ChatGPT Rates Natural Language Explanation Quality Like Humans: But on Which Scales?

Fan Huang[1], **Haewoon Kwak**[1], **Kunwoo Park**[2,3], **Jisun An**[1]
[1]Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA
[2]School of AI Convergence, Soongsil University, Seoul, South Korea
[3]Department of Intelligent Semiconductors, Soongsil University, Seoul, South Korea
huangfan@acm.org, haewoon@acm.org, kunwoo.park@ssu.ac.kr, jisun.an@acm.org

Abstract

As AI becomes more integral in our lives, the need for transparency and responsibility grows. While natural language explanations (NLEs) are vital for clarifying the reasoning behind AI decisions, evaluating them through human judgments is complex and resource-intensive due to subjectivity and the need for fine-grained ratings. This study explores the alignment between ChatGPT and human assessments across multiple scales (i.e., binary, ternary, and 7-Likert scale). We sample 300 data instances from three NLE datasets and collect 900 human annotations for both *informativeness* and *clarity* scores as the text quality measurement. We further conduct paired comparison experiments under different ranges of subjectivity scores, where the baseline comes from 8,346 human annotations. Our results show that ChatGPT aligns better with humans in more coarse-grained scales. Also, paired comparisons and dynamic prompting (i.e., providing semantically similar examples in the prompt) improve the alignment. This research advances our understanding of large language models' capabilities to assess the text explanation quality in different configurations for responsible AI development.

Keywords: Natural Language Explanation, Large Language Models, ChatGPT, Text Quality Evaluation

1. Introduction

Artificial intelligence (AI) is progressively becoming an integral part of our daily lives, emphasizing the need for transparent (Saxon et al., 2021; Wu et al., 2023) and responsible (Bergman and Diab, 2022) AI systems. An essential element in achieving transparency and building trust between such systems and users is the generation of natural language explanations (NLE) (Kumar and Talukdar, 2020). The NLEs play a crucial role in clarifying the reasoning behind AI decisions. As the significance of NLEs continues to grow, it has become increasingly important to evaluate the quality of these explanations (Yao et al., 2023a).

Traditionally, evaluating NLEs has largely relied on gathering human judgments (Clinciu et al., 2021; Yao et al., 2023a). Assessing text quality through human evaluation is a crucial yet intricate endeavor (van der Lee et al., 2019; Yao et al., 2023a). This complexity arises from two key factors: the inherently subjective nature of human text quality assessments (Yao et al., 2023a) and fine-grained ratings on a Likert scale (van der Lee et al., 2019). Furthermore, it is challenging to eliminate unintended biases in question wording (Schoch et al., 2020) or participant recruitment (Kwak et al., 2022) in collecting human responses. Consequently, human evaluation can be resource-intensive and time-consuming. Developing models capable of autonomously assessing explanation quality could be a valuable complement to human evaluations, which is a critical step toward responsible AI systems (Chiang and Lee, 2023).

The emergence of the new generation of large language models (LLMs), such as InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2022), has demonstrated remarkable ability in understanding natural language. These models have leveraged extensive knowledge accrued during training to outperform prior approaches in various tasks, including open-domain QA, document summarization, and mathematical reasoning (Qin et al., 2023; Wang et al., 2023b; Bang et al., 2023). ChatGPT has also exhibited human-level competency in generating informative and clear NLEs, especially in contexts like hate speech detection (Huang et al., 2023b). This progress naturally leads to whether LLMs can evaluate the quality of explanations. As AI-driven systems play pivotal roles in applications where explaining their decisions is imperative, ensuring the accuracy and alignment of LLM's assessments with human judgments becomes increasingly essential. While numerous studies have investigated the potential of LLMs to replace or augment human annotations, the primary focus has been on classification tasks such as topic and stance detection (Yi Liaw et al., 2023; Gilardi et al., 2023; Huang et al., 2023b), with little attention given to their ability to assign ratings or distinguish among ordinal categories, a focus of our study.

In this study, we delve into the alignment between ChatGPT's evaluation of explanation quality and human assessments using three distinct datasets:

LREC-COLING 2024, pages 3111–3132
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

e-SNLI for logical reasoning (Camburu et al., 2018), LIAR-PLUS for misinformation justification (Alhindi et al., 2018), and Latent Hatred for implicit hate speech explanation (ElSherief et al., 2021). We engage human annotators and ChatGPT to assess human-written explanations, focusing on two widely used metrics, informativeness and clarity (Gatt and Krahmer, 2018; Howcroft et al., 2020; Clinciu et al., 2021), which are in 7-point Likert scale. We employ various experiments, framing the problem as a classification task and pairwise comparisons as they can be easier (Herbrich et al., 2006) and reduce biases and errors (Narimanzadeh et al., 2023) for human annotators, to evaluate ChatGPT's capability to approximate human evaluations of explanation quality. Moreover, given the promising advantages of the in-context learning prompting approach (Xie et al., 2021), we investigate the potential benefits of dynamic prompting, in which we provide customized instruction with the most semantically similar examples.

Our findings reveal that ChatGPT's evaluations closely approximate human assessments in coarse-grained (binary and ternary classifications) but encounter challenges in fine-grained (7-way classification) assessments. Notably, ChatGPT excels in comparative evaluations akin to human judgment. The introduction of dynamic prompting shows promise in improving results, highlighting ChatGPT's potential to reduce the cost of data annotation for subjective metrics. This research contributes to our understanding of LLM's capacity to evaluate the explanations, adding valuable insights for ChatGPT's potential role in replacing or supplementing manual evaluations when building responsible AI systems.[1]

2. Related Works

Natural Language Explanation Advanced language models like GPT-3 (Brown et al., 2020) and its successors have already shown the great capability of providing natural language explanations (Wang et al., 2023a). For example, Huang et al. (2023b) find that ChatGPT can generate human-level quality text, and people prefer ChatGPT-generated text to human-written text in terms of informativeness and clarity evaluation perspectives. Also, studies have shown that it is challenging to discern ChatGPT-generated text from human-written text reliably (Sadasivan et al., 2023).

In the complex environment of evaluating advanced LLMs, choosing a tool that provides depth and precision of analysis is critical. Natural language explanation (NLE), with its inherent complexity and nuance, provides a vital avenue to perceive better the limitations and capabilities of advanced LLMs (Camburu et al., 2018; Alhindi et al., 2018; Clinciu et al., 2021; ElSherief et al., 2021). The nuanced features of NLEs allow for finer-grained evaluation from various aspects (Cambria et al., 2023), providing the fertile foundation for in-depth study of the multi-dimensional capabilities of LLMs from understanding linguistic subtleties to assessing cognitive consistency in generated responses (Camburu et al., 2018; Emelin et al., 2021; Sakai et al., 2021).

More human-preferred NLEs are generated through the ChatGPT, compared with the collected ground truth human annotations explaining the implicit online hate speech (Huang et al., 2023b). The MTurk workers also use ChatGPT to provide textual annotations, and it is hard for the task requester to distinguish the human-written and machine-generated textual annotations (Veselovsky et al., 2023). A handful of studies investigated how well those advanced LLMs can replace human annotation results (Gilardi et al., 2023; Zhu et al., 2023).

Text Quality Evaluation Despite its importance, the evaluation of text quality remains not comprehensively studied (Clinciu et al., 2021). The text quality evaluation generally falls into two categories: automatic metrics and human evaluation metrics. The automatic quality evaluation approach is based on calculating word overlap or semantic similarity between generation results and ground truth references, like BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) metrics. However, they are not always correlated with human evaluations (Clinciu et al., 2021). On the other hand, various human evaluation metrics have been proposed for text quality. Emelin et al. (2021) uses Coherence and Plausibility to measure the quality of explanations. Forbes et al. (2020) applies Relevance to measure whether the statements about social norms correlate well with the provided contextual information considering the social norms. The Valid, Satisfactory, and Shortcoming scores measure the natural language description quality under the few-shot out-of-domain scenarios (Yordanov et al., 2021). However, those metrics are domain-specific (Sakai et al., 2021; Forbes et al., 2020) or application-specific (Emelin et al., 2021; Chiyah Garcia et al., 2018).

3. Data and Annotation

To properly evaluate how well ChatGPT can judge the quality of explanations, it is important to test it with complex NLEs that might be challenging for the general public to understand. By examining how ChatGPT assesses the quality of explanations in these intricate scenarios, we can determine if the model can approximate human-like understanding and judgment.

[1] Our code and data are publicly available at https://github.com/muyuhuatang/ChatGPTRater.

We sample data instances from three existing NLE benchmark datasets. We then evaluate the quality of the explanations in the selected data instances using two widely used metrics: informativeness (Novikova et al., 2018) and clarity (van der Lee et al., 2017). Informativeness measures how relevant the information in the explanation is for understanding the underlying reasoning or justification, and clarity assesses how clearly the ideas in the explanation are expressed. Both informativeness and clarity are rated on a 7-point Likert scale, ranging from 1 (not informative/unclear) to 7 (very informative/very clear). We collect annotations from both trained annotators and ChatGPT to compare their assessments of explanation quality.

3.1. Datasets

We examine ChatGPT through the lenses of logical reasoning, misinformation justification, and implicit hate speech explanation, encompassing a spectrum of complexities and nuances of human language and cognition. We use three datasets: the e-SNLI (logical reasoning) (Camburu et al., 2018), the LIAR-PLUS (fake news justification) (Alhindi et al., 2018), and the Latent Hatred dataset (implicit hate speech explanation) (ElSherief et al., 2021).

e-SNLI dataset. One data instance includes a premise, a hypothesis, a label (contradiction, neutral, or entailment), and an explanation of the relationship between the premise and hypothesis. The label and explanations are annotated by experienced human annotators (native English speakers) hired on the Amazon Mechanical Turk platform (Camburu et al., 2018). We select the e-SNLI dataset to examine the capability of LLMs to understand and evaluate logical reasoning (Yu et al., 2023).

LIAR-PLUS dataset. Claims are labeled as pants-fire, false, barely-true, half-true, mostly-true, or true, and justifications for those claims, which may contain misinformation, are extracted from an existing corpus of fact-revealing human-written articles (Alhindi et al., 2018). We consider the justification as NLE. We select the LIAR-PLUS dataset to examine the capability of LLMs to understand and discriminate counterfactual information within context.

Latent Hatred dataset. It includes explanations of why a text is considered to indicate implicit hate within the provided context. Experienced research assistants are hired to provide the ground truth NLE annotations in the Latent Hatred dataset (ElSherief et al., 2021). The explanations in the first two datasets rely on factual information, making them inherently less subjective than those in the Latent Hatred dataset, as deciding what qualifies as hateful can vary among individuals (Gordon et al., 2022).

For the analyses in this work, we randomly sample 100 data instances from each of the three datasets, keeping the label distribution similar to the original datasets. Table 1 shows an example instance of each dataset.

3.2. Evaluation of NLE Quality

3.2.1. Human Annotation

We initially attempted to annotate the data using Amazon Mechanical Turk (MTurk). To ensure the quality of the annotations collected from MTurk, we require the workers to fulfill the following requirements: (1) to have the history approval rate equal to or more than 98%; (2) to have the history approval instances equal to or more than 5,000; and (3) to live in the United States. However, we find that the results are not reliable enough—its inter-annotator agreement score is very low with a Krippendorff's α of 0.139, which have also been observed by recent studies reporting debatable quality of annotations by MTurk workers for subjective tasks (Clinciu et al., 2021). Thus, we opted to hire trained research assistants for the annotation.

We hired experienced graduate students from the institute as research assistants (RAs). To ensure they possess similar standard rating criteria as we intended, we conducted training sessions and examined them via qualification tests. In the training session, we provided each dataset's NLE examples of low, moderate, and high quality. The RAs were then required to pass the qualification test. The qualification test comprised five NLEs asking for the informativeness and clarity scores separately, requiring them to judge the coarse-grained explanation quality (selecting from low, moderate, and high levels of quality) of given NLEs with their auxiliary information attached. The ground truth labels of those questions were selected from the data annotated by the experienced MTurk workers and then verified by three separate human experts. Only those who provide the correct answers to at least nine out of ten questions are invited for further annotations. If an RA fails this initial assessment, we provide another training session and have the second qualification test. Two of the eight candidate RAs failed the first qualification test. One of those two who did not pass the first test also failed the second test. As a result, we hired seven RAs (four females and three males) as our annotators.

The RAs evaluate the informativeness and clarity of randomly selected NLP instances. Three different annotators assess each of the 300 instances. The process takes 30 working hours, with each annotator careful not to evaluate the same instance twice. They are compensated a total of $450 based on their working hours. During annotation, RAs are instructed to assign scores to each instance within a

Dataset	Document **[Label]**	Auxiliary Information	NLE
e-SNLI (Logical reasoning)	3 young man in hoods standing in the middle of a quiet street facing the camera. **[Entailment]**	Three hood wearing people pose for a picture.	People that are facing a camera are ready to pose for a picture.
LIAR-PLUS (Misinformation justification)	A proposed tax to fund transportation projects would spend $90,000 to take a single vehicle off the road during the morning and afternoon commute. **[False]**	Spoke by Steve Brown during a forum hosted by The Atlanta Journal-Constitution	In sum: Brown based his calculation on a transit cost estimate thats open to accusations of cherry-picking. Furthermore, even Brown and Ross think their own number falls short.
Latent Hatred (Implicit hate speech explanation)	and i will point it out here when u call white people white supremacist just cause they disagree with you it s like me calling you the n word with no proof **[Implicitly hateful]**	-	The post is implicitly hateful because it implies that Blacks blame whites.

Table 1: The data instance examples of NLE and auxiliary for the three datasets.

reasonable time frame. We recommend allocating approximately 60 seconds to rating each instance. However, if they encounter instances that are long or challenging to assess, they are encouraged to take additional time to rate.

The inter-coder reliability scores for the e-SNLI, LIAR-PLUS, and Latent Hatred datasets are 0.7416, 0.6736, and 0.6755, respectively. As expected, the e-SNLI dataset, which focuses on logical reasoning statements, achieves the highest agreement score among raters. Conversely, the Latent Hatred dataset receives a lower score, suggesting that raters' differing backgrounds may lead to varied interpretations of the same text. Surprisingly, the LIAR-PLUS dataset shows an even lower agreement score, indicating that discerning the justifications linked to claims is a more challenging task than expected. This analysis demonstrates that the datasets selected for evaluation require different levels of understanding and reasoning skills, as evidenced by the inter-coder reliability scores.

Table 2 summarizes the inter-annotator agreement scores, elapsed time, the average annotation speed per data instance, and the total expenses. The inter-annotator agreement was high with an α of 0.721, considered acceptable agreement (Clinciu et al., 2021). We use these annotations as ground truth to compare with ChatGPT's annotations.

3.2.2. ChatGPT Annotation

We query the ChatGPT (gpt-3.5-turbo released on May 24, 2023) (OpenAI, 2023b) to assess the informativeness and clarity of each NLE for the 300 sampled instances. For each instance, we query three times and take the average scores using its official API provided by OpenAI (OpenAI, 2023a). The prompt has two components: task description and instructions. The task description is *Task: Scoring the quality of the given natural language explanation based on its referring tweet from the perspective of informativeness and clarity.* Then, the instructions are in six parts: (1) dataset description, (2) metric description (what informativeness and clarity are), (3) high score examples, (4) low score examples, (5) input (e.g., tweet, annotated label, and explanation), and (6) instruction (asking for the two ratings). Each generation is from a distinct session to rule out the effects of chat context. The ChatGPT annotation took two hours and cost $2 in total. While coming from the same model, the answers are relatively consistent with the α of 0.743.

4. RQ1: Does ChatGPT's evaluation of the quality of NLEs align with human assessments?

Withholding the annotated data at hand, in this section, we attempt to evaluate how accurately ChatGPT can assess the explanation quality of NLEs. In doing so, we first analyze the correlations between the two evaluation results. Then, we formulate it as a classification task: using the human (i.e., trained RAs) evaluations of the quality of explanations as ground truth data, we use ChatGPT evaluations as

Annotator Source	Krippendorff's α	Time Consumption	Annotation Speed per Instance	Expenses
Trained RAs	0.721	30 hours	120 seconds	$450
ChatGPT	0.743	2 hours	3.5 seconds	$2

Table 2: Characteristics of annotations by trained RAs and ChatGPT.

predictions and evaluate its performance in classifying the level of the quality of explanations.

4.1. Correlation Analysis

Using correlation analysis to explain the relationship between ChatGPT-generated ratings and ground-truth human ratings is one practical protocol to present the capability of the language model in providing subjective ratings of NLEs. By statistically analyzing the correlation between ChatGPT-generated ratings and RAs annotations, we can delineate the scenarios where the model acts perfectly and cases where it might require further tuning and adjustments.

Figure 1 visualizes the correlations between human and ChatGPT ratings where the blue dashed line indicates a $y = x$ relationship. We can observe strong positive correlations between the evaluations for informativeness and clarity across all three datasets. Pearson's correlation coefficients (Spearman's ranking correlation coefficients) range from 0.636 (0.502) to 0.888 (0.849). In most cases, ChatGPT tends to underestimate the quality, especially for the informativeness of all three datasets, except for the clarity on logical reasoning and the informativeness scores on implicit hate speech explanation. We also noticed that the clarity ratings show weaker correlations than the informativeness ratings, which might be due to the inherent ambiguity of the definition of clarity. ChatGPT ratings correlate well with human ratings for logical reasoning and misinformation justification datasets, while the correlation for the implicit hate speech explanation is weaker. This is unsurprising, as the implicit hate speech explanation dataset's explanations tend to be shorter than the other two datasets. Consequently, a significant portion of their scores falls within the range of 4 to 7, making it challenging to distinguish them.

4.2. Coarse- and Fine-Grained Assessment via Classification

To quantify the capability of ChatGPT in evaluating the quality of NLEs, we formulate the classification tasks on three granularities, from the binary classification of distinguishing low/high to the ternary classification of low/moderate/high, and the raw scores (1 to 7). We employ the weighted F1-score and root mean square error (RMSE) as evaluation metrics in presenting our results. While the f1-score is a common measure for classification problems, they may overlook errors arising from variations in category orders. RMSE, on the other hand, accounts for these order differences.

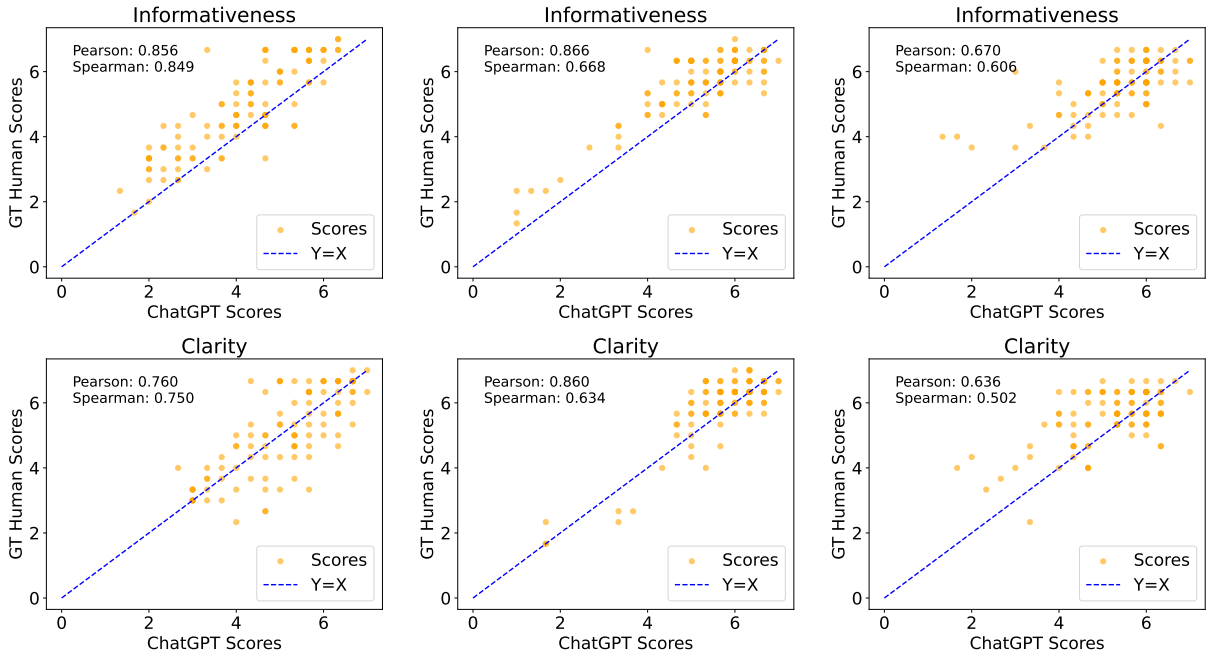
Dataset	Metric	F1-score↑	RMSE↓
e-SNLI	Info.	0.88 (**+0.28**)	0.36 (-0.17)
	Clar.	0.85 ($+0.18$)	0.37 (-0.11)
LIAR -PLUS	Info.	0.97 ($+0.09$)	0.17 (-0.11)
	Clar.	**1.00** ($+0.09$)	**0.00** (**-0.24**)
Latent Hatred	Info.	0.97 ($+0.01$)	0.20 ($+0.03$)
	Clar.	0.95 (-0.01)	0.24 ($+0.08$)

Table 3: Binary (low and high) classification results for the two metrics across the three datasets. We report the weighted f1-score and RMSE (Root Mean Square Error). The numbers in parentheses show the improvement from the baseline.

Dataset	Metric	F1-score↑	RMSE↓
e-SNLI	Info.	0.64 ($+0.34$)	0.62 (**-0.24**)
	Clar.	0.77 (**+0.34**)	0.47 (**-0.24**)
LIAR -PLUS	Info.	0.84 ($+0.07$)	0.44 (-0.14)
	Clar.	**0.90** ($+0.06$)	**0.33** (-0.21)
Latent Hatred	Info.	0.87 ($+0.13$)	0.37 (-0.05)
	Clar.	0.81 ($+0.10$)	0.47 (-0.01)

Table 4: Ternary (low, moderate, and high) classification results.

We begin with coarse granularity. We test whether ChatGPT can assess the explanation quality of NLEs in a *binary* way, i.e., high and low. We categorize informativeness and clarity scores falling between 1 and 4 as low quality and scores between 4 and 7, including 4, as high quality. Table 3 presents the binary classification result and the performance changes compared to the baseline models. We consider the majority voting our baseline model (i.e., label all the instances using the most prominent label). We expect to observe high f1-scores and low RMSE in binary classifications as the data is imbalanced (most data points are gathered in the 'high' class). The ChatGPT performs very well, especially with 100% correctness in mea-



(a) Logical Reasoning (b) Misinformation Justification (c) Implicit Hate Speech Explanation

Figure 1: Correlation between ChatGPT and human evaluations for the three datasets and the two metrics, Informativeness and Clarity. Pearson's and Spearman's correlation coefficients are in the figure.

Dataset	Metric	F1-score↑	RMSE↓
e-SNLI	Info.	0.36 (+0.26)	0.99 (−0.40)
	Clar.	0.51 (**+0.35**)	0.89 (**-0.92**)
LIAR -PLUS	Info.	0.53 (+0.21)	0.79 (−0.62)
	Clar.	**0.58** (+0.08)	**0.66** (−0.60)
Latent Hatred	Info.	0.58 (+0.30)	0.86 (+0.06)
	Clar.	0.46 (+0.23)	0.95 (+0.09)

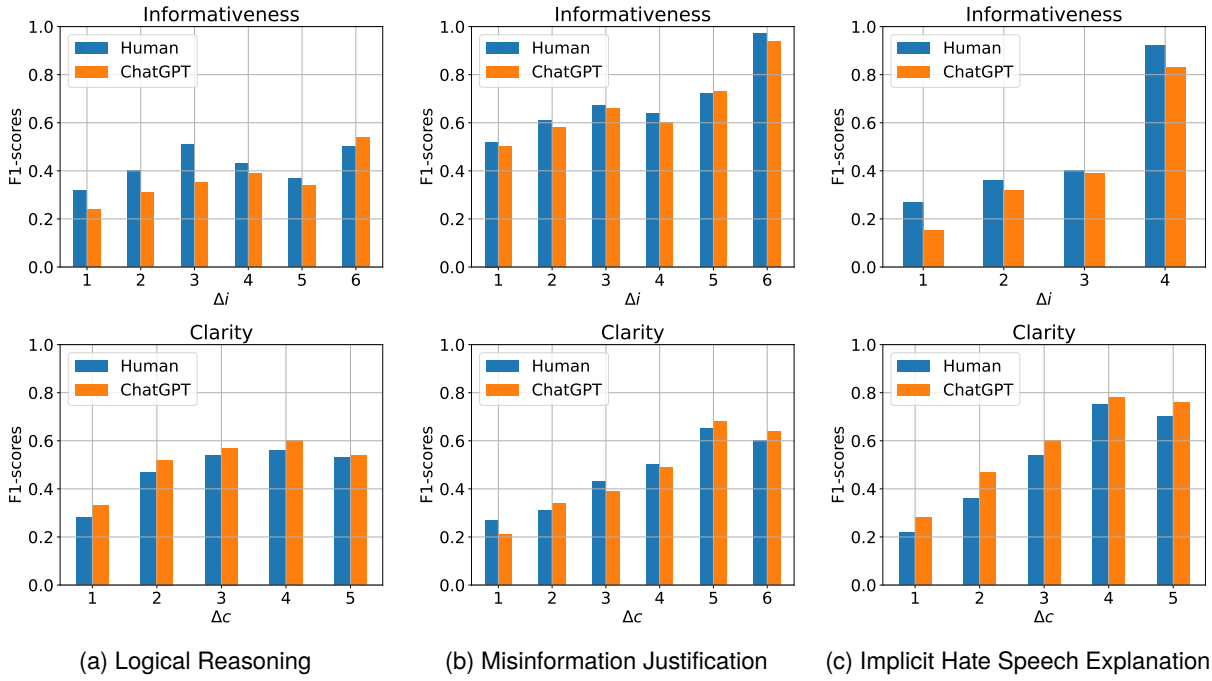
Table 5: 7-way (integer bins from 1 to 7) classification results.

suring clarity on a binary scale for the misinformation justification dataset. ChatGPT performs best for misinformation justification and worst for logical reasoning. As for the performance improvements, the biggest is for the logical reasoning dataset, and the least is for the implicit hate speech explanation dataset.

Next, we explore ChatGPT's performance in more detail. To address this, we establish a ternary classification task, with the classes defined as follows: low (ranging from 1 to 3), moderate (spanning from 3 to 5), and high (encompassing the range from 5 to 7). Table 4 shows the results of this ternary classification. Similarly, the ChatGPT performs the best for the misinformation justification dataset and the worst on the logical reasoning dataset to provide quality rating scores. Compared with the baseline, the ChatGPT provides the biggest performance improvements for the logical reasoning dataset while the least for the implicit hate speech explanation dataset.

We also use the raw informativeness and clarity score for the finest granularity of score estimation as in Table 5. We can observe similar patterns where the ChatGPT performs the best for the misinformation justification dataset. The ChatGPT performs the best for the clarity rating scores (0.58 f1-score and 0.66 RMSE value) while the worst for informativeness rating scores (0.36 f1-score and 0.99 RMSE value). As for performance improvement, ChatGPT performs best regarding clarity rating scores (+0.35 f1-score and -0.92 RMSE value). The performance disparities may arise from LLMs having different domain knowledge, as observed in existing works (Rytting and Wingate, 2021; OpenAI, 2023d; Yang et al., 2023; Huang et al., 2024).

In summary, among all three granularities, we observe a strong alignment between ChatGPT evaluation and human assessment and similar patterns of the best performance for misinformation justification rating and the worst for logical reasoning rating. However, as the granularity increases, the best f1-score drops from 1.0 to 0.58 (clarity for misinformation justification dataset), which means that the ChatGPT only performs well in the coarse-grained rating.



(a) Logical Reasoning (b) Misinformation Justification (c) Implicit Hate Speech Explanation

Figure 2: Visualization of ChatGPT pair comparison f1-scores for various Δ_i and Δ_c, from 0 to 6 rounded by integer bins, compared with additional human annotations specifically for the pair comparison task.

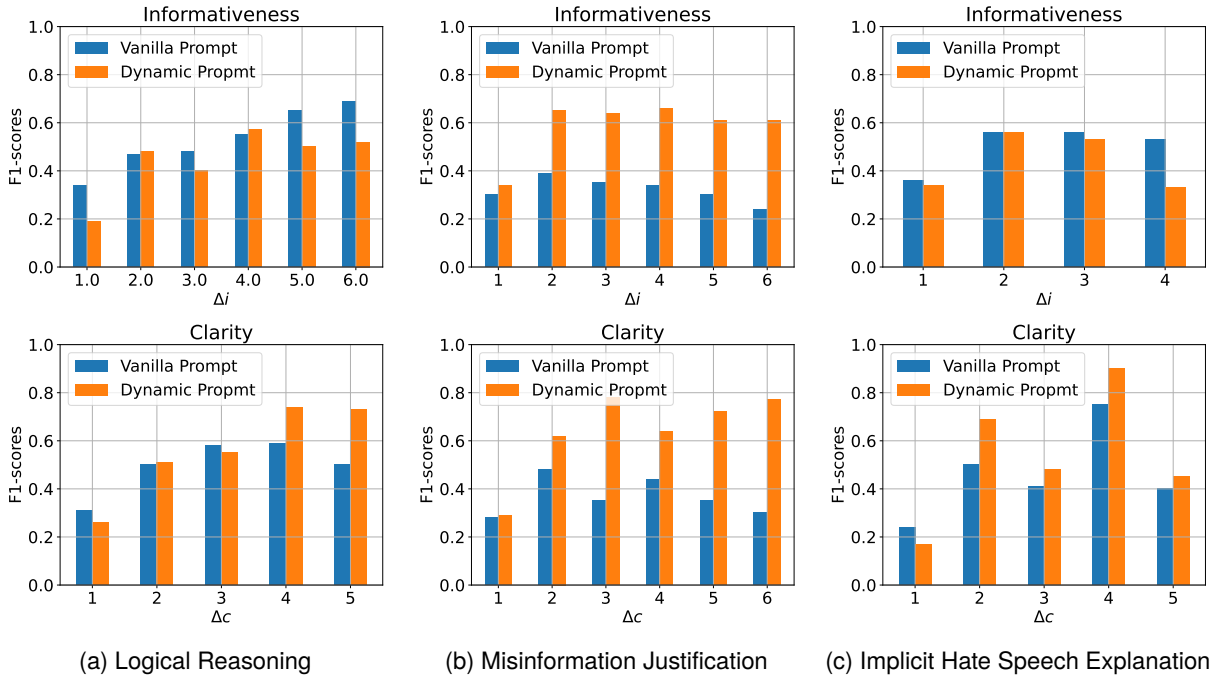
5. RQ2: Is ChatGPT capable of comparing two NLEs in terms of their explanation quality?

We evaluate ChatGPT's capability via pairwise comparison. Comparing instances is easier and more accurate than giving a score to an instance for human annotators (Herbrich et al., 2006). Our results in §4 also show that fine-grained quality assessment is challenging. If ChatGPT excels in performing pairwise comparisons close to human judgment, it becomes feasible to obtain precise estimations for all instances using pairwise comparison results (Herbrich et al., 2006).

In this experiment, rather than requesting ratings for explanation quality, we present pairs of two explanations and ask ChatGPT and human annotators which one is perceived as more informative or clear. Those pairs with small score differences are more challenging to discern than those with larger ones. We analyze the performance by varying score differences Δ. We denote the score differences of the ground-truth dataset as Δ_i for informativeness and Δ_c for clarity. Then, for each of Δ_i and Δ_c, we randomly select 100 instances out of all possible pairs. We use all pairs for cases where it does not have 100 pairs (e.g., implicit hate speech explanation dataset only has 41 pairs at most when $\Delta_i = 4$). In total, we use 520 (520), 550 (546), and 321 (372) pairs for informativeness (clarify) for logical reasoning, misinformation justifications, and implicit hate speech explanation datasets, respectively. We collected 8,346 annotations from both human annotators and ChatGPT. The same trained RAs provided the annotations and were compensated a total of $650 based on their working hours for this task. We classify the outcomes of the pairwise comparisons into three categories: 1) the first instance has a higher score, 2) the first instance has an equal score to the second, and 3) the first instance has a lower score.

Figure 2 shows how accurate the pairwise comparison is by varying Δ_i and Δ_c. As Δ_i and Δ_c increase, ChatGPT and human experts generally perform better to distinguish the difference from NLE pairs. For the sixteen Δ_i cases, human outperforms ChatGPT by 4.63%p of f1-score; for the sixteen Δ_c cases, ChatGPT outperforms human by 3.06%p f1-score. We observed a similar pattern for all three datasets: ChatGPT performs better than human experts in clarity assessment and worse in informativeness assessment. Still, for the misinformation justification dataset, the clarity assessment performs poorly when the Δ_c is one, three, or four. Human experts and ChatGPT can distinguish well for the implicit hate speech explanation when the Δ_i and Δ_c increase. Specifically for the informativeness ratings of logical reasoning and implicit hate speech explanation datasets, when Δ_i is small, ChatGPT performs poorly, and even when Δ_i becomes 3, the overall performance is still less than 0.40 f1-score.

In summary, ChatGPT performs worse than human annotations for most informativeness com-



(a) Logical Reasoning (b) Misinformation Justification (c) Implicit Hate Speech Explanation

Figure 3: Visualization of ChatGPT pair comparison f1-scores for various Δ_i and Δ_c, from 0 to 6 rounded by integer bins, showing the difference between vanilla prompting and our proposed dynamic prompting. The measurement is only based on NLE evaluation metrics of Informativeness and Clarity.

parison cases (average difference of 4.63%p) but mostly slightly better for clarity comparison cases (average difference of 3.06%p). ChatGPT performs the best for explanation quality comparison of the misinformation justification dataset among three datasets. Still, it performs significantly worse than human annotators when distinguishing a small difference in rating scores (i.e., informativeness rating of logical reasoning and implicit hate speech explanation datasets). Considering the average absolute f1-score difference among the three datasets, ChatGPT performs better in misinformation justification (2.92%p) than logical reasoning (5.64%p) and implicit hate speech explanation (5.80%p) datasets. In the next section, we will explore how to improve ChatGPT's performance on quality evaluation further.

6. RQ3: Can dynamic prompting enhance ChatGPT's ability to assess NLE quality?

As in-context learning helps LLMs to understand a task (Khalifa et al., 2022) better, we explore dynamic prompting for fine-grained assessment and pair-comparison tasks. The idea is to provide a prompt with the most semantically similar example instead of fixed examples, along with its informativeness and clarity scores. To identify the most similar example, we first split our data into a candidate example set and a test set, with the specific division method detailed later as it varies across experiments. Then, we obtain the representations of every explanation using Sentence-BERT (Reimers and Gurevych, 2019). By using a pre-trained model[2], we choose the one with the highest cosine similarity from the candidate example set for each instance in the test set.

Dataset	Metric	F1-score↑	RMSE↓
e-SNLI	Info.	0.25 (-0.09)	1.88 ($+0.45$)
	Clar.	0.19 (0.00)	2.12 ($+0.51$)
LIAR -PLUS	Info.	0.38 (-0.02)	**0.99** ($+0.11$)
	Clar.	**0.46** (**+0.11**)	1.06 (**+0.01**)
Latent Hatred	Info.	0.19 (-0.17)	1.35 ($+0.07$)
	Clar.	0.32 ($+0.05$)	1.91 ($+0.61$)

Table 6: Dynamic prompting results. The granularity used here is 7-way classification.

We evaluate the performance of dynamic prompting by fine-grained classification and pairwise comparison. We use the same ground-truth datasets built in RQ1 and RQ2. We employ a 50:50 data split for each dataset, with one half as the candidate explanation examples and the other half as the test set. In the case of the pairwise comparison, we utilize a 50:50 data split for each Δ within

[2]The paraphrase-distilroberta-base-v1 in sentence-transformers: https://github.com/UKPLab/sentence-transformers.

each dataset, each containing a maximum of 100 pairs. We compare the performance of dynamic prompting with that of the prompting used in §4 and §5. We note that the baseline results differ from those presented in Figure 2 as we use 50% of the data for testing in this analysis.

As shown in Table 6, the dynamic prompting does not improve classification performance but worsens it. Dynamic prompting works relatively equivalent to the vanilla prompt design only for the misinformation justification dataset. Generally, the clarity metric performance becomes slightly better with in-context learning, and the informativeness metric performance significantly worsens. We could not find any evidence that dynamic prompting improves the performance of explanation quality ratings.

However, in the pairwise comparison, dynamic prompting shows sizable improvement for the misinformation justification dataset (Figure 3). On average, it provides 5.13%p and 14.50%p improvements of informativeness and clarity across the different Δ among three datasets. Dynamic prompting makes it even worse for the case that showed the worst performance in the previous section: informativeness for logical reasoning and implicit hate speech explanation datasets.

In summary, dynamic prompting can further improve the performance of ChatGPT to estimate the explanation quality for the misinformation justification dataset in a comparison setting. However, the performance for the other two datasets does not always increase through dynamic prompting under different Δ_i and Δ_c. In a pairwise comparison setting, we notice a substantial performance improvement attributed to dynamic prompting. However, its effect is adverse in raw score classification. This result implies that ChatGPT can better leverage the additional contextual information for comparison than directly estimate the score, especially for the misinformation justification dataset.

7. Limitations

Our study sheds light on the alignment between ChatGPT and human assessments in different levels of logical reasoning and implication understanding, evaluated at the subjective quality assessment of natural language explanations. However, several limitations remain for further discussion and consideration.

This work primarily evaluates ChatGPT (i.e., gpt-3.5-turbo model), which is a closed-source language model. This raises concerns about the generalizability of our findings to other LLMs, particularly open-source models. Future research should explore a wider range of LLMs to understand their capabilities in assessing NLE quality and to determine the extent to which our observations regarding ChatGPT are generalizable.

We observe ChatGPT's remarkable capability in coarse-grained categorization tasks, particularly within misinformation justification datasets. Our investigation uses a simple and replicable prompt design to facilitate research reproducibility. However, it is important to acknowledge that different prompt designs may influence ChatGPT's performance. Also, for the dynamic prompting, we only add the two examples in the prompt without altering any other information. Further investigation into how changes in instructions and dynamic examples within the prompt influence ChatGPT's responses would be beneficial. This exploration will help understand the impact of different prompting designs on evaluating NLE quality, focusing on the alignment between human and machine intelligence

Though robust, the metrics used in our study, Informativeness (How helpful the NLE is while understanding the context with the given labels) and Clarity (How clearly the ideas in the NLE are expressed), might not fully capture the multifaceted nature of human language and explanation quality. Future research could incorporate diverse text quality or helpfulness evaluation metrics tailored for different context-understanding tasks, enhancing our understanding of human-machine intelligence alignments.

8. Conclusion

ChatGPT exhibited remarkable capability in coarse-grained categorization tasks, particularly within misinformation justification datasets, while simultaneously revealing potential pitfalls, especially in finer granularity assessments in various settings. While robust, the metrics we applied in this study may fail to encapsulate the multilayered complexity and subtle nuances that characterize human language. Future research could incorporate more evaluation metrics tailored to be more diverse and sensitive to the contextual variations of language.

Ethics Statement

The Institutional Review Board at Indiana University has approved the experiment design of collecting human annotations (Approval number of #18813). We provide mental consultant hotline and clinic information to all annotators in case they need it. We encouraged the human annotators to stop or quit the annotation process anytime if they felt uncomfortable. We paid the annotators based on their completed instance (for MTurk workers) or effective working hours (for hired experienced research assistants), at the amount above the average salary requirement in the US.

LLMs may reflect Western readers' views more than other regions rooted in the source of the pre-trained corpus of LLMs (Durmus et al., 2023), which could lead to inherent misunderstanding or bias towards cultural-related contexts. More future work is needed to mitigate those potential misunderstandings and biases.

Acknowledgements

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2024-RS-2022-00156360).

References

Sayed Fayaz Ahmad, Heesup Han, Muhammad Mansoor Alam, Mohd Rehmat, Muhammad Irshad, Marcelo Arraño-Muñoz, Antonio Ariza-Montes, et al. 2023. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, 10(1):1–14.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 3665–3674, New York, NY, USA. Association for Computing Machinery.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T. Wolf, Evelyn Duesterwald, Casey Dugan, Werner Geyer, and Darrell Reimer. 2021. AI-assisted human labeling: Batching for efficiency without overreliance. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

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.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1):103111.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI:

Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 175–182, New York, NY, USA. Association for Computing Machinery.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 99–108, Tilburg University, The Netherlands. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Miruna Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. *arXiv preprint arXiv:2103.08545*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Simon Hviid Del Pin and Seyed Ali Amirshahi. 2022. Subjective quality evaluation: What can be learnt from cognitive science?

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2019. Crowdsourcing interface feature design with Bayesian optimization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

A.E. Elo. 1966. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Douglas R. Hofstadter and Fluid Analogies Research Group. 1995. *Fluid Concepts And Creative Analogies*.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023a. Chain of Explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 90–93, New York, NY, USA. Association for Computing Machinery.

Fan Huang, Haewoon Kwak, and Jisun An. 2023b. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Fan Huang, Haewoon Kwak, and Jisun An. 2024. Token-Ensemble text generation: On attacking the automatic AI-generated text detection. *arXiv preprint arXiv:2402.11167*.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. E-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1244–1254.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2022. LEPUS: Prompt-based unsupervised multi-hop reranking for open-domain QA. *arXiv preprint arXiv:2205.12650*.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Few-shot reranking for multi-hop QA via language model prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15882–15897, Toronto, Canada. Association for Computational Linguistics.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 453–456, New York, NY, USA. Association for Computing Machinery.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Haewoon Kwak, Jisun An, and Kunwoo Park. 2022. Who is missing? characterizing the participation of different demographic groups in a korean nationwide daily conversation corpus. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and

Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. 2022. On the probability–quality paradox in language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–45, Dublin, Ireland. Association for Computational Linguistics.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G Smirnova, and Ted Hsuan Yun Chen. 2023. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT. OpenAI Blog.

OpenAI. 2023a. API reference - OpenAI. OpenAI Offical Website.

OpenAI. 2023b. ChatGPT — release notes. OpenAI Offical Website.

OpenAI. 2023c. GPT-4 technical report. ArXiv.

OpenAI. 2023d. GPT-4V(ision) system card. OpenAI Blog.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Christopher Rytting and David Wingate. 2021. Leveraging the inductive bias of large language models for abstract textual reasoning. In *Advances in Neural Information Processing Systems*, volume 34, pages 17111–17122. Curran Associates, Inc.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Tatsuya Sakai, Kazuki Miyazawa, Takato Horii, and Takayuki Nagai. 2021. A framework of explanation generation toward reliable autonomous robots. *Advanced Robotics*, 35(17):1054–1067.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Michael Saxon, Sharon Levy, Xinyi Wang, Alon Albalak, and William Yang Wang. 2021. Modeling disclosive transparency in NLP application

descriptions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2037, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "This is a problem, Don't you agree?" Framing and Bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Minkyu Shin, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths. 2023. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023a. Evaluating GPT-3 generated explanations for hateful content moderation. *arXiv preprint arXiv:2305.17680*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023b. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Stephen Wolfram. 2023. Wolfram|Alpha as the way to bring computational knowledge superpowers to ChatGPT. Stephen Wolfram Writings.

Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy, and Noah A. Smith. 2023. Transparency helps reveal when language models learn meaning. *Transactions of the Association for Computational Linguistics*, 11:617–634.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LLMs: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*.

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. 2023a. Are human explanations always helpful? towards objective evaluation of human natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14698–14713, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of Thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shao Yi Liaw, Fan Huang, Fabricio Benevenuto, Haewoon Kwak, and Jisun An. 2023. YouNICon: YouTube's CommuNIty of Conspiracy videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1102–1111.

Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2021. Few-shot out-of-domain transfer learning of natural language explanations. *arXiv preprint arXiv:2112.06204*.

Ping Yu, Tianlu Wang, Olga Golovneva, Badr AlKhamissi, Siddharth Verma, Zhijing Jin, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. 2023.

ALERT: Adapt language models to reasoning tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1055–1081, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

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*.

A. ChatGPT Pair Comparison Additional Details

Here, we provide the example for the prompt we will pass to ChatGPT when comparing the informativeness scores of logical reasoning NLE pairs in Table 13. The information we give to expert human annotators is the same as the above prompts. We collect three rounds of classification annotations for the logical reasoning pair comparison for informativeness and clarity. In total, the annotation number is 3300. For the misinformation justification, the annotation number here is 3120. For the implicit hatefulness explanation, the annotation number is 1926. We list the detailed pair comparison scores in Table 7, Table 8, and Table 9.

Δ_i	Human	ChatGPT
1	0.32	0.24
2	0.40	0.31
3	0.51	0.35
4	0.43	0.39
5	0.37	0.34
6	0.50	0.54
Δ_c	Human	ChatGPT
1	0.28	0.33
2	0.47	0.52
3	0.54	0.57
4	0.56	0.60
5	0.53	0.54

Table 7: Detailed pair comparison F1-scores of the E-SNLI dataset, for logical reasoning.

Δ_i	Human	ChatGPT
1	0.52	0.50
2	0.61	0.58
3	0.67	0.66
4	0.64	0.60
5	0.72	0.73
6	0.97	0.94
Δ_c	Human	ChatGPT
1	0.27	0.23
2	0.31	0.34
3	0.43	0.39
4	0.50	0.49
5	0.65	0.68
6	0.60	0.64

Table 8: Detailed pair comparison F1-scores of the LIAR-PLUS dataset, for misinformation justification.

Δ_i	Human	ChatGPT
1	0.27	0.15
2	0.36	0.32
3	0.40	0.39
4	0.92	0.83
Δ_c	Human	ChatGPT
1	0.22	0.28
2	0.36	0.47
3	0.54	0.60
4	0.75	0.78
5	0.70	0.76

Table 9: Detailed pair comparison F1-scores of the Latent Hatred dataset, for hate speech explanation.

B. Human Annotation Details

We applied the MTurk design similar to the work of ElSherief et al. (2021) and Huang et al. (2023b). To ensure the good quality of the collected human annotation scores, we only hire the AMT Masters who fit the qualifications of: (1) have an approval rate greater than 98%, (2) have more than 5000 HITs approved, (3) located in the United States. The screenshots of our designed interface used for the Amazon Mechanical Turk platform are shown in Figure 4, 5, 6, and 7.

C. Prompts for ChatGPT Rating Scores

We showcased the prompt examples fulfilled by the dataset instance example listed in Table 1 in the below three tables (Table 10 for logical reasoning, Table 11 for misinformation justification, and Table 12 for implicit hatefulness explanation).

D. Dynamic Prompting

In the pair-comparison task, we randomly selected 50 instances from each dataset to form the database to provide candidate dynamic examples for the remaining 50 testing data instances. For each testing data instance, we applied the Sentence-BERT to create the embedding of itself and all instances in the database. Then we select the one instance from the database with the highest cosine similarity score with the testing instance, which will be used to formulate the example in the prompt fed into the ChatGPT.

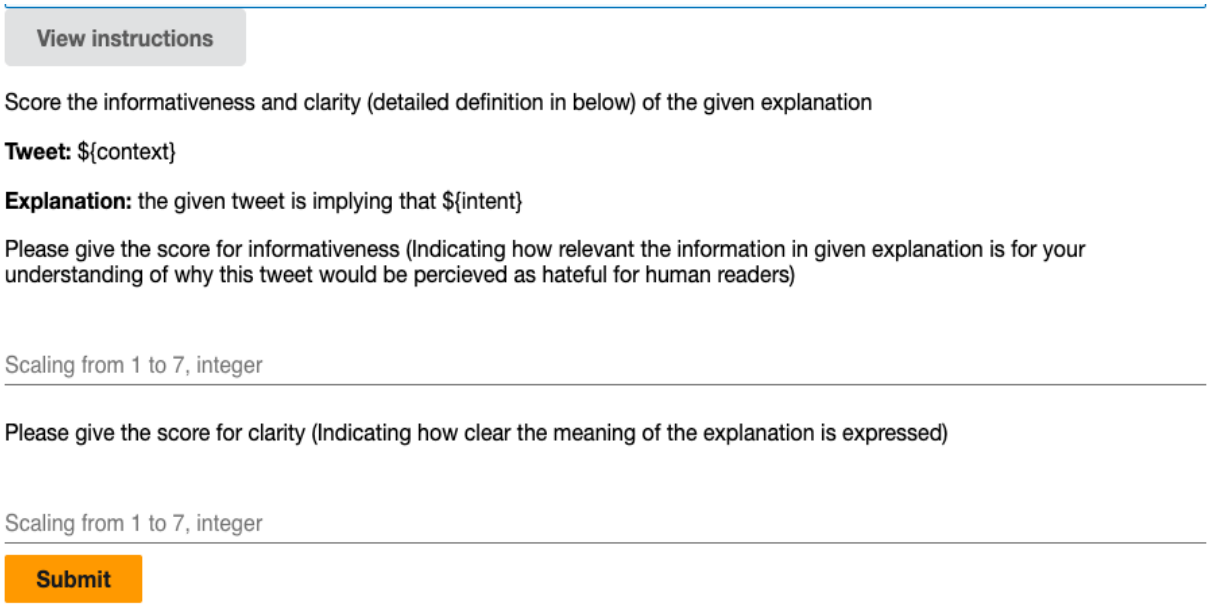


Figure 4: The screenshot of our design user interface used to collect the human evaluation scores of informativeness and clarity.

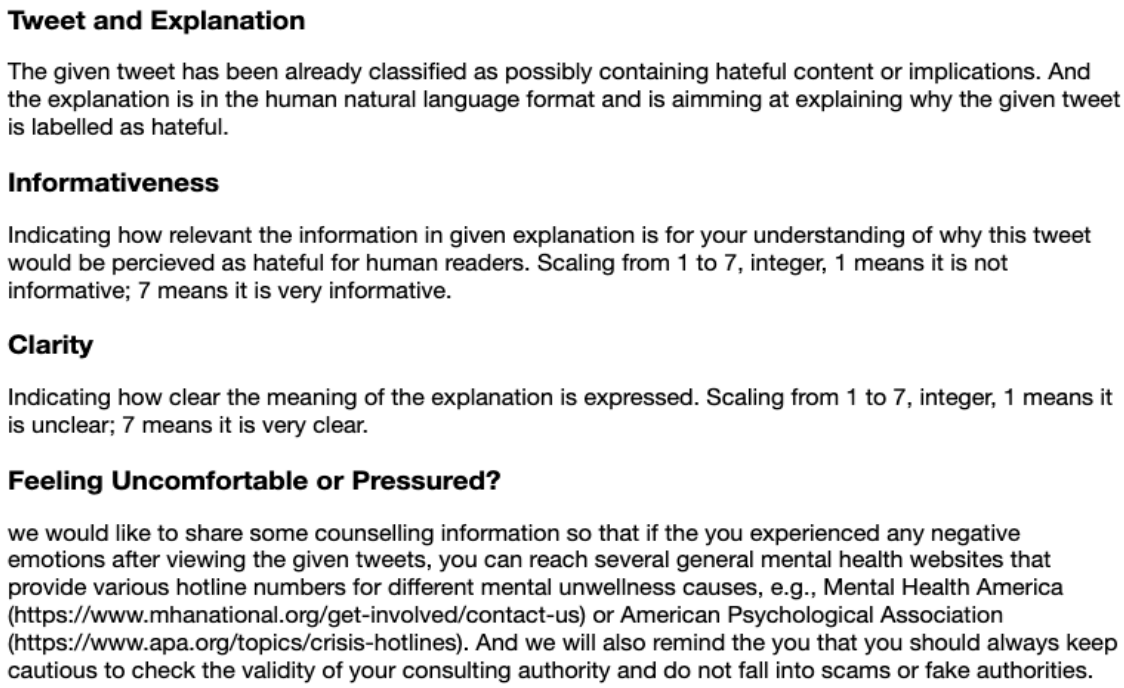


Figure 5: The screenshot of the detailed instruction information in MTurk.

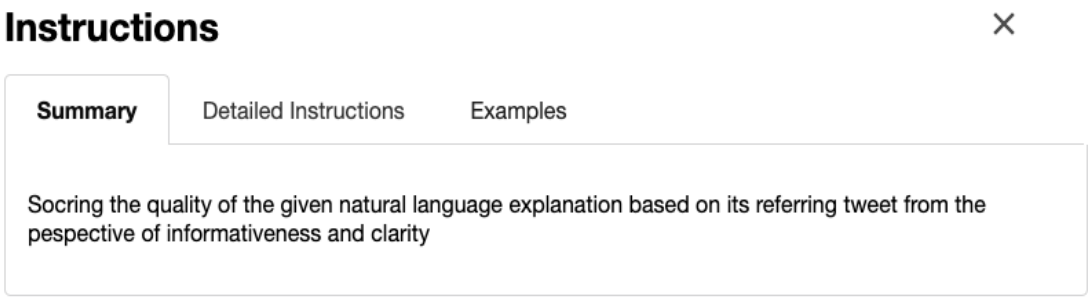


Figure 6: The screenshot of the summary of instruction information in MTurk.

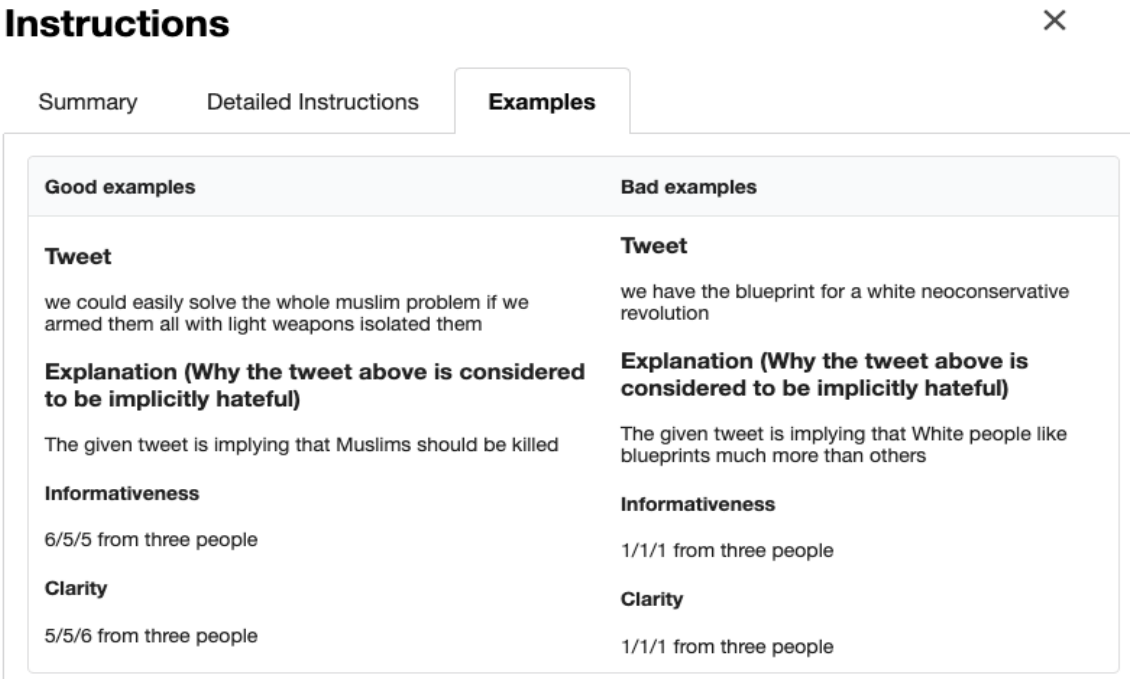


Figure 7: The screenshot of the detailed example information in MTurk.

Task: Scoring the quality of the given natural language explanation based on its referring short text from the perspective of informativeness and clarity. Instructions: [Short Texts and Explanation] The given short texts fit in entrailment/neutral/contridicting relation between Premiss and Hypothesis, related only to commonsense. And the explanation is in the human natural language format and is aimming at explaining the relationship between given Premiss and Hypothesis. [Informativeness] Indicating how relevant the information in given explanation is for your understanding of explaining the relationship between given Premiss and Hypothesis. Scaling from 1 to 7, integer, 1 means it is not informative; 7 means it is very informative. [Clarity] Indicating how clear the meaning of the explanation is expressed. Scaling from 1 to 7, integer, 1 means it is unclear; 7 means it is very clear. High score example: [Premiss] A female dancer wearing all white dancing while a photographer wearing all white takes her picture. [Hypothesis] Some players are playing cricket in ground. [Relationship] contradiction [Explanation] Cricket is not played by female dancers. [Informativeness] 6 [Clarity] 6. Low score example: [Premise] A man in a black leather jacket and a book in his hand speaks in a classroom. [Hypothesis] A man is teaching science in a classroom. [Relationship] neutral [Explanation] doesnt́ mean itś a science class [Informativeness] 3 [Clarity] 2 Score the informativeness and clarity (detailed definition in below) of the given explanation: [Premise] (**3 young man in hoods standing in the middle of a quiet street facing the camera**) [Hypothesis] (**Three hood wearing people pose for a picture.**) [Relationship] (**Entailment**) [Explanation] (**People that are facing a camera are ready to pose for a picture.**) Please give the score for informativeness (Indicating how relevant the information in given explanation is for your understanding of explaining the relationship between given Premise and Hypothesis) and the score for clarity (Indicating how clear the meaning of the explanation is expressed) scale from 1 to 7, integer. And then explain the reason for your scores in one or two sentences."

Table 10: Prompt example we used for logical reasoning (Camburu et al., 2018) to collect ChatGPT rating scores for informativeness and clarity metrics. We use the dataset example in Table 1 to showcase how the prompt is fulfilled, highlighted in bold.

Task: Scoring the quality of the given natural language explanation based on its referring news information from the perspective of informativeness and clarity.

Instructions:
[News and Explanation] The given news statement with pants-fire/flase/barely-true/half-true/mostly-true/true label, considering the related information of speaker and context. And the explanation is in the human natural language format and is aimming at explaining why the given news is considered as pants-fire/flase/barely-true/half-true/mostly-true/true.
[Informativeness] Indicating how relevant the information in given explanation is for your understanding of why the given news is considered as pants-fire/flase/barely-true/half-true/mostly-true/true. Scaling from 1 to 7, integer, 1 means it is not informative; 7 means it is very informative.
[Clarity] Indicating how clear the meaning of the explanation is expressed. Scaling from 1 to 7, integer, 1 means it is unclear; 7 means it is very clear.

High score example: [Statement] Says the governor is going around the state talking about how we should fund an income tax cut that benefits higher income earners and not lower income earners [Label] false [Speaker] john-burzichelli [Context] an interview with NJToday [Explanation] In reality, the Affordable Care Act calls for a slowed growth of Medicare funding, not a slash to current funds. The savings from this approach will be used to offset Obamacare costs. [Informativeness] 6 [Clarity] 6.
Low score example: [Statement] Says an EPA permit languished under Strickland but his new EPA director got it done in two days. [Label] barely-true [Speaker] john-kasich [Context] a news conference [Explanation] Points of Light has a unique mission carved out by President Bush – mobilizing volunteers around the world. Nunn has used her stewardship of the non-profit to burnish her leadership credentials in the race against businessman Perdue. But her political ad undercuts how Points of Light describes itself on its website – as the largest organization in the world dedicated to volunteer service. Thats substantially different from the worlds largest volunteer organization. [Informativeness] 3 [Clarity] 2.

Score the informativeness and clarity (detailed definition in below) of the given explanation: [News statement] (**A proposed tax to fund transportation projects would spend $90,000 to take a single vehicle off the road during the morning and afternoon commute.**) [Label of correctness] (**False**) [Speaker] (**Steve Brown**) [Context] (a forum hosted by The Atlanta Journal-Constitution) [Explanation] (**In sum: Brown based his calculation on a transit cost estimate thats open to accusations of cherry-picking. Furthermore, even Brown and Ross think their own number falls short.**)
Please give the score for informativeness (Indicating how relevant the information in given explanation is for your understanding of why the given news is considered as the label [label]) and the score for clarity (Indicating how clear the meaning of the explanation is expressed) scale from 1 to 7, integer. And then explain the reason for your scores in one or two sentences.

Table 11: Prompt example we used for misinformation justification (Alhindi et al., 2018) to collect ChatGPT rating scores for informativeness and clarity metrics. We use the dataset example in Table 1 to showcase how the prompt is fulfilled, highlighted in bold.

Task: Scoring the quality of the given natural language explanation based on its referring tweet from the perspective of informativeness and clarity.

Instructions:
[Tweet and Explanation] The given tweet has been already classified as possibly containing hateful content or implications. And the explanation is in the human natural language format and is aiming at explaining why the given tweet is labelled as hateful.
[Informativeness] Indicating how relevant the information in given explanation is for your understanding of why this tweet would be perceived as hateful for human readers. Scaling from 1 to 7, integer, 1 means it is not informative; 7 means it is very informative.
[Clarity] Indicating how clear the meaning of the explanation is expressed. Scaling from 1 to 7, integer, 1 means it is unclear; 7 means it is very clear.

High score example: [Tweet] we could easily solve the whole muslim problem if we armed them all with light weapons isolated them [Explanation (Why the tweet above is considered to be implicitly hateful)] The given tweet is implying that Muslims should be killed [Informativeness] 6 [Clarity] 5.
Low score example: [Tweet] we have the blueprint for a white neoconservative revolution [Explanation (Why the tweet above is considered to be implicitly hateful)] The given tweet is implying that White people like blueprints much more than others [Informativeness] 1 from three people [Clarity] 1 from three people.

Score the informativeness and clarity (detailed definition in below) of the given explanation: [Tweet] (**and i will point it out here when u call white people white supremacist just cause they disagree with you it s like me calling you the n word with no proof**) [Explanation] (**The post is implicitly hateful because it implies that Blacks blame whites.**)
Please give the score for informativeness (Indicating how relevant the information in given explanation is for your understanding of why this tweet would be perceived as hateful for human readers) and the score for clarity (Indicating how clear the meaning of the explanation is expressed) scale from 1 to 7, integer. And then explain the reason for your scores in one or two sentences.

Table 12: Prompt example we used for implicit hate speech explanation (ElSherief et al., 2021) to collect ChatGPT rating scores for informativeness and clarity metrics. We use the dataset example in Table 1 to showcase how the prompt is fulfilled, highlighted in bold.

Task: Comparing the quality of the given natural language explanation based on its referring short text from the perspective of informativeness and clarity. Instructions: [Short Texts and Explanation] The short texts fit in entrailment/neutral/contridicting relation between Premiss and Hypothesis, related only to commonsense. And the explanation is in the human natural language format and is aimming at explaining the relationship between given Premiss and Hypothesis. [Informativeness] Indicating how relevant the information in given explanation is for your understanding of explaining the relationship between given Premiss and Hypothesis. Scaling from 1 to 7, integer, 1 means it is not informative; 7 means it is very informative. [Clarity] Indicating how clear the meaning of the explanation is expressed. Scaling from 1 to 7, integer, 1 means it is unclear; 7 means it is very clear. High score example: [Premiss] A female dancer wearing all white dancing while a photographer wearing all white takes her picture. [Hypothesis] Some players are playing cricket in ground. [Relationship] contradiction [Explanation] Cricket is not played by female dancers. [Informativeness] 6 [Clarity] 6. Low score example: [Premise] A man in a black leather jacket and a book in his hand speaks in a classroom. [Hypothesis] A man is teaching science in a classroom. [Relationship] neutral [Explanation] doesnt́ mean itś a science class [Informativeness] 3 [Clarity] 2. Compare the informativeness or clarity (detailed definition in above) of the given explanation The First: [Premise] ($premise_1$) [Hypothesis] ($hypothesis_1$) [Relationship] ($relationship_1$) [Explanation] ($explanation_1$) The Second: [Premise] ($premise_2$) [Hypothesis] ($hypothesis_2$) [Relationship] ($relationship_2$) [Explanation] ($explanation_2$) Please answer Yes or No or Neutral for whether the first instance is more informative (Indicating how relevant the information in the given explanation is for your understanding of why this tweet would be perceived as hateful for human readers) than the second instance. And then explain the reason for your scores in one or two sentences.

Table 13: Prompt example we used to collect ChatGPT answers when comparing the informativeness scores for logical reasoning NLE pairs (Camburu et al., 2018).