ARTICLE: Annotator Reliability Through In-Context Learning

Sujan Dutta[1], **Deepak Pandita**[1], **Tharindu Cyril Weerasooriya**[1], **Marcos Zampieri**[2], **Christopher M. Homan**[1], **Ashiqur R. KhudaBukhsh**[1*]

[1]Rochester Institute of Technology
[2]George Mason University
{sd2516, cmhvcs}@rit.edu, {deepak, cyril, khudabukhsh}@mail.rit.edu, mzampier@gmu.edu

Abstract

Ensuring annotator quality in training and evaluation data is a key piece of machine learning in NLP. Tasks such as sentiment analysis and offensive speech detection are intrinsically subjective, creating a challenging scenario for traditional quality assessment approaches because it is hard to distinguish disagreement due to poor work from that due to differences of opinions between sincere annotators. With the goal of increasing diverse perspectives in annotation while ensuring consistency, we propose ARTICLE, an in-context learning (ICL) framework to estimate annotation quality through self-consistency. We evaluate this framework on two offensive speech datasets using multiple LLMs and compare its performance with traditional methods. Our findings indicate that ARTICLE can be used as a robust method for identifying reliable annotators, hence improving data quality.

Code — https://github.com/Suji04/ARTICLE

Introduction

From classical supervised systems (Carbonell, Michalski, and Mitchell 1983) to the RLHF framework (Christiano et al. 2017), human input plays a central role in human value-aligned AI and NLP systems. Crowdsourcing is a well-studied, affordable, and distributed framework that allows data collection from broad and diverse annotator pools within a short period (Gray and Suri 2019; Kahneman, Sibony, and Sunstein 2021; Wang, Hoang, and Kan 2013). The benefits of crowdsourcing notwithstanding, enforcing quality control and estimating annotation quality remain a long-standing challenge (Lease 2011; Huang, Fleisig, and Klein 2023).

Conventional approaches to distinguish *high* from *poor* quality annotators are typically based on outlier detection, where the divergence from aggregate opinions is considered a signal of poor quality annotation (Dumitrache et al. 2018a; Leonardelli et al. 2021; Davani, Díaz, and Prabhakaran 2022; Ustalov, Pavlichenko, and Tseitlin 2024). However, for subjective tasks (Passonneau et al. 2012; Pavlick and Kwiatkowski 2019; Uma et al. 2021; Nie, Zhou, and Bansal 2020; Jiang and Marneffe 2022; Deng et al. 2023), such outlier-based approaches can potentially muffle minority or unique perspectives, leading to annotation echo chambers. Consider a war corpus where annotators hail from countries $\mathcal{A}$ and $\mathcal{B}$. Even simple questions like *who is winning the war* could have drastically different responses depending on which country the annotator belongs to. If a pool has an overwhelming presence of $\mathcal{A}$, any perspective that annotators from $\mathcal{B}$ could contribute to will be eliminated since their responses will be visibly different from the majority view.

This paper introduces an alternative path to estimate annotator quality through the lens of self-consistency. Prior work in this domain explored to address it through the lens of annotation patterns of individual annotators (Dawid and Skene 1979; Hovy et al. 2013; Ustalov, Pavlichenko, and Tseitlin 2024), without taking into account what is being annotated (context) and information of the annotator. Suppose we are interested in collecting a dataset of offensive speech. If we observe that a given annotator has marked one instance that attacks an ethnic group as highly offensive while marking another instance with an even sharper attack on the same group as not offensive, the annotator's responses could be self-consistent. Incorporating self-consistency into the annotation quality estimation process has the following benefits. First, it bypasses the requirement of having annotations from multiple other annotators to compute divergence from aggregate opinion, thus promising to be more resource-efficient. Second, this approach preserves unique but self-consistent perspectives, which outlier-based methods might eliminate.

While the notion of self-consistency has been applied to diverse settings (see, e.g., Wang et al. (2023); Cooper et al. (2024a)), to our knowledge, this paper first applies self-consistency for rater quality estimation on subjective annotation tasks. The introduction of large language models (LLM) with larger context lengths for language understanding has also led to research on utilizing LLMs (Gilardi, Alizadeh, and Kubli 2023; He et al. 2024) as human annotators. However, prior research has focused on using the LLM (He et al. 2024) to replace the majority opinion of data annotation but not the intricate annotator-level labels.

Contributions. Our contributions are the following.

1. We introduce ARTICLE, a novel framework to estimate annotator quality through self-consistency;

*Ashiqur R. KhudaBukhsh is the corresponding author.

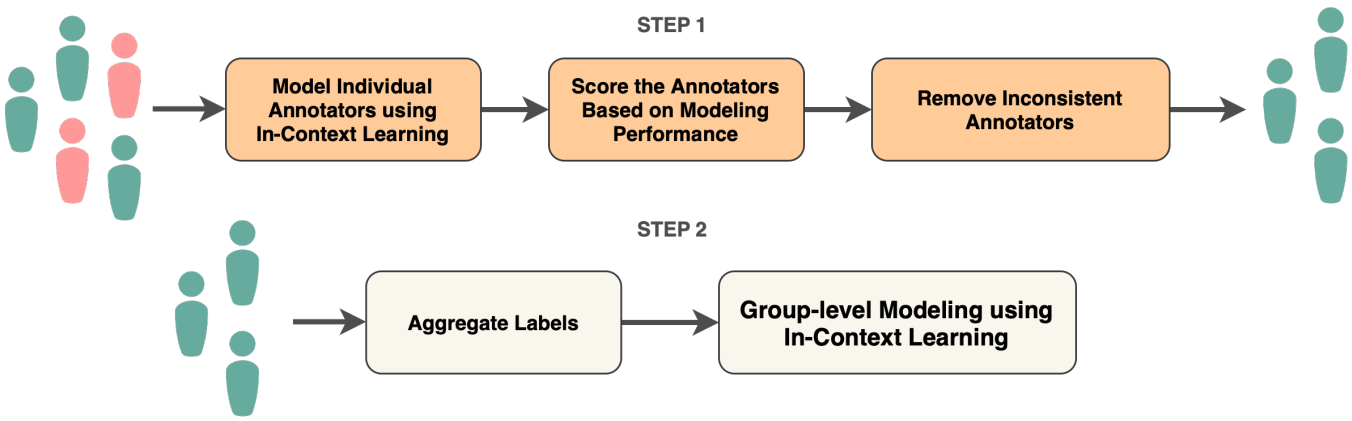


Figure 1: Schematic Diagram of ARTICLE.

2. We evaluate this framework on two well-known English offensive speech datasets: (1) Toxicity Ratings (Kumar et al. 2021) henceforth $\mathcal{D}_{\text{TR}}$; and (2) VOICED (Weerasooriya et al. 2023a) henceforth $\mathcal{D}_{\text{VOICED}}$ and we contrast our approach with CrowdTruth (CT) (Dumitrache et al. 2018a).

Related Work

Crowdsourcing platforms such as Amazon Mechanical Turk, Toloka, and Prolific have played a critical role over the years for collecting annotations for training models (Kahneman, Sibony, and Sunstein 2021). However, just as with any task with human annotators in the loop, prior research has identified instances when annotators have been inconsistent with providing information (Huang, Fleisig, and Klein 2023; Abercrombie, Hovy, and Prabhakaran 2023). Röttger *et al.* (2021) demonstrated the impact of the paradigm (subjective or prescriptive) used during the survey on the (dis)agreement level of the annotations. These characteristics have led to research for modeling annotators and rating them for reliability.

Dawid and Skene (1979) presented the initial two-stage generative model for inferring ground truth from unreliable annotators. The model assumes each annotator has a concealed error rate and utilizes expectation maximization to iteratively estimate these error rates along with the most probable ground truth labels based on the current error rate estimates. Hovy et al. (2013) extended this model with a bipartite annotator model that distinguishes between spammers and non-spammers. CrowdTruth (Dumitrache et al. 2018b) is another method for measuring the reliability of the annotators and the entire dataset as a whole based on their overall agree-ability with other annotators.

However, a limitation of prior work is not taking into consideration the content of the data item that is being annotated for scoring the performance of the annotators and how consistent the annotator is in terms of annotating (Cooper et al. 2024b). In our research, we explore how to utilize the capabilities of the LLM to understand and identify inconsistencies of the annotators utilizing the context of the annotation task. The use of LLMs as judges in various tasks (Zheng et al. 2023; Tan et al. 2024; Huang et al. 2024) has recently garnered much attention. While most of the current work focuses on improving the abilities of smaller language models, we see a potential for utilizing these LLM judges as an evaluator of annotation consistency. At the same time, we remain cognizant of inherent limitations, such as model bias, and address these issues in our Limitations section. This paper studies consistency in human annotations; however, the proposed method can also be used for LLM-generated annotations.

Methodology

We propose **ARTICLE** (**A**nnotator **R**eliability **T**hrough **I**n-**C**ontext **Le**arning) – a two-step framework (Figure 1) to identify reliable annotators and model the perception of offense for different political groups. In the fist step, we identify the annotators who exhibit inconsistency in labeling and remove them from the dataset. In the second step, based on the aggregated responses of the consistent annotators, we model the group-level perception of offense.

Step 1: Identifying Inconsistent Annotators

We hypothesize that annotators who show inconsistent annotation patterns are difficult to model. We individually model each annotator using a well-known and high-performance LLM, Mistral-7B-instruct (Jiang et al. 2023), and utilize the model's performance (ease of modeling) as a proxy for the annotator's consistency. For each annotator, we randomly split their annotations into two sets – the first set (training set) contains 10 data points, and the second (test set) contains the rest. Using the training set as in-context learning (ICL) (Dong et al. 2022; Min et al. 2022) examples, we prompt Mistral-7B-instruct to predict the labels for the test set. The detailed prompt can be found in Figure 4. Then, we compute the F1-score to evaluate the model's performance. A high F1 score indicates the annotator is easy to model and, hence, consistent, and a low score indicates the opposite. We define a hyperparameter (k) that acts as a threshold. If, for a given annotator, the F1-score is less than k, we mark them as inconsistent and remove them from the

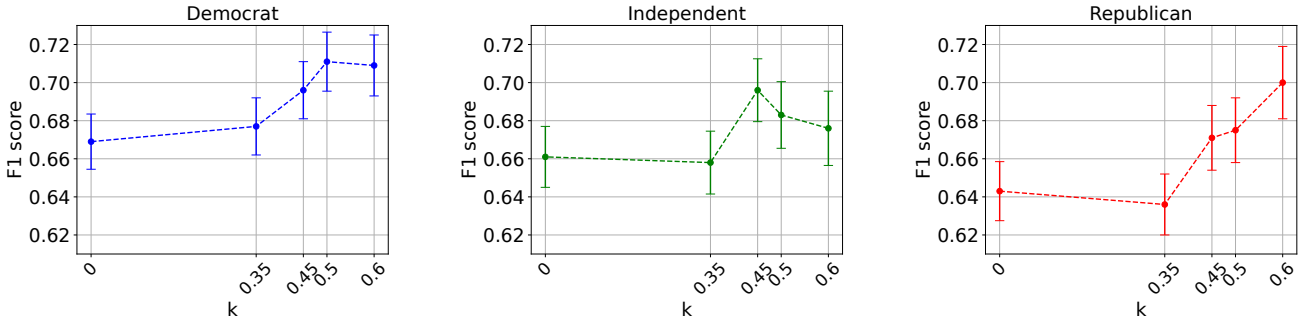


Figure 2: Group-level model performance at different k values in $\mathcal{D}_{\texttt{TR}}$. The error bars indicate 95% confidence interval.

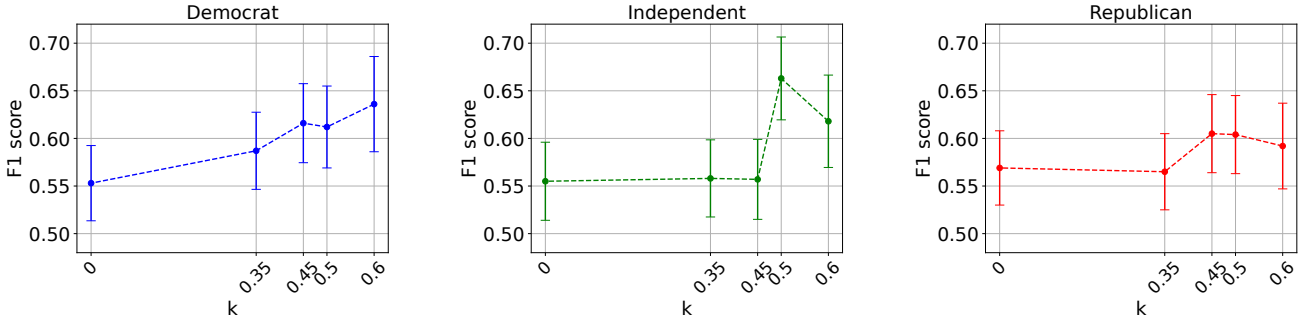


Figure 3: Group-level model performance at different k values in $\mathcal{D}_{\texttt{VOICED}}$. The error bars indicate 95% confidence interval.

```
You are an expert in guessing my response against
a social media comment. Your task is to analyze and
predict my response against the comment after <<<>>>
into one of the following pre-defined categories:

offensive
non-offensive

###
Here are some examples:

Comment: {comment text}
Response: {response} (offensive/non-offensive)
       ...
(10 few-shot examples)
       ...
<<<
Comment: {test comment text}
>>>
```

Figure 4: Prompt designed for ARTICLE.

dataset.

Step 2: Modeling Group-level Perception

After removing the inconsistent annotators from all political groups, we recompute the aggregate labels for each group. We again use ICL to model the group-level perception of offense. For each group, we construct a training set using 70% of the data. The rest is used for testing. For each test instance, we randomly sample 15 examples from the training set and use them as in-context examples. The same `Mistral-7B-instruct` model is used in this step.

Experimental Setup

Datasets

Political Leaning	$\mathcal{D}_{\texttt{TR}}$	$\mathcal{D}_{\texttt{VOICED}}$
Democrat	43%	34%
Republican	28%	36%
Independent	29%	30%

Table 1: Distribution of political leanings of the annotators in $\mathcal{D}_{\texttt{TR}}$ and $\mathcal{D}_{\texttt{VOICED}}$.

We consider two datasets on web toxicity: $\mathcal{D}_{\texttt{TR}}$ and $\mathcal{D}_{\texttt{VOICED}}$. $\mathcal{D}_{\texttt{TR}}$ contains 107,620 comments from multiple social web platforms (Twitter, Reddit, and 4chan) collectively annotated by 17,280 annotators. We sample 20,000 comments from $\mathcal{D}_{\texttt{TR}}$ for our experiments ensuring that each set of 20 comments is annotated by the same five annotators, thereby retaining the structure of the original dataset. $\mathcal{D}_{\texttt{VOICED}}$ (Weerasooriya et al. 2023b) includes 2,338 YouTube comments on three major US cable news networks (KhudaBukhsh et al. 2021) annotated by 726 annotators. Both datasets include annotators from diverse political backgrounds with at least 28% (Table 1) representation from each major political affiliation – Democrats, Republicans, and Independents. In both datasets, comments are rated

on a five-point scale of toxicity. To avoid rare classes, we convert these categories into binary labels. The lowest two toxic categories are mapped to *non-offensive* class, and the rest are mapped to *offensive* class.

Models

We primarily use `Mistral-7B-instruct` for the proposed framework; however, we have conducted experiments with `Llama3-8B-instruct` (Touvron et al. 2023) and `GPT-3.5-turbo` (OpenAI 2022) to ensure the stability of the method.

Computing Environment

We run all our experiments in a Google Colab (pro+) environment with a single A100 GPU (40 GB) and 52 GB RAM.

Inconsistent Annotation Examples

Before we discuss our results and findings, we make a short detour to highlight why consistency matters through a few illustrative examples. We present annotations from a few inconsistent annotators as identified by `ARTICLE`. For instance, in the first example, both the comments express that the government should not control abortion; however, the annotator labeled them differently. In the second example, the annotator labeled a seemingly harmless comment offensive; however, assigned a non-offensive label to a comment containing a racial slur.

EXAMPLE ANNOTATOR #1

Comment: `Abortion should be between the woman and God. I do agree the government should not control it. I'm not saying I agree with Abortion what so ever but the rioting and stuff is just ridiculous.`
Label: *offensive*

Comment: `The 'state' has no business deciding what women can or cannot do to their bodies. The fact that so many in the USA feel that abortion should be banned speaks more to a moronic pandering to the right wing religious fringe crazies who want to impose THEIR beliefs on everyone.`
Label: *non-offensive*

EXAMPLE ANNOTATOR #2

Comment: `Amazing to see some of our ex students still continuing on their passion for the arts`
Label: *offensive*

Comment: `yea lil ass went Mia that boy a hog. n`[redacted]`gg`[redacted]`r finna average 3 assists a game a piece`
Label: *non-offensive*

EXAMPLE ANNOTATOR #3

Comment: `If I Was A Cruddy n`[redacted]`gg`[redacted]`r Like I USED To Be And I Knew A Girl With An OnlyFans Account. I'd Cyber Pimp That Hoe`
Label: *non-offensive*

Comment: `Omfg is it like engagement season or something??? Like everyone is getting engaged right now. Wtf`
Label: *offensive*

EXAMPLE ANNOTATOR #4

Comment: `Oh you wanna be part of my business venture? You can help fill the twinkies with c`[redacted]`m`
Label: *non-offensive*

Comment: `Can't wait to see you guys`
Label: *offensive*

Evaluation

Modeling Performance

We evaluate the proposed framework on $\mathcal{D}_{\texttt{TR}}$ and $\mathcal{D}_{\texttt{VOICED}}$. In each dataset, we model the perception of offense for each political group: Democrat, Republican, and Independent. As mentioned earlier, our framework requires setting a value for the hyperparameter k. To study the impact of k, we run experiments for the following values of $k : \{0, 0.35, 0.45, 0.5, 0.6\}$. The case $k = 0$ serves as the baseline where we do not remove any annotators from the dataset. Figures 2 and 3 illustrate the performance (F1-score on the test set) at various values of k for $\mathcal{D}_{\texttt{TR}}$ and $\mathcal{D}_{\texttt{VOICED}}$, respectively. In general, in both the datasets, across all political groups, we observe an upward trend in the F1-score as the value of k increases with noticeable fluctuations for Independents. In almost all instances, the F1-score achieved with $k = 0.45$ surpassed the baseline performance, suggesting the effectiveness of the proposed method. We also note for most cases with $k > 0.5$, the performance either plateaus or declines slightly. It suggests that while increasing k generally improves model performance up to a point, there may be a threshold beyond which further increase in k does not yield additional benefits and might even be detrimental.

Data Loss

While increasing k improves modeling performance, the annotations lost in this process merit investigation. We first compute the percentage of the annotators remaining at various values of k. From Figures 5a and 5b, we note that $\mathcal{D}_{\texttt{VOICED}}$ undergoes a sharper decline in annotators compared to $\mathcal{D}_{\texttt{TR}}$. However, at $k = 0.45$, we still retain the

majority ($\sim$ 70% in $\mathcal{D}_{\texttt{TR}}$ and $\sim$ 55% in $\mathcal{D}_{\texttt{VOICED}}$) of the annotators in both datasets, with Democrats generally showing the highest retention rates.

Next, we focus on the number of comments remaining as we increase k. We again compute this at group level for $\mathcal{D}_{\texttt{TR}}$ (Figure 5c) and $\mathcal{D}_{\texttt{VOICED}}$ (Figure 5d). ARTICLE at $k = 0.45$, retains more than 80% of the comments in both datasets.

Comparison with CT

Political Leaning	CT (*WQS* ≥ 0.6)	ARTICLE ($k \geq 0.45$)
Democrat	0.669 ± 0.016	0.696 ± 0.015
Republican	0.642 ± 0.018	0.671 ± 0.017
Independent	0.665 ± 0.018	0.696 ± 0.017

Table 2: Group-level modeling performance (F1-score on test set) comparison between ARTICLE and CT in $\mathcal{D}_{\texttt{TR}}$. The results are computed over five runs with different random seeds.

Political Leaning	CT (*WQS* ≥ 0.7)	ARTICLE ($k \geq 0.45$)
Democrat	0.449 ± 0.036	0.616 ± 0.041
Republican	0.435 ± 0.032	0.605 ± 0.042
Independent	0.453 ± 0.036	0.557 ± 0.042

Table 3: Group-level modeling performance (F1-score on test set) comparison between ARTICLE and CT in $\mathcal{D}_{\texttt{VOICED}}$. The results are computed over five runs with different random seeds.

We compare our framework with CT, a well-known method of estimating the quality of annotations (Dumitrache et al. 2018a). CT computes multiple metrics on the annotated dataset, among which *WQS* measures the quality of the annotators. The value of *WQS* ranges between $[0, 1]$. We consider annotators who score more than (or equal to) a specific *WQS* value and model their aggregated annotations following the second step of ARTICLE. Using $\mathcal{D}_{\texttt{TR}}$, we choose $WQS = 0.6$, as in this setting, CT retains a similar percentage ($\sim$ 70%) of annotators to ARTICLE ($k = 0.45$). Table 2 shows that ARTICLE outperforms CT across all groups. The results for $\mathcal{D}_{\texttt{VOICED}}$ are presented in Table 3. Here, too, we notice a significant performance improvement with ARTICLE over CT.

We further investigate the overlap between ARTICLE and CT. Figures 6 and 7 show the venn diagram between the low-quality annotators identified by the two methods in $\mathcal{D}_{\texttt{TR}}$ and $\mathcal{D}_{\texttt{VOICED}}$. We observe that while there is a substantial overlap between the two methods, there are annotators who are flagged as low-quality by one but not by the other. This suggests that these methods measure slightly different aspects of the annotation quality, and future work should explore ways to combine them in a single pipeline.

Stability across LLMs

Beyond `Mistral-7B-instruct`, we study the robustness of the ARTICLE framework across multiple LLMs. We consider two additional models: `Llama3-8B-instruct` (Touvron et al. 2023) (open-sourced) and `GPT-3.5-turbo` (OpenAI 2022) (v. 0125, proprietary). To study the stability of our framework, we look at the overlap between the inconsistent annotators found by different LLMs. More precisely, we compute the Jaccard similarity between the sets of inconsistent annotators identified using a pair of LLMs. To ensure a fair evaluation, for each LLM, we consider the annotators who score less than the median as the inconsistent annotators. Table 4 and 5 present the Jaccard similarities among the LLMs pairs in $\mathcal{D}_{\texttt{TR}}$and $\mathcal{D}_{\texttt{VOICED}}$, respectively. We find a substantial (≥ 0.60) similarity between every pair of LLM in both datasets, suggesting the stability of the framework. We also report the similarity between the annotators found by different LLMs with CT. The similarities between each LLM and CT are much lower (≤ 0.40) than between any two LLMs. This result indicates that CT does not identify many inconsistent annotators as poor-quality annotators. On the other hand, ARTICLE does not remove many of the annotators deemed unreliable by CT.

	Mistral	Llama3	CT
Mistral	-	0.60	0.35
Llama3	0.60	-	0.40
CT	0.35	0.40	-

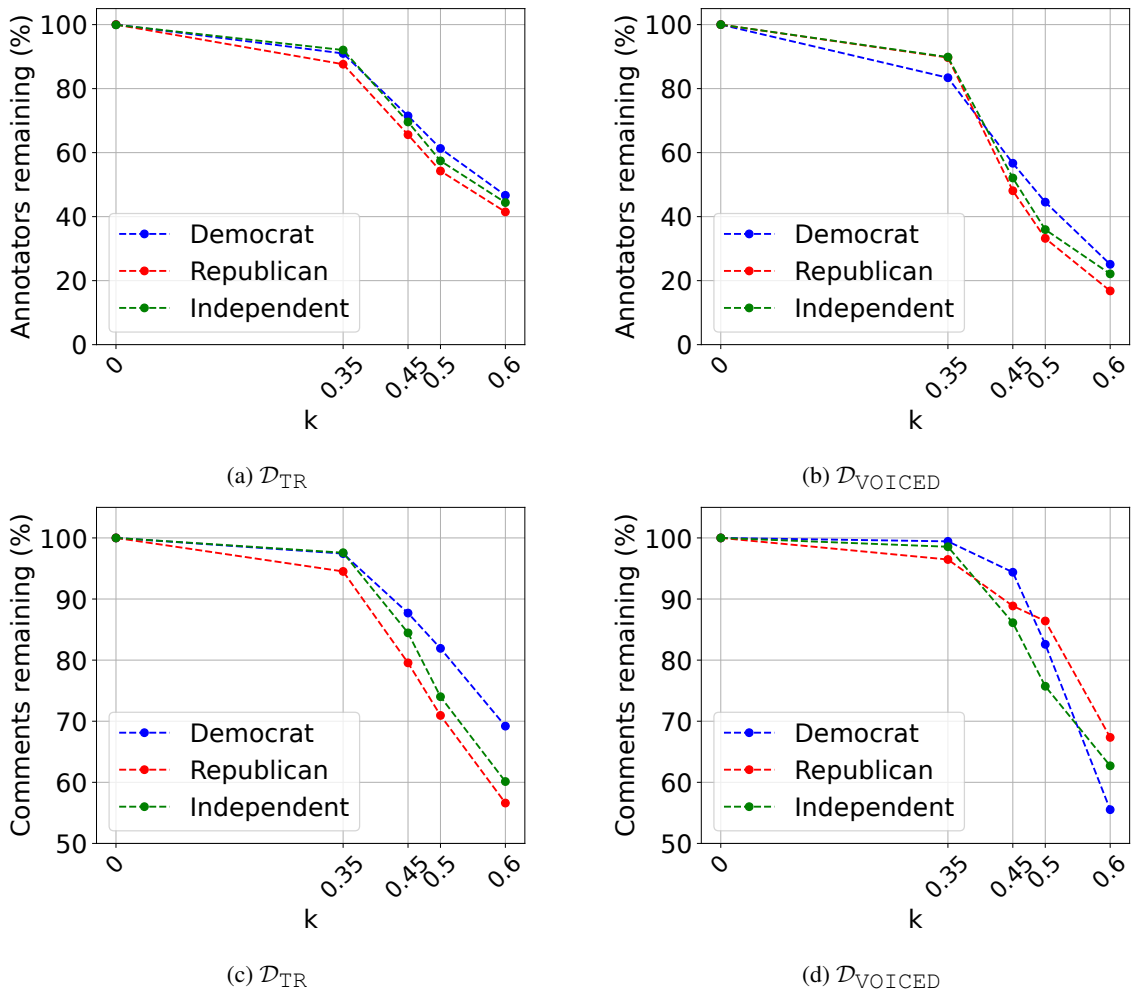
Table 4: Jaccard similarities between inconsistent annotators identified by ARTICLE using different LLMs in $\mathcal{D}_{\texttt{TR}}$. It also includes similarities between each LLM and CT. Due to resource limitations, GPT was not used for this dataset.

	Mistral	Llama3	GPT	CT
Mistral	-	0.68	0.65	0.18
Llama3	0.68	-	0.65	0.16
GPT	0.65	0.65	-	0.20
CT	0.18	0.16	0.20	-

Table 5: Jaccard similarities between inconsistent annotators identified by ARTICLE using different LLMs in $\mathcal{D}_{\texttt{VOICED}}$. It also includes similarities between each LLM and CT.

Conclusion

We introduce ARTICLE, a novel framework for estimating annotator quality through self-consistency. Our approach marks a significant shift from traditional outlier-based methods. Evaluations across two offensive speech datasets demonstrate that ARTICLE effectively identifies reliable annotators while preserving unique, self-consistent viewpoints that might be overlooked. Furthermore, the consistent performance of ARTICLE across multiple language models highlights its robustness. Focusing on self-consistency reduces the dependence on larger annotator pools, potentially lowering costs and increasing the feasibility of deploying quality control mechanisms in annotation tasks. The ongoing development of ARTICLE aims to enhance our understanding and management of the subjective nature of annotation, paving the way for more reliable and inclusive data



(a) $\mathcal{D}_{\texttt{TR}}$ (b) $\mathcal{D}_{\texttt{VOICED}}$

(c) $\mathcal{D}_{\texttt{TR}}$ (d) $\mathcal{D}_{\texttt{VOICED}}$

Figure 5: Percentage of annotators and comments remaining at various value of k in $\mathcal{D}_{\texttt{TR}}$ and $\mathcal{D}_{\texttt{VOICED}}$.

collection methods.

Limitations

While ARTICLE introduces a promising approach to annotator quality assessment, several limitations warrant further investigation:

Model Bias. Reliance on LLMs for evaluating self-consistency could introduce biases inherent to these models (Bommasani et al. 2022; Dutta et al. 2024). These biases may affect the framework's ability to accurately estimate the quality of annotations, especially in contexts involving linguistic or cultural nuances that LLMs might not fully capture.

Handling Justified Disagreement. ARTICLE currently lacks a robust mechanism to distinguish between justified disagreements and genuine inconsistencies in annotations which merits deeper exploration.

Generalizability Across Domains. While tested on datasets involving offensive speech, the generalizability of the framework to other types of annotation tasks, such as medical image annotation or legal document analysis, remains unverified. Different domains may present unique challenges that require adaptations of the framework.

Dependency on Annotation Volume. The effectiveness of ARTICLE is constrained by the volume of data available for each annotator. In scenarios where annotators contribute a low number of annotations, the assessment of self-consistency could be less reliable.

Ethics Statement

ARTICLE's approach to annotation quality assessment through self-consistent intends to help mitigate potential biases towards minor perspectives in NLP systems. In this work, we used two publicly available datasets referenced in the paper. No new data collection has been carried out as part of this work. The datasets used do not reveal any identifiable information about the annotators.

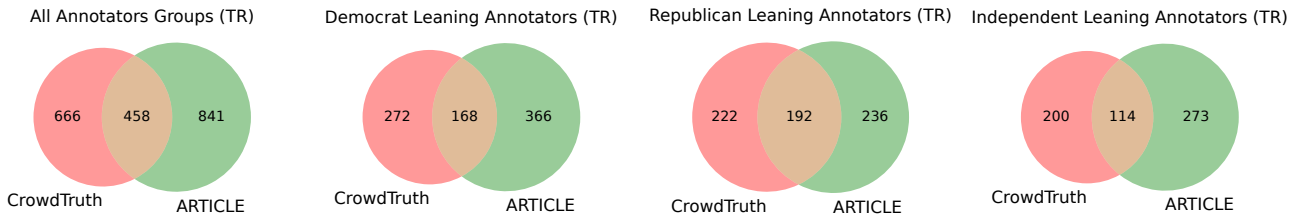


Figure 6: Annotators that are identified as unreliable based on CT and ARTICLE scores for $\mathcal{D}_{\texttt{TR}}$. The last three Figures show the same inconsistent annotators broken down by their political leaning. For $\mathcal{D}_{\texttt{TR}}$, CT ($WQS \geq 0.6$) and ARTICLE ($k \geq 0.45$).

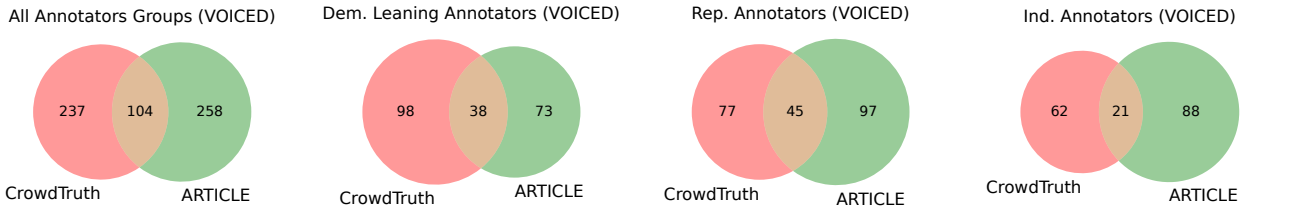


Figure 7: Annotators that are identified as unreliable based on CT and ARTICLE scores for $\mathcal{D}_{\texttt{VOICED}}$. The last three Figures show the same inconsistent annotators broken down by their political leaning. For $\mathcal{D}_{\texttt{VOICED}}$, CT ($WQS \geq 0.86$) and ARTICLE ($k \geq 0.46$).

References

Abercrombie, G.; Hovy, D.; and Prabhakaran, V. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *17th Linguistic Annotation Workshop 2023*, 96–103. Association for Computational Linguistics.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35: 3663–3678.

Carbonell, J. G.; Michalski, R. S.; and Mitchell, T. M. 1983. An overview of machine learning. *Machine learning*, 3–23.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; De Sa, C.; Grimmelmann, J.; Kleinberg, J.; Sen, S.; and Zhang, B. 2024a. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22004–22012.

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; De Sa, C.; Grimmelmann, J.; Kleinberg, J.; Sen, S.; and Zhang, B. 2024b. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. ArXiv:2301.11562 [cs, stat].

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1): 20.

Deng, N.; Zhang, X.; Liu, S.; Wu, W.; Wang, L.; and Mihalcea, R. 2023. You Are What You Annotate: Towards Better Models through Annotator Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12475–12498. Association for Computational Linguistics.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Dumitrache, A.; Inel, O.; Aroyo, L.; Timmermans, B.; and Welty, C. 2018a. CrowdTruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*.

Dumitrache, A.; Inel, O.; Aroyo, L.; Timmermans, B.; and Welty, C. 2018b. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. ArXiv:1808.06080 [cs].

Dutta, A.; Khorramrouz, A.; Dutta, S.; and KhudaBukhsh, A. R. 2024. Down the Toxicity Rabbit Hole: A Framework to Bias Audit Large Language Models with Key Emphasis on Racism, Antisemitism, and Misogyny. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, To appear. ijcai.org.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. ArXiv:2303.15056 [cs].

Gray, M. L.; and Suri, S. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Boston: Houghton Mifflin Harcourt. ISBN 978-1-328-56628-7.

He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4.

Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2nd Workshop on Computational Linguistics for Literature, CLfL 2013 at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2013*, 1120–1130. Atlanta, Georgia: Association for Computational Linguistics. ISBN 978-1-937284-47-3.

Huang, H.; Qu, Y.; Liu, J.; Yang, M.; and Zhao, T. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.

Huang, O.; Fleisig, E.; and Klein, D. 2023. Incorporating Worker Perspectives into MTurk Annotation Practices for NLP. ArXiv:2311.02802 [cs].

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, N.-J.; and Marneffe, M.-C. d. 2022. Investigating reasons for disagreement in natural language inference. *TACL*, 10: 1357–1374.

Kahneman, D.; Sibony, O.; and Sunstein, C. R. 2021. *Noise: a flaw in human judgment*. New York: Little, Brown Spark, first edition edition. ISBN 978-0-316-45140-6 978-0-316-26665-9. OCLC: on1249942231.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2021. We Don't Speak the Same Language: Interpreting Polarization through Machine Translation. In *AAAI 2021*, 14893–14901.

Kumar, D.; Kelley, P. G.; Consolvo, S.; Mason, J.; Bursztein, E.; Durumeric, Z.; Thomas, K.; and Bailey, M. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 299–318.

Lease, M. 2011. On quality control and machine learning in crowdsourcing. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*.

Leonardelli, E.; Menini, S.; Aprosio, A. P.; Guerini, M.; and Tonelli, S. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10528–10539. ArXiv:2109.13563 [cs].

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Nie, Y.; Zhou, X.; and Bansal, M. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *EMNLP*, 9131–9143.

OpenAI. 2022. ChatGPT (Jun 14). https://chat.openai.com. Gpt-3.5-turbo-0125.

Passonneau, R. J.; Bhardwaj, V.; Salleb-Aouissi, A.; and Ide, N. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46: 219–252.

Pavlick, E.; and Kwiatkowski, T. 2019. Inherent disagreements in human textual inferences. *TACL*, 7: 677–694.

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

Tan, S.; Zhuang, S.; Montgomery, K.; Tang, W. Y.; Cuadron, A.; Wang, C.; Popa, R. A.; and Stoica, I. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Uma, A. N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470.

Ustalov, D.; Pavlichenko, N.; and Tseitlin, B. 2024. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96): 6227.

Wang, A.; Hoang, C. D. V.; and Kan, M.-Y. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47: 9–31.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Weerasooriya, T.; Dutta, S.; Ranasinghe, T.; Zampieri, M.; Homan, C.; and KhudaBukhsh, A. 2023a. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11648–11668. Singapore: Association for Computational Linguistics.

Weerasooriya, T. C.; Dutta, S.; Ranasinghe, T.; Zampieri, M.; Homan, C. M.; and KhudaBukhsh, A. R. 2023b. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. arXiv. ArXiv:2301.12534 [cs].

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.