding development sets. Their consistent behavior across languages supported the selection of a single configuration for all three training sets. Table 1 reports the results for the selected configuration.

Applying the selected configuration produced the class distributions shown in Table 2. Compared to the human annotations, Claude assigned more items to the *inappropriate* class and fewer to the *offensive* and *violent* classes. Thus, similar overall proportions of *appropriate* content conceal systematic differences among the less frequent categories.

3.3 Model Fine-tuning and Evaluation

We fine-tune two classifiers from the same pretrained XLM-R model [3]: `Human-label-XLMR` on the human annotations and `Claude-label-XLMR` on the Claude annotations. Rather than evaluating isolated models, we compare two realistic training pipelines. The architecture, data partitions, and base hyperparameters are held constant, while dataset-dependent elements such as class loss weights and validation checkpoint selection are computed separately within each pipeline.

We partition the training set into training (80%) and validation (20%) sets, stratified by language and label. Both labels for an item always occur in the same partition, thereby preventing data leakage between the training and validation sets.

Both classifiers are trained for three epochs using AdamW, a learning rate of 6×10^{-6}, and a batch size of 8. To account for class imbalance, we use weighted cross-entropy with weights inversely proportional to class frequencies in the respective training set. For each classifier, we select the checkpoint with the highest validation macro-F1. Fine-tuning was conducted on an NVIDIA A100 GPU.

Each test item has two human annotations, which are treated as two evaluation instances. The model therefore produces one prediction per unique item, and this prediction is compared separately with each of the two human judgments.

We report three complementary metrics. Accuracy is the proportion of model–annotation pairs with identical labels. Macro-F1 is the unweighted mean of the four class-specific F1 scores and therefore gives equal importance to frequent and infrequent classes. Krippendorff's ordinal alpha [8] measures agreement beyond chance while accounting for the ordered severity scale: disagreement between *appropriate* and *violent*, for example, is penalized more strongly than disagreement between adjacent categories. Following Kralj Novak et al. [7] and Grigor et al. [6], ordinal alpha is the primary metric because it accounts for chance agreement and permits comparison between model–annotator and inter-annotator agreement.

4 Results

`Human-label-XLMR` outperforms `Claude-label-XLMR` on all metrics and in all three languages (Table 3), with the smallest difference for Italian and the largest for Slovenian.

Table 3: Evaluation results for both models across all three languages. Alpha: Krippendorff's ordinal alpha. F1: macro-averaged F1 score.

Language	Ann. method	Alpha	Macro F1	Accuracy
EN	Human	**0.618**	**0.563**	**0.831**
	Claude	0.513	0.511	0.813
IT	Human	**0.621**	**0.663**	**0.842**
	Claude	0.544	0.580	0.785
SI	Human	**0.594**	**0.602**	**0.809**
	Claude	0.412	0.445	0.741

Figure 2 presents paired bootstrap estimates of the difference between the human-trained and Claude-trained models, with positive values favoring the former. All 95% confidence intervals exclude zero, and all comparisons remain statistically significant after Holm correction for multiple comparisons.

Table 4 reports the class-specific F1 scores for both models. The performance gap is negligible for the *appropriate* class but widens considerably for the hate-speech classes. It is largest for *violent* in Slovenian, where `Claude-label-XLMR` scores only 0.176, compared with 0.364 for `Human-label-XLMR`.

Table 4: Per-class F1 scores on the human-labeled test set. The best score for each language and class is shown in bold.

Lang.	Ann. method	Approp.	Inapprop.	Offensive	Violent
EN	Human	**0.888**	**0.325**	**0.713**	**0.327**
	Claude	**0.888**	0.280	0.584	0.292
IT	Human	**0.912**	**0.601**	**0.629**	**0.512**
	Claude	0.878	0.476	0.558	0.409
SI	Human	**0.866**	**0.469**	**0.710**	**0.364**
	Claude	0.843	0.264	0.498	0.176

5 Discussion

The results illustrate that prompt-level agreement is insufficient for evaluating training labels produced by a low-cost LLM. Although the selected prompt achieved reasonable agreement with human labels, the classifier trained on Claude labels performed consistently worse, with all comparisons remaining significant after Holm correction. Annotation agreement and downstream utility are therefore related but distinct evaluation criteria.

The class-level results suggest how annotation differences can propagate into downstream classifier behavior. Claude assigned a similar proportion of *appropriate* labels but more *inappropriate* and fewer *offensive* and *violent* labels than the human annotators (Table 2). Correspondingly, the difference between the classifiers

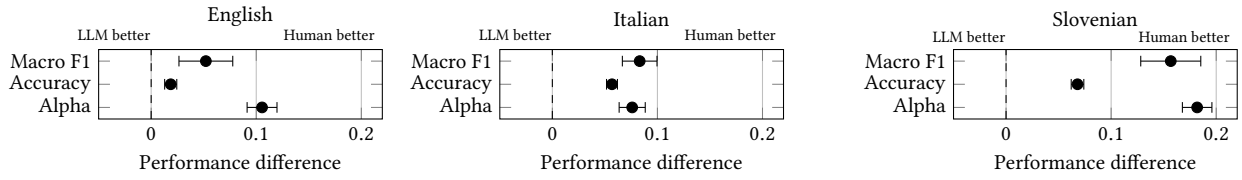

Figure 2: Paired bootstrap comparison between the human-trained and LLM-trained models. Dots denote the observed paired difference (human − LLM), and horizontal bars indicate the 95% bootstrap confidence intervals obtained from 5,000 paired resamples of test items with replacement. The dashed vertical line marks zero. All comparisons remain statistically significant after Holm correction for multiple comparisons.

was small for the *appropriate* class and larger for the less frequent classes, particularly *offensive* and *violent*. This pattern does not establish a causal relationship, but it indicates that relatively small systematic differences in training labels can produce substantial differences in class-specific performance.

Annotator diversity may also contribute to the observed performance gap. Despite persona prompting, agreement between the two Claude runs remained substantially higher than between human annotators. The Claude-annotated training data may therefore represent ambiguous cases less effectively than the human diamond standard. Future work should examine whether greater diversity can be obtained by combining annotations from different LLMs rather than eliciting two perspectives from the same model. Comparing single-LLM, multi-LLM, and human annotation regimes would help separate the effects of correspondence with human judgments from the value of disagreement itself. An earlier English-only experiment using a single LLM annotation per item similarly found that a classifier trained on LLM labels performed worse than one trained on human labels [10]. However, whether two LLM annotations per item provide better training supervision than one remains untested.

The economic advantage of LLM annotation remains substantial. Generating all 319,931 Claude labels, including the prompt-development experiments, cost less than USD 65 through the Anthropic API (June 2026). The present results therefore reveal a concrete trade-off: LLM annotation offers considerable savings and scalability, but these benefits were accompanied by lower downstream performance, especially for less frequent classes.

This study is limited to one small LLM (Claude Haiku 4.5), a single training run per model, and three languages. The larger performance gap observed for Slovenian may arise from the available language resources, the source platform of the data, or other annotation properties. Given the substantial cost advantage of this setup, future work should evaluate whether more capable models and practices such as chain-of-thought reasoning and systematic persona ablations can make agreement a more reliable proxy for downstream utility.

6 Conclusion

We evaluated LLM-generated training labels through both their agreement with human annotations and their effect on a downstream classifier. In this multilingual hate-speech experiment, XLM-R trained on Claude labels performed worse than XLM-R trained on human diamond-standard annotations, especially for less frequent classes. At the same time, the low annotation cost demonstrates the practical appeal of LLM-generated labels. These findings show that agreement on a small prompt-development sample may not adequately represent the downstream loss in quality. Robust protocols are therefore needed to estimate the utility of LLM-generated training labels from limited human-annotated data.

Acknowledgements

The authors acknowledge partial financial support from the Slovenian Research Agency (research core funding no. P2-103).

References

[1] Alexander Brown. 2017. What is Hate Speech? Part 2: Family Resemblances. *Law and Philosophy*, 36, 5, (Oct. 2017), 561–613. doi:10.1007/s10982-017-9300-x.

[2] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 6, (June 26, 2023), 6860–6868. doi:10.1609/aaai.v37i6.25840.

[3] Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, (July 2020), 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[4] Paula Fortuna, Juan Soler, and Leo Wanner. [n. d.] Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6786–6794. Retrieved Apr. 13, 2026 from.

[5] 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, (July 25, 2023), e2305016120. doi:10.1073/pnas.2305016120.

[6] Patricia-Carla Grigor, Bojan Evkoski, and Petra Kralj Novak. 2024. Multilingual Hate Speech Modeling by Leveraging Inter-Annotator Disagreement. In *Proceedings of Data Mining and Data Warehouses – SiKDD 2024*. Jožef Stefan Institute. doi:10.70314/is.2024.sikdd.7.

[7] 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*. Springer International Publishing, Cham, 681–695. doi:10.1007/978-3-031-08974-9_54.

[8] Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc. doi:10.4135/9781071878781.

[9] Seul Lee and Anne Gilliland. 2024. Evolving Definitions of Hate Speech: The Impact of a Lack of Standardized Definitions. In *Wisdom, Well-Being, Win-Win*. Springer Nature Switzerland, Cham, 141–156. doi:10.1007/978-3-031-57860-1_11.

[10] Dario Mammana. 2026. *Human vs LLM labels for fine tuning hate speech classification: a cost-effective comparative analysis*. Bachelor's Thesis. Central European University, Vienna, Austria.

[11] Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, St. Julian's, Malta, (Mar. 2024), 179–192. doi:10.18653/v1/2024.eacl-short.17.

[12] Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. ChatGPT for Text Annotation? Mind the Hype! (Oct. 4, 2023). Pre-published.

[13] Andraž Pelicon, Syrielle Montariol, and Petra Kralj Novak. 2023. Don't start your data labeling from scratch: OpSaLa-optimized data sampling before labeling. In *International Symposium on Intelligent Data Analysis*. Springer, 353–365.

[14] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 11, 2024), 1–21. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3641960.