Assessing Crowdsourced Annotations with LLMs: Linguistic Certainty as a Proxy for Trustworthiness

Tianyi Li
Purdue University
li4251@purdue.edu

Divya Sree
Purdue University
divyasree080220@gmail.com

Tatiana Ringenberg
Purdue University
tringenb@purdue.edu

Abstract

Human-annotated data is fundamental for training machine learning models, yet crowdsourced annotations often contain noise and bias. In this paper, we investigate the feasibility of employing large language models (LLMs), specifically GPT-4, as evaluators of crowdsourced annotations using a zero-shot prompting strategy. We introduce a certainty-based approach that leverages linguistic cues categorized into five levels (Absolute, High, Moderate, Low, Uncertain) based on Rubin's framework—to assess the trustworthiness of LLM-generated evaluations. Using the MAVEN dataset as a case study, we compare GPT-4 evaluations against human evaluations and observe that the alignment between LLM and human judgments is strongly correlated with response certainty. Our results indicate that LLMs can effectively serve as a preliminary filter to flag potentially erroneous annotations for further expert review.

1 Introduction

Human-annotated data remains a cornerstone for training datasets in machine learning applications. However, crowdsourced annotations are often noisy and contain biases (Demszky et al., 2020; Edwin Chen, 2022; Stoica et al., 2020; Zhang et al., 2023; Wang et al., 2020). Additionally, the context and perspective of annotators can limit the accuracy and comprehensiveness of these annotations (Mena et al., 2020; Cao et al., 2023). In digital humanities, where subtle nuances and context is critical to interpretive accuracy, such biases and errors can undermine research findings. Therefore, it is essential to continuously evaluate and improve the quality of human annotations in existing datasets.

When working with existing annotations, researchers often lack access to the original annotators or their decision rationale. Even though dataset documentation frameworks (e.g. data statements or Data Cards) aim to improve transparency (Pushkarna et al., 2022), in practice most public datasets only provide the final labels. Validating crowdsourced annotations after the fact typically requires domain experts or trained annotators to re-annotate a sample and measure agreement (Davani et al., 2022). This approach is reliable but labor-intensive and costly, especially as datasets grow in size.

Recently, large language models (LLMs) have shown exceptional performance across various data annotation tasks (Tan et al., 2023; Jeblick et al., 2023; Gilardi et al., 2023; Goel et al., 2023). Yet, human input remains crucial in most annotation efforts. This paper explores the potential of LLMs in delivering reliable evaluations and supporting continuous improvements in the quality of crowdsourced data annotations.

We propose leveraging linguistic cues from LLM-generated evaluations to gauge the certainty of responses, using this certainty as an indicator of the trustworthiness of the evaluations. This paper presents preliminary results from applying this method, utilizing a zero-shot prompting strategy with GPT-4 to evaluate a general domain event dataset containing event-labeled sentences.

2 Related Work

2.1 Challenges in Crowdsourced Annotations

Significant research has focused on improving the quality crowdsourced annotations by identifying individual annotator patterns (Mena et al., 2020) and deriving reliability scores based on annotator expertise and task complexity (Cao et al., 2023). However, these approaches are primarily designed to optimize the annotation process itself. Furthermore, despite these efforts, many crowdsourced datasets still exhibit a significant number of labeling errors. For instance, the TACRED relation extraction dataset has an estimated 23.9% error rate (Stoica et al., 2020), the GoEmotions dataset may contain up to 30% incorrect labels (Demszky et al.,

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 191–201

2020; Edwin Chen, 2022), and a study by Zhang et al. found an error rate of approximately 25.79% in a sample of 10,000 instances from the MAVEN dataset (Zhang et al., 2023; Wang et al., 2020).

2.2 Supporting Data Annotation with LLMs

Previous studies (Brown et al., 2020) have demonstrated that a pre-trained LLM can achieve benchmark performance for NLP tasks like question answering (Tan et al., 2023), document summarization (Jeblick et al., 2023), text annotation tasks (Gilardi et al., 2023), without the need for fine-tuning. The research community is actively investigating the role of LLMs in data annotation and its advantages and disadvantages in different annotation tasks (Gilardi et al., 2023; Goel et al., 2023).

This paper contributes to this growing body of knowledge by investigating the possibility of using LLMs as evaluators, rather than annotators, of previously crowdsourced annotations.

3 Method

We propose leveraging linguistic cues from LLM-generated evaluations to measure response certainty, using this metric as a proxy for the evaluations' trustworthiness.

Prior research has introduced several methods for assessing certainty in LLM outputs. For example, logit-based approaches (Guo et al., 2017; Jiang et al., 2021) are frequently used to quantify uncertainty at the token level. Meanwhile, methods based on verbalized confidence (Lin et al., 2022; Kadavath et al., 2022) and consistency (Wang et al., 2022; Xiong et al., 2023) have been developed to evaluate overall response accuracy. However, LLMs can exhibit notable overconfidence when expressing uncertainty (Tanneru et al., 2023), and consistency-based methods tend to be computationally expensive (Chen and Mueller, 2023).

To evaluate the trustworthiness of LLM-generated assessments of data annotations, we propose an approach that uses **linguistic cues** to determine the certainty of these evaluations. This method is inspired by epistemic uncertainty theory, which notes that humans often signal their level of confidence with phrases like "I guess" or "It's likely." This method is based on the assumption that, although LLMs do not possess true epistemic certainty – they generate responses based on statistical likelihood – they nonetheless reflect uncertainty through similar linguistic markers. In this study, we adopt the standard Rubin's framework (Rubin, 2006) to identify such cues in LLM responses.

3.1 Theoretical Backgrounds

Existing literature on pragmatics and discourse addresses textual certainty through various interrelated linguistic concepts. For example, *hedging* refers to the use of words that render a phrase more ambiguous, thereby introducing speculation (Lakoff, 1973). Vincze (Vincze, 2014) and Szarvas (Szarvas et al., 2012) categorize it under semantic and discourse certainty, and Sauri links textual certainty to factuality (Saurí and Pustejovsky, 2012). Rubin (Rubin, 2006) synthesized these perspectives, clarifying that certainty can be understood through three main linguistic dimensions: epistemic modality, evidentiality, and hedging.

Epistemic modality refers to the speaker's degree of confidence in a proposition, typically expressed through words such as "think" or "may" (Coates, 1987). Statements that include these markers are explicitly qualified for certainty, while those lacking them are implicitly certain. For example, "His feet were blue" is implicitly certain, whereas "His feet were *sort of* blue" is explicitly uncertain due to the hedge "sort of."

Evidentiality evaluates the trustworthiness of information by considering its source. This concept overlaps with epistemic modality by incorporating the speaker's attitude toward knowledge (Chafe and Nichols, 1986). Chafe expands evidentiality to encompass both the evidence supporting a claim and the attitude toward that evidence, a perspective that Rubin uses to interpret textual certainty.

Hedging serves to introduce uncertainty or soften assertions, using single words or phrases such as "in my opinion" (Vincze, 2014; Hyland, 1998; Brown and Levinson, 1987). Rubin's framework leverages these concepts by identifying certainty markers and categorizing them as Absolute, High, Moderate, Low, and Uncertain.

We applied Rubin's guidelines (Rubin, 2006) to identify these markers and assign corresponding certainty levels to LLM responses. These aggregated certainty levels then provide a means to evaluate the trustworthiness of LLM-generated annotation evaluations.

3.2 Study Design

In this study, we investigate the use of linguistic cues to assess the certainty level of LLM responses

as a metric for assessing LLM-generated annotation evaluations. Specifically, we aim to answer the following research questions (RQs):

RQ1: How effectively can a large language model like GPT-4 identify correct vs. incorrect annotations in a crowdsourced dataset? We measure effectiveness by comparing the LLM's judgments with those of human evaluators on the same data.

RQ2: In the context of annotation evaluation, how does GPT-4 linguistically express certainty or uncertainty about its judgments? We qualitatively and quantitatively examine the language used in GPT-4's responses (e.g., usage of modal verbs, hedges, or confident assertions).

3.2.1 Dataset and Baseline

We evaluated the crowdsourced annotations in the MAVEN dataset (Wang et al., 2020), a general-domain event detection (ED) resource comprising annotations for 4,480 Wikipedia documents. The dataset features a diverse array of trigger words paired with event types, as defined by the frames in FrameNet (Baker et al., 1998). A trigger word is typically a verb or noun that signals the occurrence of an event, while an event label is a predefined category in the MAVEN event schema assigned to that trigger word (Consortium et al., 2005).

Zhang et al. (Zhang et al., 2023) evaluated the MAVEN annotations and flagged disagreements with the crowd-sourced labels as debatable. For example, consider the sentence:

> 46 seconds later the plane crashed (CATASTROPHE) and burned (BODILY_HARM) 1335 meters from the threshold.

In this case, crowd workers identified crashed and burned as trigger words, assigning the labels CATASTROPHE and BODILY_HARM, respectively. While evaluators agreed that crashed correctly indicates a CATASTROPHE event, they disputed the BODILY_HARM label for burned, marking it as a debatable annotation. Although the evaluators did not propose an alternative label, it can be inferred that burned describes the condition of the plane rather than implying bodily harm.

All debatable annotations identified by the evaluators (Zhang et al., 2023) are publicly available[1]. In our study, we use these human evaluations as the baseline to assess the quality of LLM-generated evaluations of crowd annotations.

3.2.2 LLM Configuration

State-of-the-art LLMs vary in their training strategies, model architectures, and intended use cases, with performance largely influenced by factors such as pre-training, fine-tuning, and test data (Yang et al., 2024). In our study, we employed OpenAI's GPT-4 to evaluate crowd-sourced annotations, given its strong performance across multiple benchmarks (Brown et al., 2020; Tan et al., 2023; Jeblick et al., 2023). The experiments were conducted from January to March 2024.

For each API request, we set the temperature to 0.6, following the recommendations in the ChatGPT-4 technical report (Achiam et al., 2023). We then iteratively refined our prompts based on several considerations (DAIR.AI, 2024). First, we used clear command verbs—such as "Assess," "Evaluate," and "Identify", to instruct the model. We found that "Evaluate" yielded the best results. We also focused on positive, specific instructions rather than emphasizing what the model should avoid. After testing several iterations of prompt design, we settled on the following prompt:

> *Evaluate the choice of the word <Trigger word> as a trigger word signifying the event <Event Label> in the sentence <Sentence>. Please explain. If you disagree with the event label for the word <Trigger word>, propose a new event label.*

In the study, <Trigger word>, <Event Label> and <Sentence> are replaced by the actual trigger words, event labels, and sentences.

3.2.3 Evaluation Generation

We randomly selected a sample of 40 sentences from the list of debatable annotations (Zhang et al., 2023). This sample contained a total of 113 event labels identified by crowd annotators. Among these, the human evaluators from Zhang's (Zhang et al., 2023) study agreed with 86 of the crowd event labels and disagreed with 27. In other words, around 24% of the crowd-sourced annotations were considered as debatable. We sent 113 API requests to OpenAI's GPT-4 model, each containing a unique prompt with the trigger words, event labels, and sentences from the sample annotations.

[1] http://edx.leafnlp.org/event_detection/data/debatable_annotations

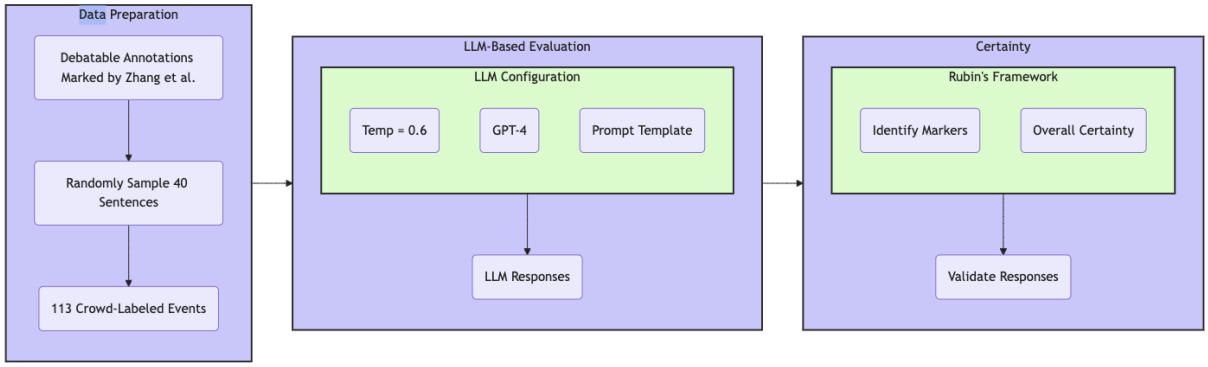

Figure 1: Study procedure of using LLM to evaluate crowdsourced annotations. The study followed three main steps: (1) sampling crowdsourced annotations for evaluation, (2) configuring LLM and generating evaluations, and (3) assessing certainty of LLM-generated evaluations.

The responses are then translated into evaluation results based on whether GPT-4 *agreed* or *disagreed* with the crowd-sourced event labels. After that, each response was labeled with a certainty level based on linguistic cues.

3.3 Certainty Analysis

Previous studies have provided lexical lists of certainty markers (Prokofieva and Hirschberg, 2014), but these lists are domain-specific and not directly applicable to our annotation evaluation scenario. To address this limitation, we extended the existing lists by carefully examining the context surrounding each sentence in the GPT-4 responses and identifying additional certainty markers. Each marker was then assigned one of five certainty levels: *Absolute*, *High*, *Moderate*, *Low*, or *Uncertain*, using Rubin's annotation guidelines (Rubin, 2006).

Identification of Certainty Markers. We analyzed each GPT-4 response to detect both explicit and implicit expressions of certainty. If no explicit markers were present, the sentence was considered implicitly certain. For example, in the sentence "I *think* he bought it for $100," the word "think" explicitly indicates that the statement is an opinion, suggesting moderate certainty. In contrast, the sentence "I *am positive* it was he who bought a mower last week" contains the marker "am positive," conveying high certainty. A sentence such as "Wayne Storick, a 35-year-old contract laborer, bought his mower for $100" lacks any certainty marker and is therefore treated as implicitly certain.

Handling Ambiguity. For sentences where it was unclear whether an explicit certainty marker was present, we adopted the following strategies: *(1) Paraphrasing:* We reworded the sentence to determine if the conveyed confidence level changed. *(2) Auditory Assessment:* We read the sentence aloud to assess its inherent certainty. *(3) Marker Removal:* We evaluated the impact of removing potential markers to observe any shift in the certainty level. *(4) Consistency Check:* We compared sentences within the broader evaluation context to ensure consistency.

For these sentences, the labeling decision was reviewed and discussed by all authors until consensus was reached.

Sentence	Certainty level
He is ***destined*** to be famous	Absolute certainty
He foresaw a ***probable*** loss	High certainty
I ***think*** he bought it for $100	Moderate certainty
He ***may*** need more work	Low certainty
We can ***not know*** what will happen	Uncertain

Table 1: Examples of certainty markers for the five certainty levels

Assignment of Certainty Levels. Based on the classification defined in Rubin's study (Rubin, 2006), each certainty marker was assigned to one of the five levels: Absolute, High, Moderate, Low, or Uncertain. Table 1 shows example markers corresponding to each level.

4 Results

From the 113 LLM-generated evaluations, each for one event label, we identified a list of 67 certainty markers to assess the certainty levels of LLM responses. We found 60 (52%) of the evaluations expressed absolute and high certainty, while 45 (39%) of the evaluations expressed moderate certainty and only 8 (7%) of the evaluations expresses low certainty. Overall, GPT-4 agreed with 87 labels

and disagreed with 26 labels.

4.1 Distribution of Certainty Markers

Each evaluation contained multiple instances of certainty markers, with a total of 490 instances identified across all responses. Among these, 35 instances indicated absolute certainty, and these were nearly evenly distributed between cases where GPT-4 agreed with and disagreed from the crowd annotations (see Figure 2). One-third of the markers (N=163) signified high certainty, with 68.71% appearing in evaluations that agreed with the crowd annotations. More than half of the markers (N=248) expressed moderate certainty, with 77.42% found in cases where GPT-4 concurred with the crowd annotations. Finally, the 43 instances indicating low certainty were evenly distributed between agreement and disagreement cases, and there was only one instance of uncertainty in evaluations where GPT-4 agreed with the crowd.

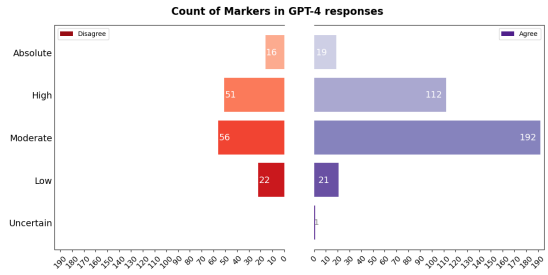

Figure 2: Distribution of certainty markers in GPT-4 responses. All the certainty markers in the entire responses are counted.

4.2 Certainty Levels of LLM Evaluations

Since LLM responses often contained multiple sentences, the certainty markers from different sentences of the same response could express different certainty levels. In addition, we have observed cases where GPT-4 initially agrees with the given crowd annotation but as it explains its response, it changes its decision. In these cases, one evaluation could contain two conflicting decisions and hence, making it difficult to aggregate the corresponding certainty markers. Thus, the certainty level of each evaluation cannot be represented by counting the certainty markers. The frequency of these markers, the relative length of the responses, as well as the sentiment of the decisions expressed, all influence the certainty of the LLM-generated evaluations.

Therefore, we use the claim sentences where GPT-4 clearly stated its agreement or disagreement with crowd labels to determine the certainty of each evaluation response.

For example, in the following GPT-4 response, the first sentence is the claim sentence that expresses agreement with the crowd annotation. In this sentence, there is a certainty marker "somewhat", that expresses LOW CERTAINTY. Although this response also had other certainty markers like "more" and "could", this evaluation is considered as LOW CERTAINTY.

> *The choice of the word captured as a trigger word signifying the event Conquering in this context is somewhat (*LOW CERTAINTY*) inaccurate.* Although capturing can be a part of the conquering process, it does not necessarily (HIGH CERTAINTY) denote conquering itself. ..... A more (HIGH CERTAINTY) accurate event label could (MODERATE CERTAINTY) be Possessing or Seizing.

Using this method, we classified the certainty level of each LLM-generated evaluation (Figure 3). Most of the evaluations agreeing with crowd annotations expressed moderate, high, or absolute certainty. When disagreeing with crowds, the evaluations are usually moderate to low certainty.

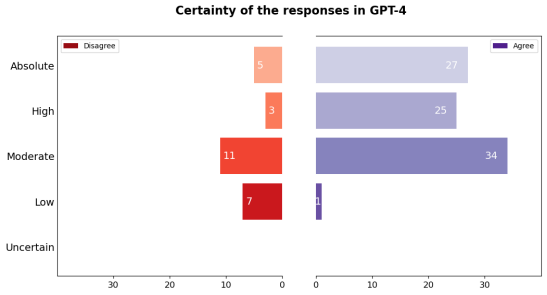

Figure 3: Certainty of GPT-4 Evaluations. The certainty markers appeared in the claim statement (indicating whether GPT-4 agrees or disagrees with crowd annotations) were used to determine the evaluation certainty.

4.3 Comparing LLM and Human Evaluations

Figure 4 shows the agreement between LLM and crowds versus human evaluator from (Zhang et al., 2023) and crowds, where the overlapping areas refer to where the evaluations agree with crowd annotations.

Overall, LLM agreed with 87 of the 113 crowd annotations and human evaluators agreed with 86. Among those, 72 event labels overlap, where both

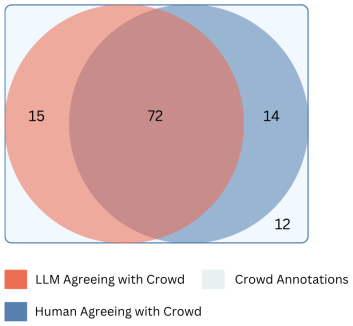

Figure 4: LLM evaluation vs. Human Evaluation. The rectangle shape presents all the crowd annotations being evaluated. The left circle represents cases where LLM agrees with crowd annotations. The right circle represents cases where human evaluators agree with crowd annotations.

the LLM and human evaluators agreed with the crowd annotations. Among the 26 event labels where human evaluators disagreed with the crowd annotations, the LLM also disagreed with 12 of them. This leads to an observed agreement of 74.3% ($\frac{72+12}{113}$) between LLM and human evaluators. However, after accounting for the possibility of agreement occurring by chance, as measured by Cohen's Kappa, the agreement beyond what would be expected by chance is only fair ($\kappa = 0.286$). Thus, we further analyzed the discrepancies between LLM and human evaluators.

There are 14 crowd annotations with which **human evaluators agreed, but the LLM disagreed with crowd-sourced labels**. For 6 of those differences, we found that the discrepancies were likely due to the LLM's lack of awareness of the semantic frames associated to certain event types in the MAVEN dataset. As the definitions of event labels [2] were not provided in the prompts, GPT-4 evaluated the crowd annotations with the literal meaning of the event labels rather than the rules defined in the annotation guidelines. Sometimes LLM can also go wrong as they make up the assumption around the context.

For example, LLM disagreed with crowd-sourced event label INFLUENCED as it assumed that the road was damaged due to ravines and suggested DAMAGE event label.

> In Mehedini County, county road DJ607C and local road DC4 were affected (INFLUENCED) , due to the formation of transversal and longitudinal ravines.

Sometimes, the event labels suggested by LLM could also be reasonable, and thus point to ambiguous cases. For example, the following sentence was annotated to have a trigger word "district" that mention a PLACING event label by crowd. "Placing" event label is associated with the BEING_LOCATED semantic frame. BEING_LOCATED is defined as "A Theme is in a stable position with respect to a Location. " (Baker et al., 1998). Without this knowledge, GPT-4 disagreed with crowd event label and suggested "Location" or "Geographical_Entity" as the correct event label.

> Fighting was mostly concentrated in the inner city Chinese business district (PLACING) of Cholon.

Conversely, it appears that the human evaluations might contain mistakes as well. For instance, human evaluators have agreed with the crowd event label for the following sentence.

> Although 6 RAR ultimately prevailed (CONVINCING) , the vicious fighting at "ap my an" was probably the closest the australian army came to a major defeat during the war.

The event label CONVINCING refers to act of persuading someone (Consortium et al., 2005), however, the trigger word prevailed refers to act of being victorious (Consortium et al., 2005). GPT-4's evaluation concurred with this and provided "Winning" as the correct event label.

Furthermore, there are 15 event labels with which **human evaluators disagreed but LLM agreed with crowd-sourced labels**. Similar to the where LLM mistakenly disagree with crowd annotations, it may mistakenly agree with the provided crowd label without recognizing a more accurate alternative due to lack of the overall semantic frames. For example,

> Vehicles and houses were burned and stores owned by Chinese were plundered (THEFT).

while "theft" captures the act of stealing, a more appropriate label would be "robbery," as it specifies the use of force (Consortium et al., 2005).

[2] https://arxiv.org/pdf/2004.13590

On the other hand, the LLM might overlook key contextual information and thus fail to identify incorrect annotations. It might be prioritizing specific keywords within event labels without considering the broader context of the sentence. For instance,

> Although the helicopter's loss was initially blamed on enemy action, a subsequent inquiry (CRIMINAL_INVESTIGATION) found Cardiff's missile to be the cause.

the lack of context regarding legal charges or criminal activity makes "criminal_investigation" inaccurate. Yet LLM agreed with this annotation, probably focused solely on the keyword "investigation" and disregarded the surrounding context, leading it to agree with the crowd-sourced label.

Four of these cases, however, revealed human evaluators' mistakes. For example,

> The exclusion zone was later increased to radius when a further 68000 people were evacuated (EMPTYING)from the wider area.

EMPTYING reflects the act of removing represented by the evacuated. We also did not find a more appropriate event label in the MAVEN schema to disagree with the event label.

4.4 LLM is more certain when the evaluation is aligned with human evaluators

	Value	df	AS(2)
Pearson Chi-Square	11.688	4	.020
Likelihood Ratio	10.111	4	0.039
No of Valid Cases	113		

Table 2: Association between certainty of claim sentences (X) vs GPT-4's agreement with dedicated annotators' evaluation (Y). AS(2) means Asymptotic Significance (2-sided).

We conducted Chi-squared test of independence and found that there is a significant association between the LLM evaluation certainty and whether the evaluation agrees with human evaluations (p=0.02, $\chi^2 = 11.69$). Further analysis using logistic regression shows that the GPT-4 evaluations are more likely to express low certainty or uncertainty when the evaluation results are different from human evaluators (p =0.019, Exp(B) = 7.912). This indicates that low certainty markers can potentially act as an indictor of discrepancies between LLM and human evaluators, and thus prioritize human evaluation efforts to those low-certainty cases.

5 Discussion

In this study, we investigated the feasibility of employing a large language model (LLM) as an evaluator of crowd-sourced annotations using a zero-shot prompting strategy. We introduced a certainty-based approach to assess the trustworthiness of LLM-generated evaluations. Guided by Rubin's framework (Rubin, 2006), we developed a list of certainty markers, categorized their certainty levels, and analyzed the relationship between evaluation quality and these certainty measurements. To validate our approach, we compared the LLM-generated evaluations with human evaluations from (Zhang et al., 2023) as the baseline.

5.1 Using LLMs to Support Human Evaluators

Our findings reveal that the alignment between LLM and human evaluations is strongly correlated with response certainty. Consequently, LLMs can serve as an effective preliminary filter for detecting potential errors in crowd-sourced annotations. By evaluating annotations and flagging those that exhibit low certainty markers, LLMs can identify cases requiring expert review. This targeted approach enables human evaluators to focus their efforts on these flagged instances, thereby enhancing overall efficiency.

Identifying Initial Errors Our results indicate that the alignment between LLM and human evaluations is significantly correlated with response certainty. Therefore, LLMs can serve as an initial filter to identify potential errors in crowd-sourced annotations. By evaluating annotations and flagging those with low certainty markers, LLMs can highlight cases that require expert attention. This allows human evaluators to focus on reviewing these flagged cases and improve efficiency.

Providing Additional Insights When LLMs disagree with crowd annotations, they often provide detailed explanations. These explanations can offer valuable insights and alternative perspectives that human evaluators might not have considered, enriching the evaluation process and potentially leading to more accurate conclusions.

Enhancing Consistency and Reducing Bias Human evaluators can introduce biases and inconsis-

	B	S.E.	Wald	df	Sig.	Exp(B)
Absolute	-0.144	0.885	0.027	1	0.870	0.866
High	-0.194	0.442	0.192	1	0.661	0.824
Moderate	-0.202	0.437	0.214	1	0.644	0.817
Low	2.068	0.885	5.468	1	**0.019**	7.912
Constant	-1.088	0.399	7.433	1	**0.006**	0.337

Table 3: Logistic regression model results: the impact of predictor variables (frequency of the five levels of certainty markers: Absolute, High, Moderate, Low, Uncertain) on whether the LLM agrees with human evaluations. B refers to the regression coefficient, S.E. (Standard Error) is the Variability of the coefficient estimate, wald test is coefficient divided by its standard error, the Sig. value is the p-value statistic and Exp(B) represents the odds ratio.

tencies in their assessments due to subjective interpretations or fatigue. LLMs, with their ability to process large amounts of data consistently, can help mitigate these issues. By cross-referencing LLM evaluations with human assessments, discrepancies can be identified and addressed, promoting greater consistency and reducing individual biases in the annotation process.

Facilitating Continuous Improvement The use of LLMs in the evaluation process can contribute to continuous improvement in data quality. An interesting and important future direction is to investigate the use of LLMs to conduct more targeted evaluations, checking for specific biases or other fairness concerns in existing labels. This iterative process is crucial for both human and AI components of the evaluation system to evolve and improve progressively.

5.2 Implications for Annotation Evaluation

5.2.1 LLM configuration

In designing the prompt for our study, we have only provided the instruction in the user message, deliberately excluding information related to MAVEN Event Schema. This approach aimed to avoid overloading the user message and to maintain task specificity, while also reducing the cost of each API request. For larger scale evaluation, future research is needed to investigate the balance between model fine-tuning and system/user prompt engineering to optimize the performance and costs of using LLM as an annotation evaluator.

Additionally, we chose not constrain the response format, and did not set any limit for max_token, to let the LLM use the context length without having any constraints. This approach aimed to give the LLM the freedom to generate responses using linguistic cues and observe its natural tendencies in providing evaluations. Our results suggest that this method is effective but future research could further investigate optimal settings to enhance performance.

5.2.2 Certainty marker identification

In our study, we manually identified certainty markers within LLM responses following the Rubin's framework (Rubin, 2006). This is because the existing lexical lists of certainty markers were tailored for specific writing styles and domains, and there were no pre-existing list designed specifically used by LLMs. However, this manual identification can introduce biases.

We have developed and shared a list of certainty markers for assessing the certainty of LLM's annotation evaluations. This serves as a first step towards a collaborative research effort aimed at enhancing the list of certainty markers and potentially training machine learning models to automatically detect these markers within LLM responses.

6 Conclusion

This study has explored the feasibility and potential of using large language models (LLMs), specifically GPT-4, to evaluate crowdsourced annotations. Our findings indicate that LLMs can significantly align with human evaluators, achieving a substantial portion (74.3%) of agreement. This opens exciting avenues for utilizing LLMs as a complementary tool to assess the quality of crowdsourced data especially in domains where manual validation is expensive or time-consuming.

Our approach of using linguistic cues for certainty assessment proved effective, providing a reliable metric for assessing the LLM-generated evaluation quality. In conclusion, this research paves the way towards the integration of LLMs like GPT-4 into crowdsourced annotation validation workflows. Further research on improving LLM certainty calibration and targeted training for specific annotation tasks can further enhance their reliability and effectiveness in data validation tasks.

Limitations

Despite the promising findings, several limitations warrant discussion. First, the non-deterministic nature of LLMs means that responses can vary with repeated queries, making consistency a challenge. Additionally, GPT-4's performance is sensitive to prompt design, and the absence of domain-specific semantic frame information in the prompts may lead to ambiguous or incorrect evaluations. Future work should explore variance in responses over multiple queries and prompting strategies.

Our approach also relies on manually curated certainty markers. While guided by Rubin's framework, this may introduce subjective bias and limit reproducibility across different domains or datasets.

Moreover, when dealing with confidential or sensitive information, the use of third-party LLMs raises privacy concerns due to potential data exposure. Finally, evidence of stereotyping bias within LLM responses suggests that while LLMs can serve as effective initial filters, they should not replace human oversight; instead, they must be integrated into a hybrid evaluation framework that leverages the complementary strengths of both human expertise and artificial intelligence.

Ethics Statement

Although studies such as Gilardi et al. (Gilardi et al., 2023) have suggested that LLMs like GPT-4 can outperform crowd annotators in certain tasks, their performance varies depending on the task, dataset, and label set employed (Zhu et al., 2023; Wang et al., 2024). Moreover, when dealing with confidential or sensitive information, using LLMs as evaluation tools poses risks related to data exposure to third parties that own the models.

Our findings further reveal the presence of stereotyping bias in LLM responses. For example, consider the sentence:

> "Although the helicopter's loss was initially blamed on enemy action, a subsequent inquiry (*Criminal_Investigation*) found Cardiff's missile to be the cause."

Here, the term "inquiry" is used in a non-criminal context. However, due to its frequent association with criminal investigations in the training data, the model stereotypes "inquiry" as primarily linked to criminal contexts, regardless of the actual usage.

These limitations indicate that relying solely on LLMs for evaluation is not advisable. Human oversight is crucial to ensure fairness and mitigate potential biases in the evaluation process. Our results suggest that LLMs can serve as an initial filter to assess crowd work and complement human evaluators, thereby focusing human effort on reviewing cases that require additional scrutiny. An important avenue for future research is to further explore human-AI collaboration, leveraging the complementary strengths of both to promote fairness and reduce bias.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhi Cao, Enhong Chen, Ye Huang, Shuanghong Shen, and Zhenya Huang. 2023. Learning from crowds with annotation reliability. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2103–2107.

Wallace L Chafe and Johanna Nichols. 1986. *Evidentiality: The linguistic coding of epistemology*, volume 20. Ablex Publishing Corporation Norwood, NJ.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness.

Jennifer Coates. 1987. Epistemic modality and spoken discourse. *Transactions of the Philological society*, 85(1):110–131.

Linguistic Data Consortium et al. 2005. Ace (automatic content extraction) english annotation guidelines for events version 5.4. 3. *ACE*.

DAIR.AI. 2024. General tips for designing prompts.

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

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Edwin Chen. 2022. 30 percent of google's emotions dataset is mislabeled.

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

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ken Hyland. 1998. Hedging in scientific research articles. *Hedging in scientific research articles*, pages 1–317.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2023. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, pages 1–9.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Francisco Mena, Ricardo Ñanculef, and Carlos Valle. 2020. Collective annotation patterns in learning from crowds. *Intelligent Data Analysis*, 24(S1):63–86.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

Victoria L Rubin. 2006. Identifying certainty in texts. *Unpublished Doctoral Thesis, Syracuse University, Syracuse, NY*.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational linguistics*, 38(2):261–299.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2020. Re-tacred: A new relation extraction dataset. In *Proceedings of the 4th Knowledge Representation and Reasoning Meets Machine Learning Workshop (KR2ML 2020), at NeurIPS, Virtual*, pages 11–12.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Quantifying uncertainty in natural language explanations of large language models. *arXiv preprint arXiv:2311.03533*.

Veronika Vincze. 2014. Uncertainty detection in natural language texts. *PhD, University of Szeged*, 141.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. *arXiv preprint arXiv:2004.13590*.

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 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Wenlong Zhang, Bhagyashree Ingale, Hamza Shabir, Tianyi Li, Tian Shi, and Ping Wang. 2023. Event detection explorer: An interactive tool for event detection exploration. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 171–174.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.