

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



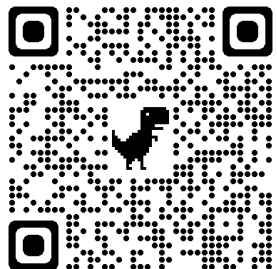
Fan Huang<sup>1</sup>, Haewoon Kwak<sup>1</sup>, Kunwoo Park<sup>2</sup>, Jisun An<sup>1</sup>

<sup>1</sup>Indiana University Bloomington, USA

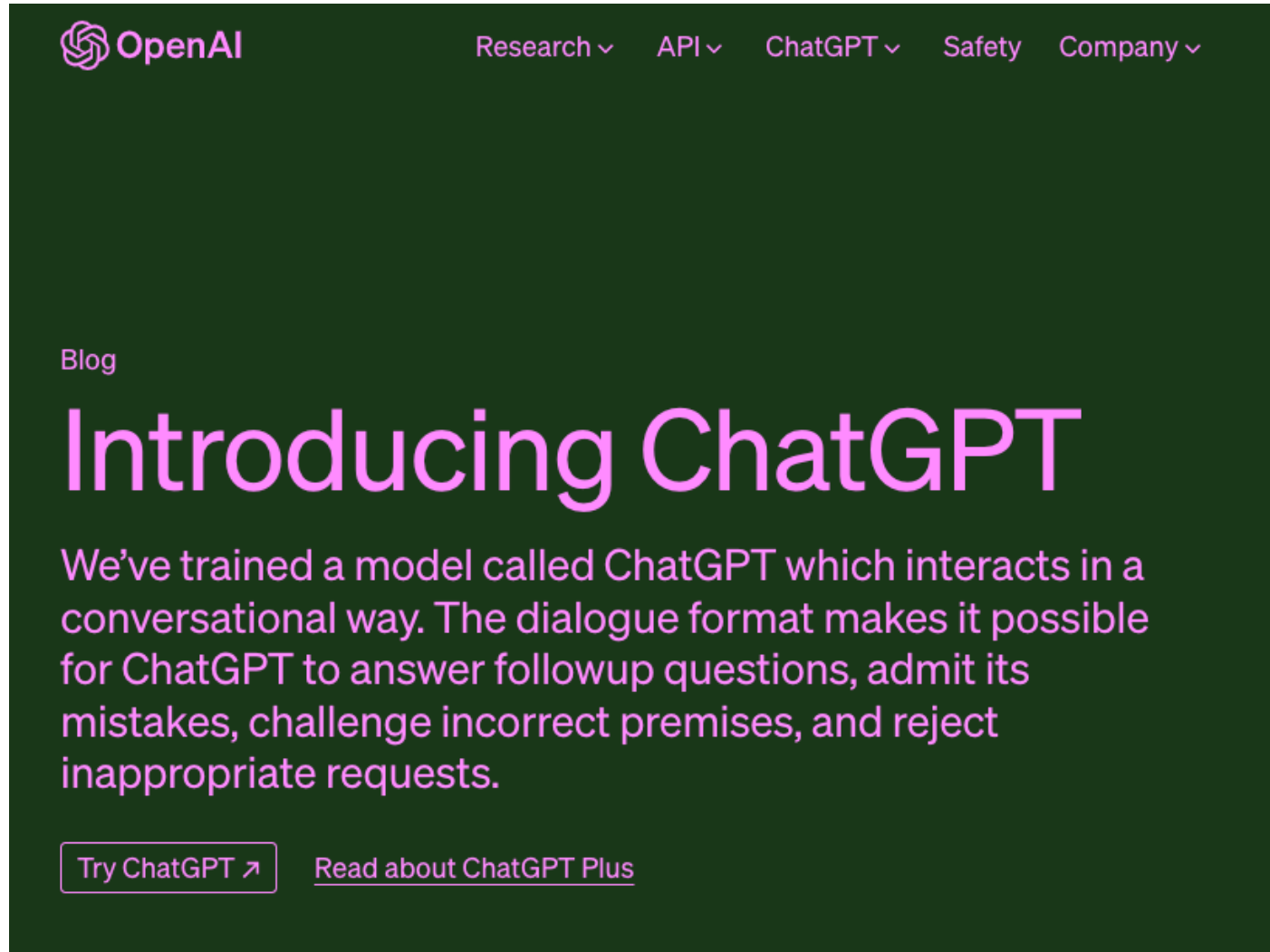
<sup>2</sup>Soongsil University, South Korea



Presenter: Fan Huang  
[fanhuan@iu.edu](mailto:fanhuan@iu.edu)



# Era of Large Language Models



The image is a screenshot of the OpenAI website's blog page. At the top left is the OpenAI logo, and to its right are navigation links for Research, API, ChatGPT, Safety, and Company. Below the navigation is a 'Blog' section header. The main heading is 'Introducing ChatGPT' in a large, bold, black font. Below the heading is a paragraph of text describing the model's conversational capabilities. At the bottom of the content area are two buttons: 'Try ChatGPT' and 'Read about ChatGPT Plus'.

OpenAI Research API ChatGPT Safety Company

Blog

## Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

[Try ChatGPT](#) [Read about ChatGPT Plus](#)

# Powerful Large Language Models are changing everything

Already outperform all previous NLP models in basically all NLP tasks, sometimes reaching human performance.



**What does this sign mean?**

- There is an underpass ahead
- There is a winding road ahead
- There is a narrow bridge ahead
- The road ahead will merge

Let's think about it together. When designed to communicate specific information in conditions ahead, right? If you choose "ahead," consider what type of information to convey. Is there another option that conveys the conditions or actions drivers should watch out for based on the standard road sign? Reflecting on the purpose of different signs, clarify why this particular answer might be described. What does each sign type indicate you should watch out for?

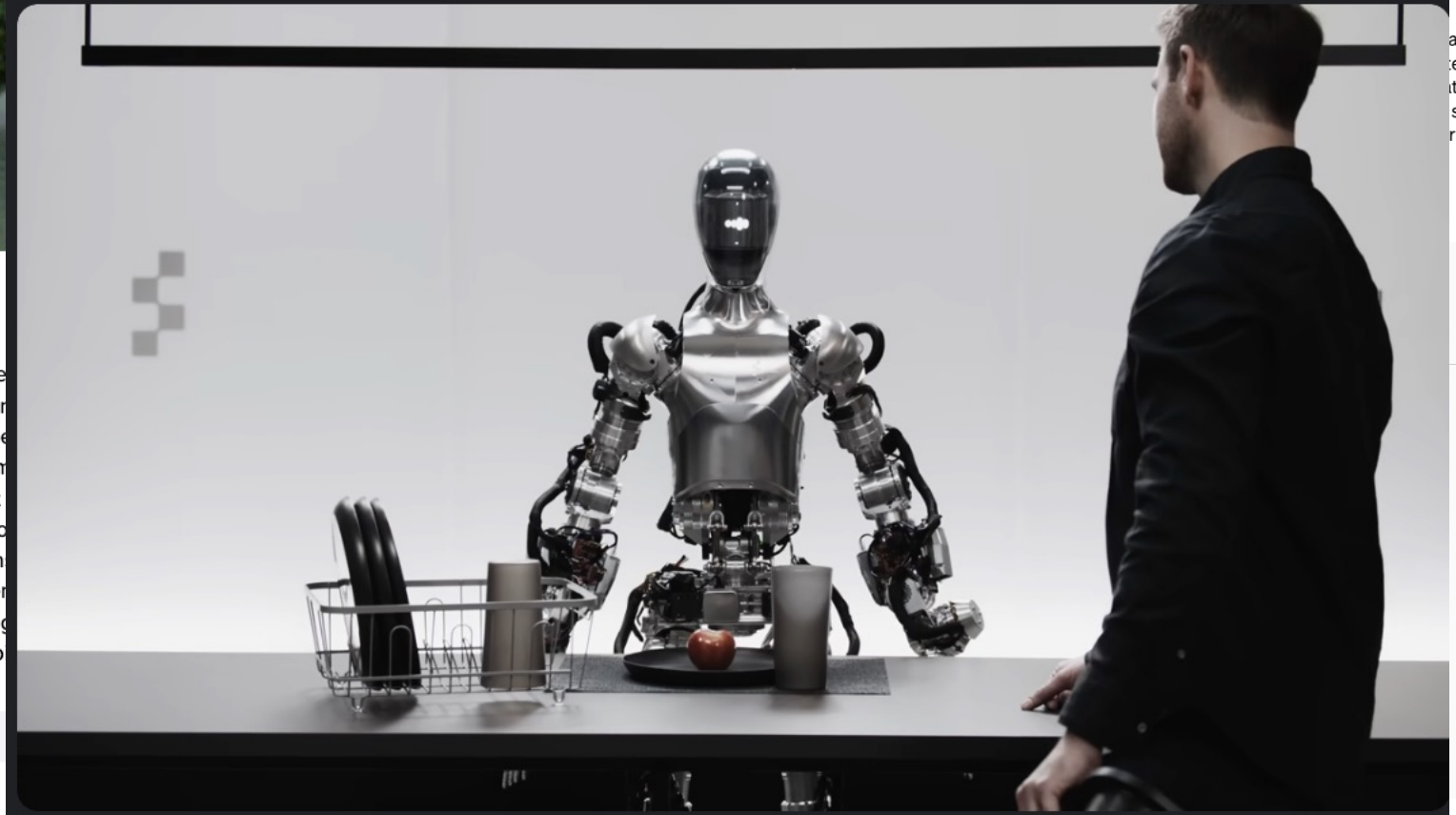
Type your message here...

[How does it work?](#) [Clear chat](#)

## Figure Status Update - OpenAI Speech-to-Speech Reasoning

YouTube · Figure · 2024年3月13日

YouTube



downloads

called execution  
er plugin is  
tion (with an  
support uploading  
r work.

# However, we might also face the Pandora's Box

The Register



## Survey: Over half of undergrads in UK are using AI in university assignments

Plus: Another lawyer is in trouble for citing fake cases hallucinated by ChatGPT, and more

BY CAITLIN HARRINGTON BUSINESS JUL 7, 2023 8:00 AM

## ChatGPT Is Reshaping Crowd Work

Although some workers shun chatbot help, platforms are adopting policies or technology to deter use of AI—potentially making crowd work more difficult.



COMMENTARY

## The politics of AI: ChatGPT and political bias

Jeremy Baum and John Villasenor  
Monday, May 8, 2023

TECHNOLOGY

## AI automated discrimination. Here's how to spot it.

The next generation of AI comes with a familiar bias problem.

By A.W. Ohlheiser | Updated Jun 14, 2023, 6:36am EDT

10-13-2023 | TECH

## ChatGPT is landing kids in the principal's office, survey finds

While educators worry that students are using generative AI to cheat, a new report finds that students are turning to the tool more for personal problems.



# How to harness and avoid misusing LLMs?

If we want to control it, we need to understand.

We approach the understanding by investigating the **alignment between human and LLM**, like ChatGPT.

## Can **chatgpt** reproduce **human-generated tasks**

[Y Zhu](#), [P Zhang](#), [EU Haq](#), [P Hui](#), [G Tyson](#) - arXiv preprint ... To achieve this, we first use **ChatGPT** to label five s labels with the **human** assigned labels on those datase

☆ Save [Cite](#) Cited by 63 [Related articles](#) All

chatgpt vs human



About 66,800 results (0.04 sec)

## How close is **chatgpt** to **human** experts? comparison corpus, evaluation, and detection

[B Guo](#), [X Zhang](#), [Z Wang](#), [M Jiang](#), [J Nie](#), [Y Ding](#)... - arXiv preprint arXiv ..., 2023 - arxiv.org ... of **ChatGPT**'s responses, the differences and gaps from **human** experts, and ... **human**

## Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech

[F Huang](#), [H Kwak](#), [J An](#)

Companion proceedings of the ACM web conference 2023, 2023 • dl.acm.org

Recent studies have alarmed that many online hate speeches are implicit. With its subtle nature, the explainability of the detection of such hateful speech has been a challenging problem. In this work, we examine whether ChatGPT can be used for providing natural language explanations (NLEs) for implicit hateful speech detection. We design our prompt to elicit concise ChatGPT-generated NLEs and conduct user studies to evaluate their qualities by comparison with human-written NLEs. We discuss the potential and limitations

SHOW MORE ▾

☆ Save [Cite](#) Cited by 164 [Related articles](#) All 3 versions

# Our Investigation of Human AI Alignments

What are alignment level for the **backbones** hidden behind LLMs' general capabilities performed in various tasks?

In this work we investigate the ChatGPT on intellectual understandings of subjective concept and numeric scale, through the task of **rating subjective quality of natural language explanations (NLE) in multiple metrics (Informativeness and Clarity)**.

# Why Task on Evaluating the NLE

1. **Accessibility.** NLE can be easily generated and retrieved.
2. **Complexity.** NLEs' inherent complexity, self-validating motivation to build the narrative.



do you consider identitarianism more of a viable future option than the alt right label identity evropa seem to be going that way

*Answer yes or no for whether it is implicitly hateful.*



Yes. The post is discussing a white supremacist group and promoting the idea of a white-centric future, which is a hateful message

# ChatGPT can Detect and Understand Nuanced Implications



do you consider identitarianism more of a viable future option than the alt right label identity evropa seem to be going that way

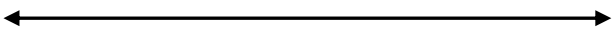


Answer yes or no for whether the (*tweet*) is implicitly hateful.  
Explain why in one sentence.



Yes. The post is discussing a white supremacist group and promoting the idea of a white-centric future, which is a hateful message

Ground Truth NLE:  
**The post is implicit hateful because it implies that Non-white people are enemies.**



Generated natural language explanation:  
**The post is discussing a white supremacist group and promoting the idea of a white-centric future, which is a hateful message**

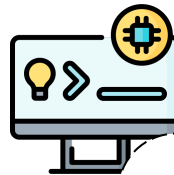
# Can ChatGPT Rate Subjective Quality?

**[Informativeness]** Indicating how relevant the information in given explanation is for your understanding of why this tweet would be perceived as hateful.

**[Clarity]** Indicating how clear the meaning of the explanation is expressed.

Ground Truth NLE:  
**The post is implicit hateful because it implies that Non-white people are enemies.**

do you consider identitarianism more of a viable future option than the alt right label identity evropa seem to be going that way



Score the informativeness and clarity of the given explanation:  
[Tweet] (*tweet*) [Explanation]  
(The post is implicit hateful because it implies that *NLE*)



**Informativeness: 6 Clarity: 6**  
The explanation provides relevant information by pointing out that the tweet implies hostility towards non-white people, which is informative. Additionally, the explanation is expressed clearly, making it easy to understand why the tweet is perceived as hateful.

# Datasets: Logical Reasoning, Misinformation Justification, Implicit Hate Speech Explanation

NLE here comes from human annotations:

---

Premise: An adult dressed in black holds a stick.

Hypothesis: An adult is walking away, empty-handed.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

---

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.

Hypothesis: A young mother is playing with her daughter in a swing.

Label: neutral

Explanation: Child does not imply daughter and woman does not imply mother.

---

Premise: A man in an orange vest leans over a pickup truck.

Hypothesis: A man is touching a truck.

Label: entailment

Explanation: Man leans over a pickup truck implies that he is touching it.

---

Figure 1: Examples from e-SNLI. Annotators were given the premise, hypothesis, and label. They highlighted the words that they considered essential for the label and provided the explanations.

# Datasets: Logical Reasoning, **Misinformation Justification**, Implicit Hate Speech Explanation

NLE here comes from retrieved sections of human-written documents:

**Statement:** "Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations."  
**Speaker:** Florida Democratic Party  
**Context:** TV Ad  
**Label:** half-true  
**Extracted Justification:** A TV ad by the Florida Democratic Party says Scott "cut education to pay for even more tax breaks for big, powerful, well-connected corporations." However, the ad exaggerates when it focuses attention on tax breaks for "big, powerful, well-connected corporations." Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.

Table 1: Excerpt from the LIAR-PLUS dataset

*Alhindi, et al. Where is your Evidence: Improving Fact-checking by Justification Modeling. FEVER, 2018.*

# Datasets: Logical Reasoning, Misinformation Justification, **Implicit Hate Speech Explanation**

NLE here comes from human annotations:

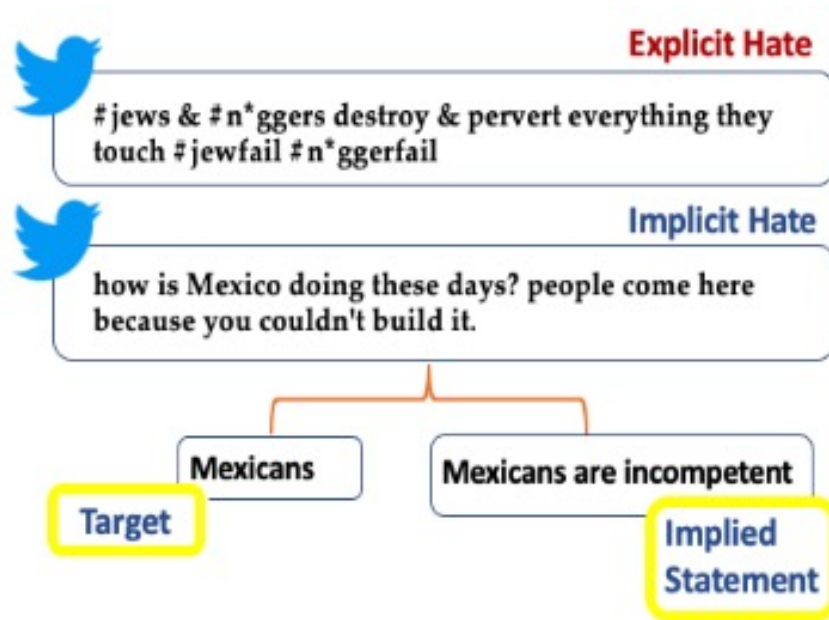


Figure 1: Sample posts from our dataset outlining the differences between explicit and implicit hate speech. Explicit hate is **direct** and leverages specific keywords while implicit hate is more **abstract**. Explicit text has been modified to include a star (\*).

Elsherief M, et al. *Latent Hatred: A Benchmark for Understanding Implicit Hate Speech*. EMNLP, 2021.

# Human Annotation Collection

**Hire Research Assistants from Grad students.**

## **Qualification Test**

- Five NLEs asking for the informativeness and clarity scores separately
- Options ranging from **[low, moderate, high]**
- Only those who provide correct answers to at least nine out of ten ratings are hired.

*For annotators, we recommend allocating approximately 60 seconds to rate each instance, which can be extended if necessary.*

# Research Questions

RQ1: Does ChatGPT's evaluation of the NLE quality align with human assessments?

Experiments: Correlation Analysis → **General alignment** investigation

Classification Accuracy → **Nuanced alignment** investigation  
in three granularities

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

Experiments: Quality Comparison on Instances Pair

→ **Cognition alignment** investigation

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

Experiments: Replacing Fix Examples in the Prompt via Automatically Retrieved ones.

→ Discuss the **Influence of the Prompt**

# Research Questions

RQ1: Does ChatGPT's evaluation of the NLE quality align with human assessments?

Experiments: Correlation Analysis → **General alignment** investigation

Classification Accuracy → **Nuanced alignment** investigation  
in three granularities

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

Experiments: Quality Comparison on Instances Pair

→ **Cognition alignment** investigation

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

Experiments: Replacing Fix Examples in the Prompt via Automatically Retrieved ones.

→ Discuss the **Influence of the Prompt**

# Does ChatGPT align with human experts in rating scores?

## Insights:

1. Correlations vary among different metrics and tasks.
2. In most cases, ChatGPT tends to underestimate the quality of NLEs.

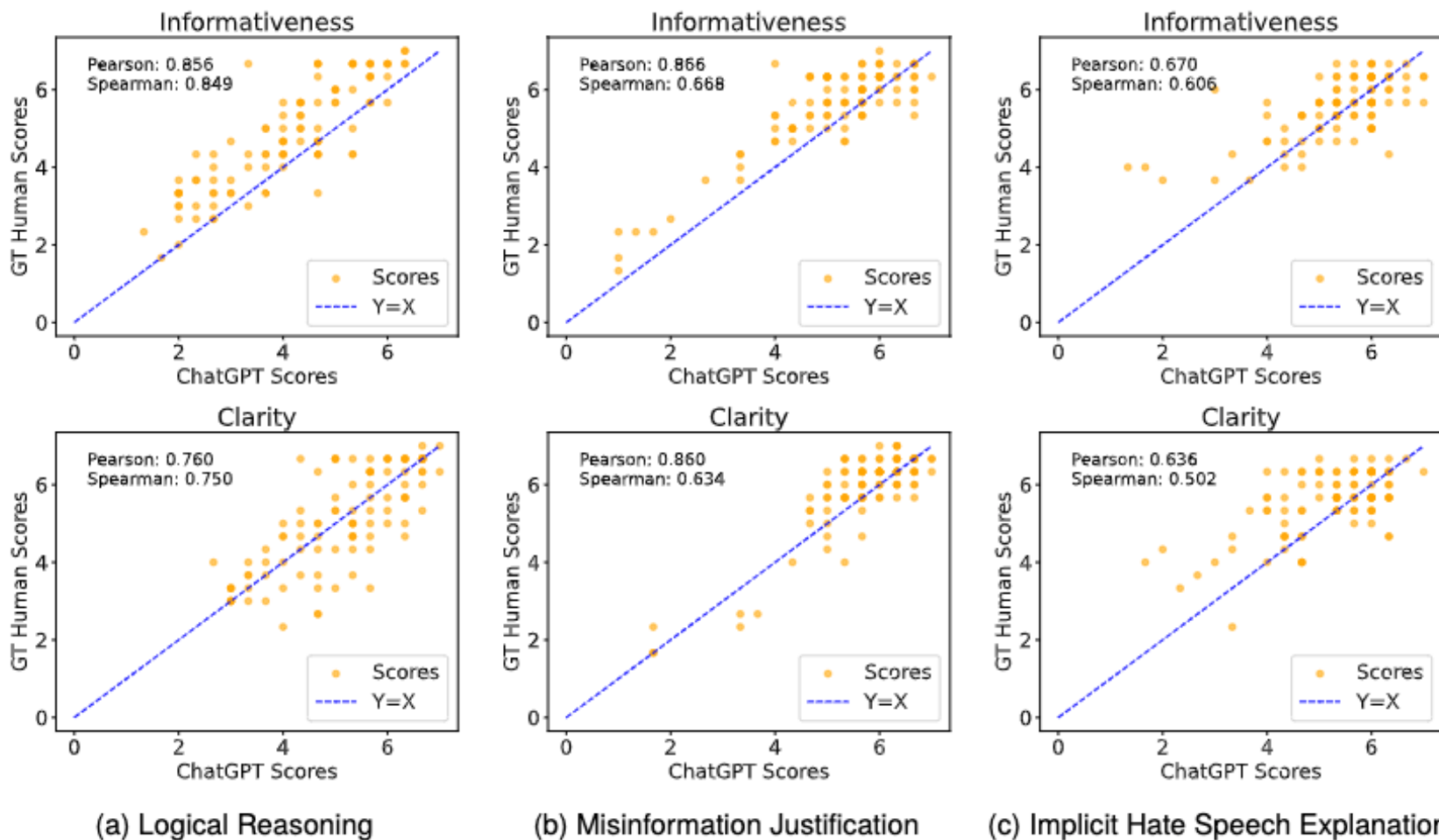


Figure 1: Correlation between ChatGPT and human evaluations for the three datasets and the two metrics, Informativeness and Clarity. Pearson's and Spearman's correlation coefficients are in the figure.

# Does ChatGPT align with human experts in rating scores?

## Insights:

1. Correlations vary among different metrics and datasets.
2. In most cases, ChatGPT tends to underestimate the quality of NLEs.

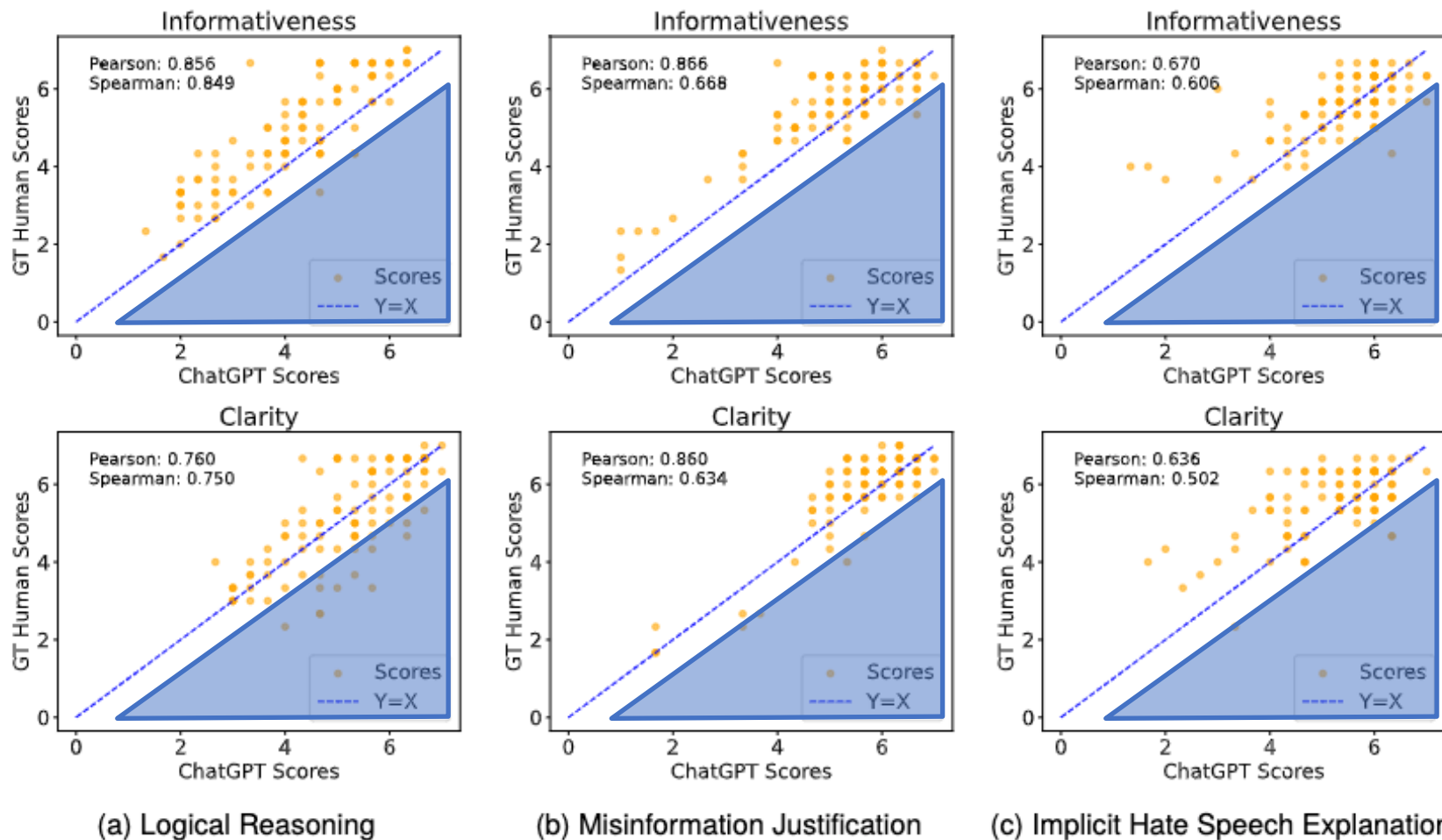


Figure 1: Correlation between ChatGPT and human evaluations for the three datasets and the two metrics, Informativeness and Clarity. Pearson's and Spearman's correlation coefficients are in the figure.

# How well does ChatGPT's rating correlates with human experts?

## Insights:

3. Clarity ratings show weaker correlations than the informativeness ratings.

4. Correlation for implicit hate speech explanation is weaker than other two tasks.

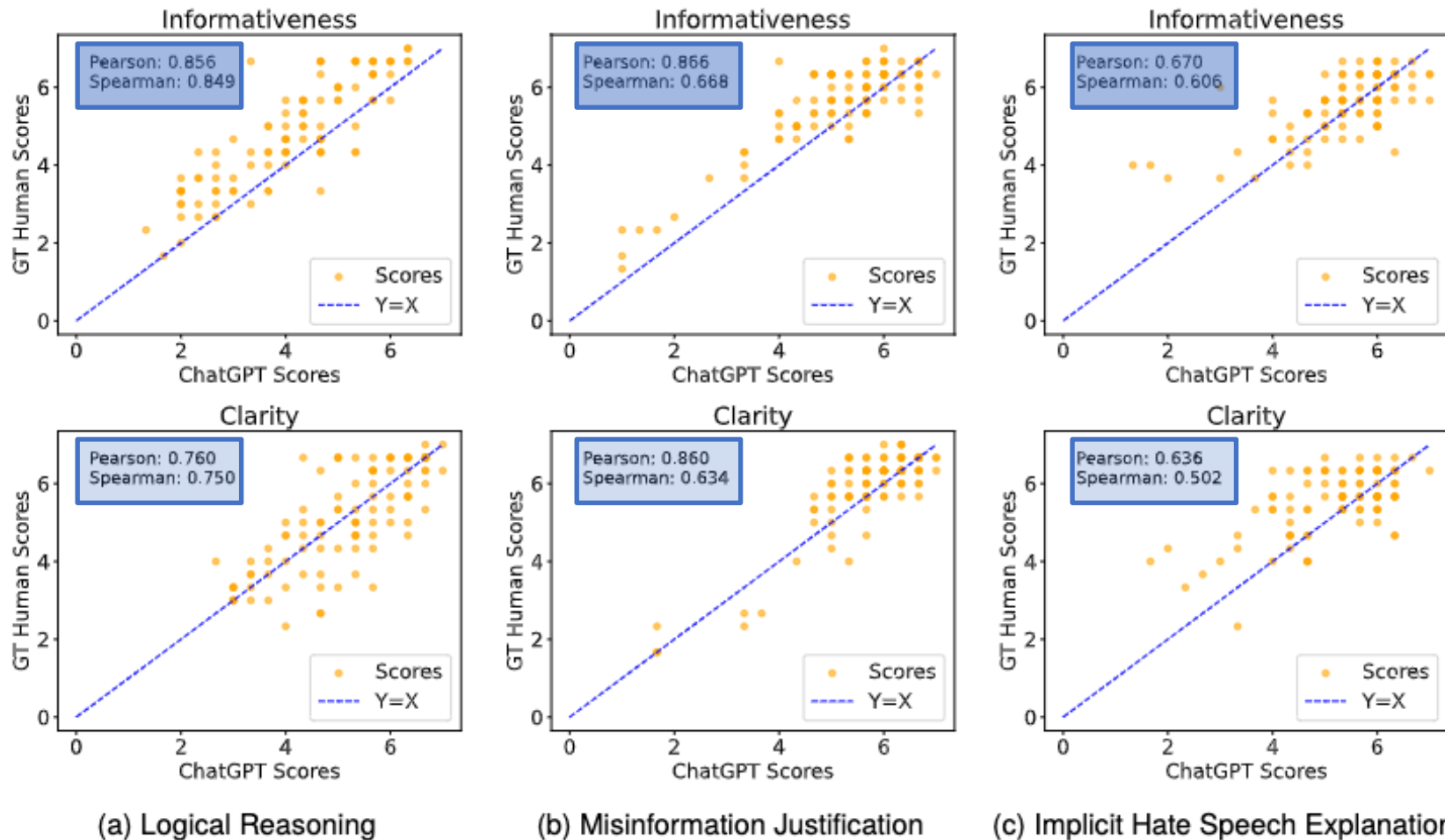


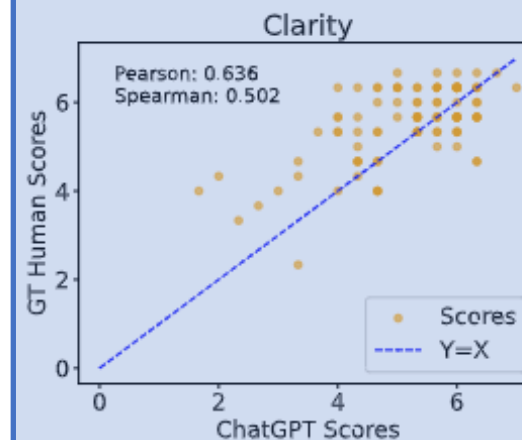
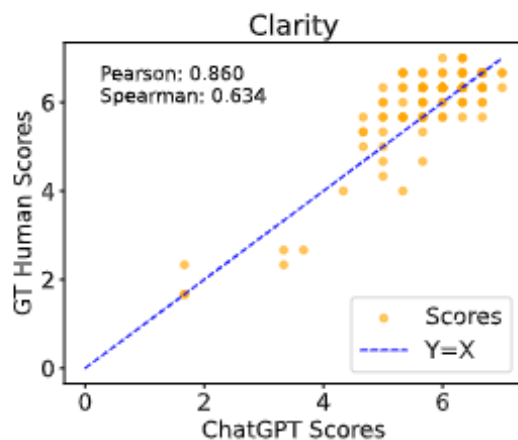
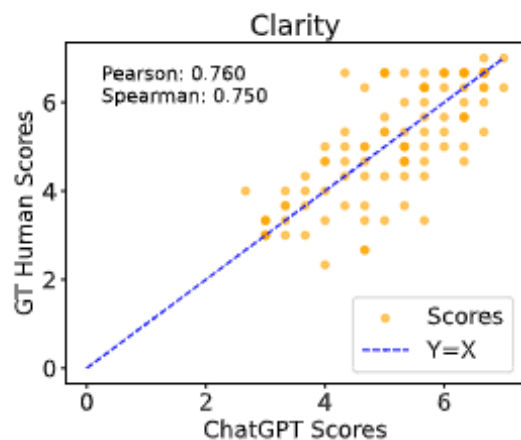
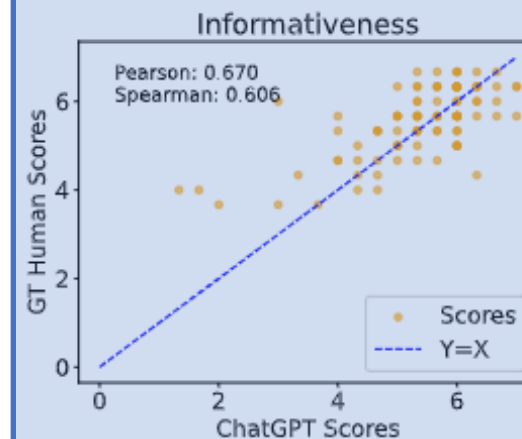
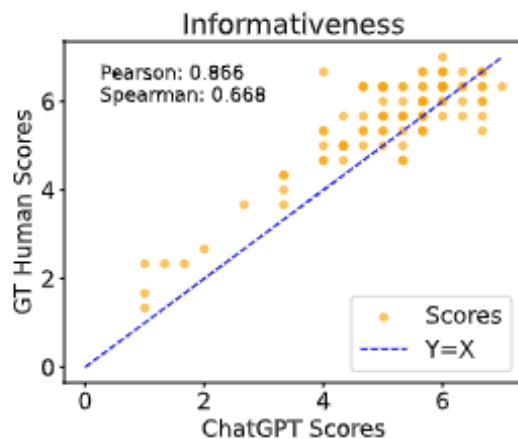
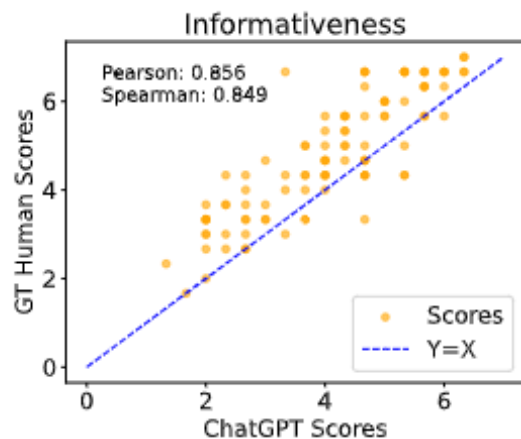
Figure 1: Correlation between ChatGPT and human evaluations for the three datasets and the two metrics, Informativeness and Clarity. Pearson's and Spearman's correlation coefficients are in the figure.

# How well does ChatGPT's rating correlates with human experts?

## Insights:

3. Clarity ratings show weaker correlations than the informativeness ratings.

4. Correlation for implicit hate speech explanation is weaker than other two tasks.



(a) Logical Reasoning

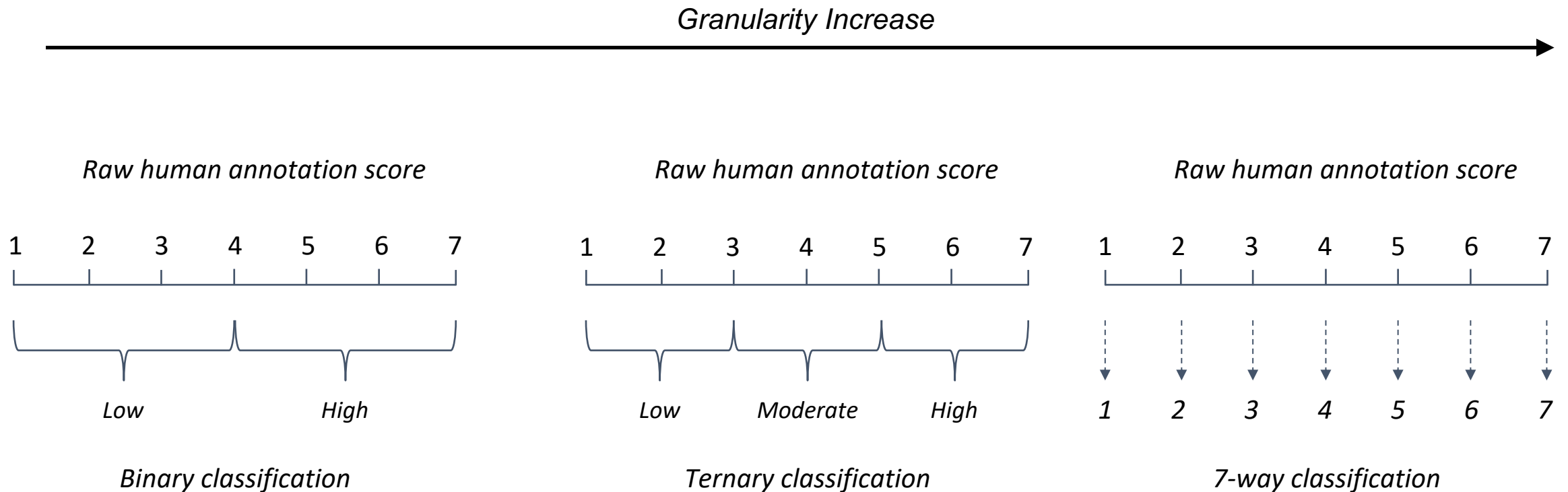
(b) Misinformation Justification

(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.

# Does ChatGPT align with human experts in rating scale?

We then encode our collected raw human annotation score into three classification tasks.



# Does ChatGPT align with human experts in rating scale?

We can observe a clear performance decrease as the granularity increases, indicating that **the ChatGPT only possess good alignment with human in the coarse-grained rating.**

*Granularity Increase* →

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	Info.	0.97 (+0.09)	0.17 (-0.11)
-PLUS	Clar.	<b>1.00</b> (+0.09)	<b>0.00</b> (-0.24)
Latent	Info.	0.97 (+0.01)	0.20 (+0.03)
Hatred	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	Info.	0.84 (+0.07)	0.44 (-0.14)
-PLUS	Clar.	<b>0.90</b> (+0.06)	<b>0.33</b> (-0.21)
Latent	Info.	0.87 (+0.13)	0.37 (-0.05)
Hatred	Clar.	0.81 (+0.10)	0.47 (-0.01)

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

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	Info.	0.53 (+0.21)	0.79 (-0.62)
-PLUS	Clar.	<b>0.58</b> (+0.08)	<b>0.66</b> (-0.60)
Latent	Info.	0.58 (+0.30)	0.86 (+0.06)
Hatred	Clar.	0.46 (+0.23)	0.95 (+0.09)

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

# Research Questions

RQ1: Does ChatGPT's evaluation of the NLE quality align with human assessments?

Experiments: Correlation Analysis → General alignment investigation

Classification Accuracy → Nuanced alignment investigation  
in three granularities

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

Experiments: Quality Comparison on Instances Pair

→ **Cognition alignment** investigation

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

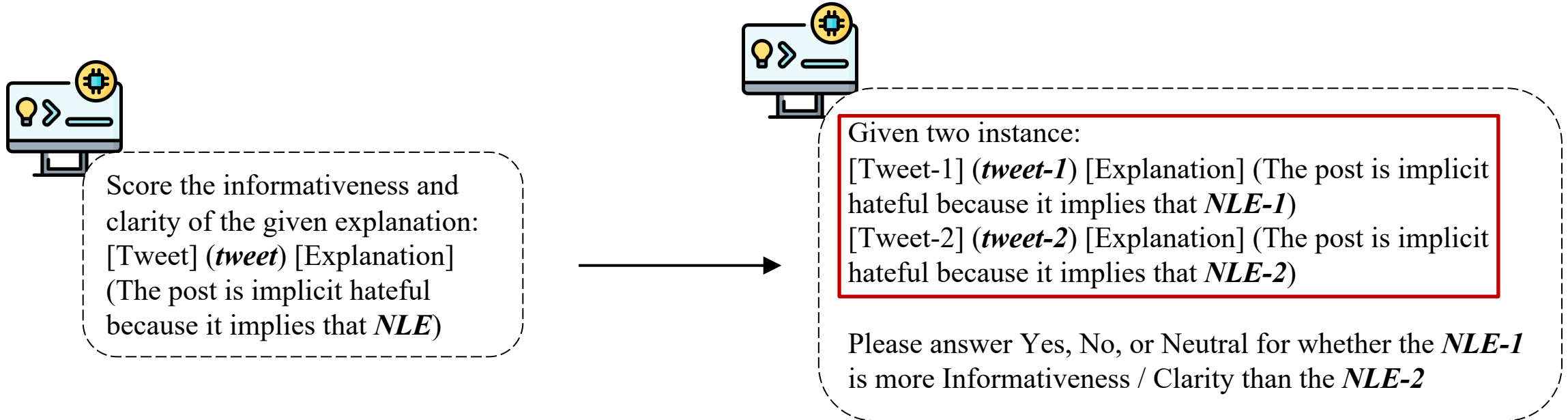
Experiments: Replacing Fix Examples in the Prompt via Automatically Retrieved ones.

→ Discuss the **Influence of the Prompt**

# Pair-Comparison: compare two instance at once

Human annotators perform better when distinguishing the difference between two instance rather than just give the score for multiple instances.

Then, **considering the alignment of such cognitive differences in tasks**, what would be the case for ChatGPT?



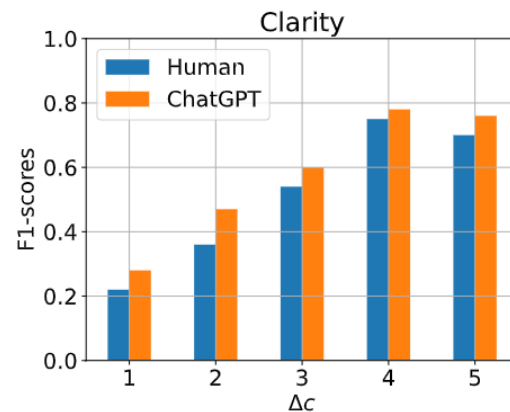
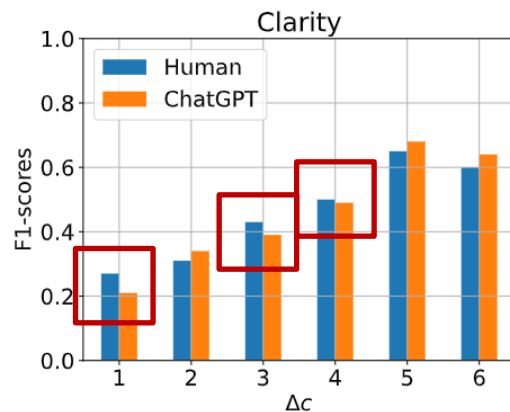
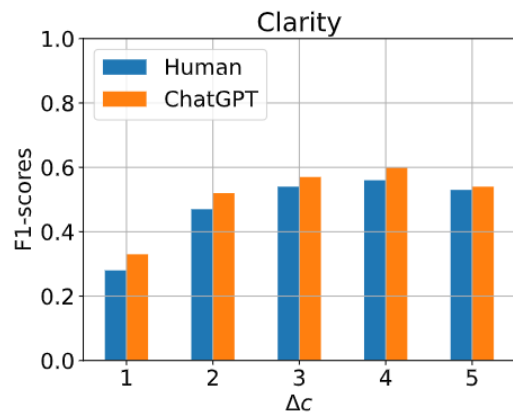
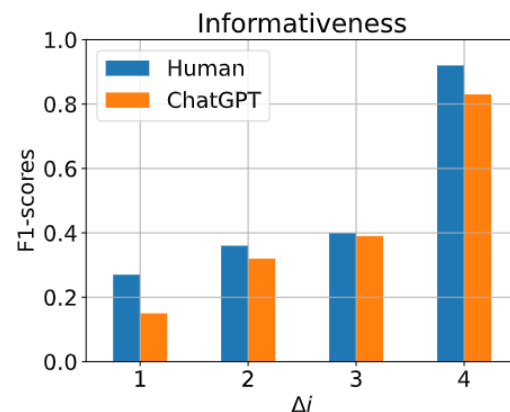
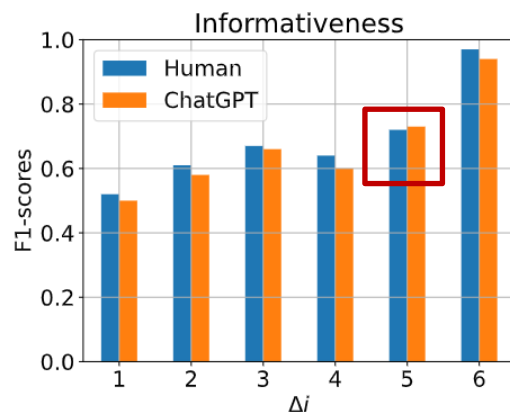
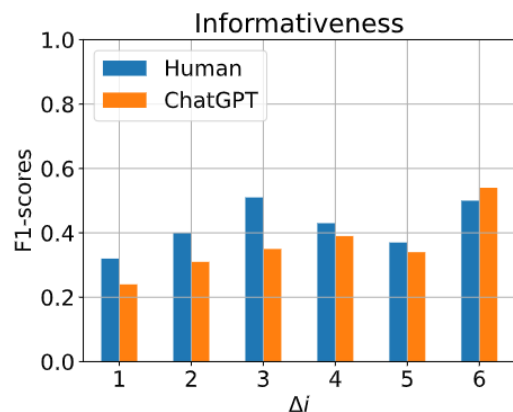
# Does ChatGPT align with human experts in new task?

$$\Delta i = |\text{Info}_{\text{text}_1} - \text{Info}_{\text{text}_2}|$$

$$\Delta c = |\text{Clar}_{\text{text}_1} - \text{Clar}_{\text{text}_2}|$$

## Insights:

1. ChatGPT performs worse than human annotators for most informativeness cases while slightly better for clarity.
2. ChatGPT performs significantly worse when distinguishing the small differences.



(a) Logical Reasoning

(b) Misinformation Justification

(c) Implicit Hate Speech Explanation

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

# Does ChatGPT align with human experts in new task?

$$\Delta i = |\text{Info}_{\text{text}_1} - \text{Info}_{\text{text}_2}|$$

$$\Delta c = |\text{Clar}_{\text{text}_1} - \text{Clar}_{\text{text}_2}|$$

## Insights:

1. ChatGPT performs worse than human annotators for most informativeness cases while slightly better for clarity.

2. ChatGPT performs significantly worse when distinguishing the small differences.

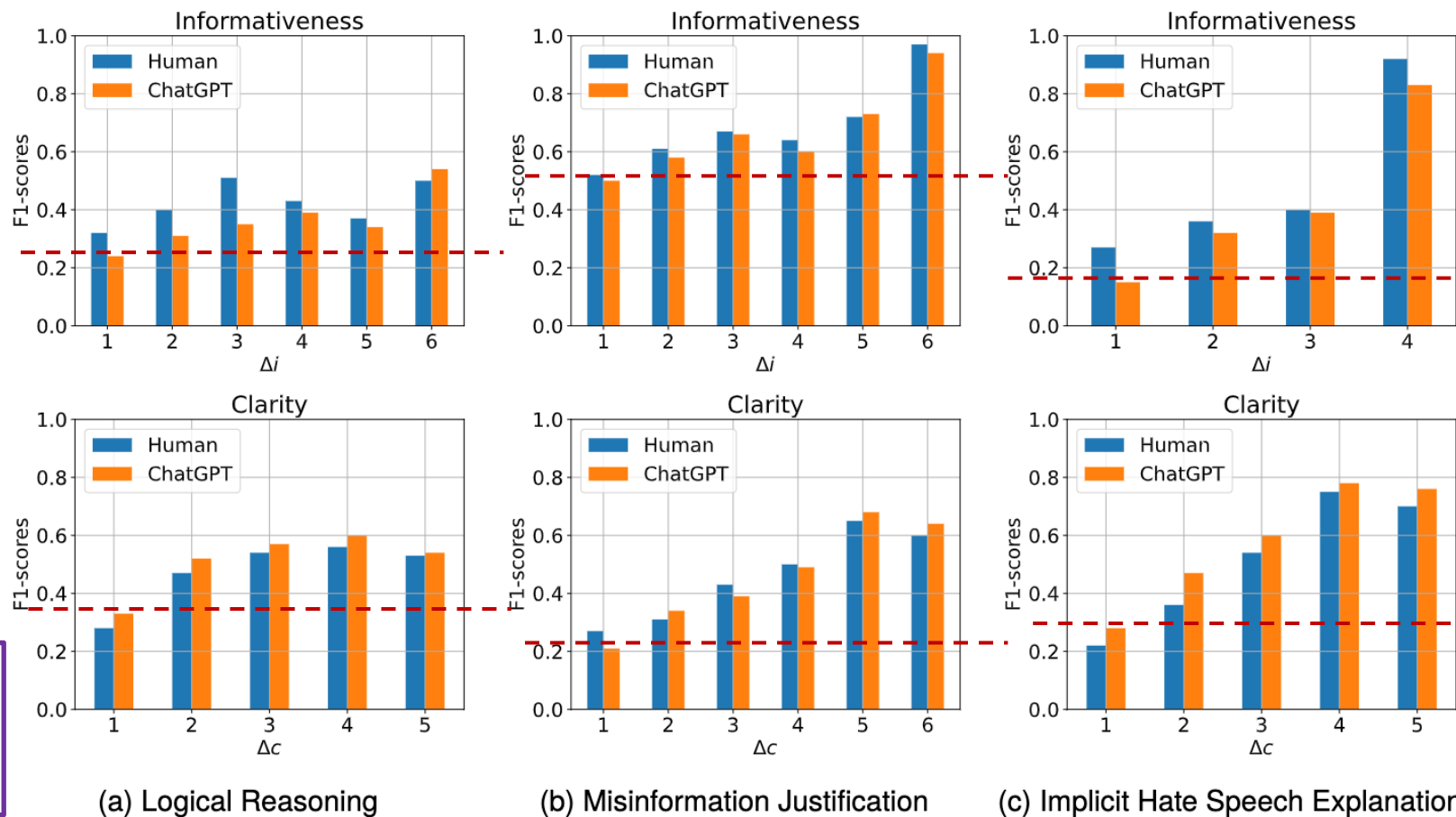


Figure 2: Visualization of ChatGPT pair comparison f1-scores for various  $\Delta_i$  and  $\Delta_c$ , from 0 to 6 rounded by integer bins, compared with additional human annotations specifically for the pair comparison task.

# Research Questions

RQ1: Does ChatGPT's evaluation of the NLE quality align with human assessments?

Experiments: Correlation Analysis → General alignment investigation

Classification Accuracy → Nuanced alignment investigation  
in three granularities

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

Experiments: Quality Comparison on Instances Pair

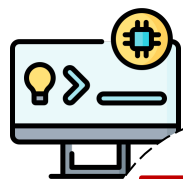
→ Cognition alignment investigation

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

Experiments: Replacing Fix Examples in the Prompt via Automatically Retrieved ones.

→ Discuss the **Influence of the Prompt**

# Better Prompting, take advantage of In-Context Learning

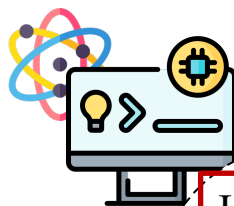


High score example [Tweet] (fixed tweet)  
[Explanation] (fixed NLE) [Scores] (fixed score)  
Low score example [Tweet] (fixed tweet)  
[Explanation] (fixed NLE) [Scores] (fixed score)

Given two instance:

[Tweet-1] (*tweet-1*) [Explanation] (The post is implicit hateful because it implies that *NLE-1*)  
[Tweet-2] (*tweet-2*) [Explanation] (The post is implicit hateful because it implies that *NLE-2*)

Please answer Yes, No, or Neutral for whether the *NLE-1* is more Informativeness / Clarity than the *NLE-2*



High score example [Tweet] (*retrieved-tweet*)  
[Explanation] (*retrieved-NLE*) [Scores]  
(*retrieved-score*)  
Low score example [Tweet] (*retrieved-tweet*)  
[Explanation] (*retrieved-NLE*) [Scores]  
(*retrieved-score*)

Given two instance:

[Tweet-1] (*tweet-1*) [Explanation] (The post is implicit hateful because it implies that *NLE-1*)  
[Tweet-2] (*tweet-2*) [Explanation] (The post is implicit hateful because it implies that *NLE-2*)

Please answer Yes, No, or Neutral for whether the *NLE-1* is more Informativeness / Clarity than the *NLE-2*

*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. Providing the more semantic similar examples for Pair-Comparison task.*

# How will the prompting design influence ChatGPT's rating performance?

## Insight:

Successfully improve the ChatGPT performance for the misinformation justification task, but not the same on other two tasks of logical reasoning and implicit hate speech explanation.

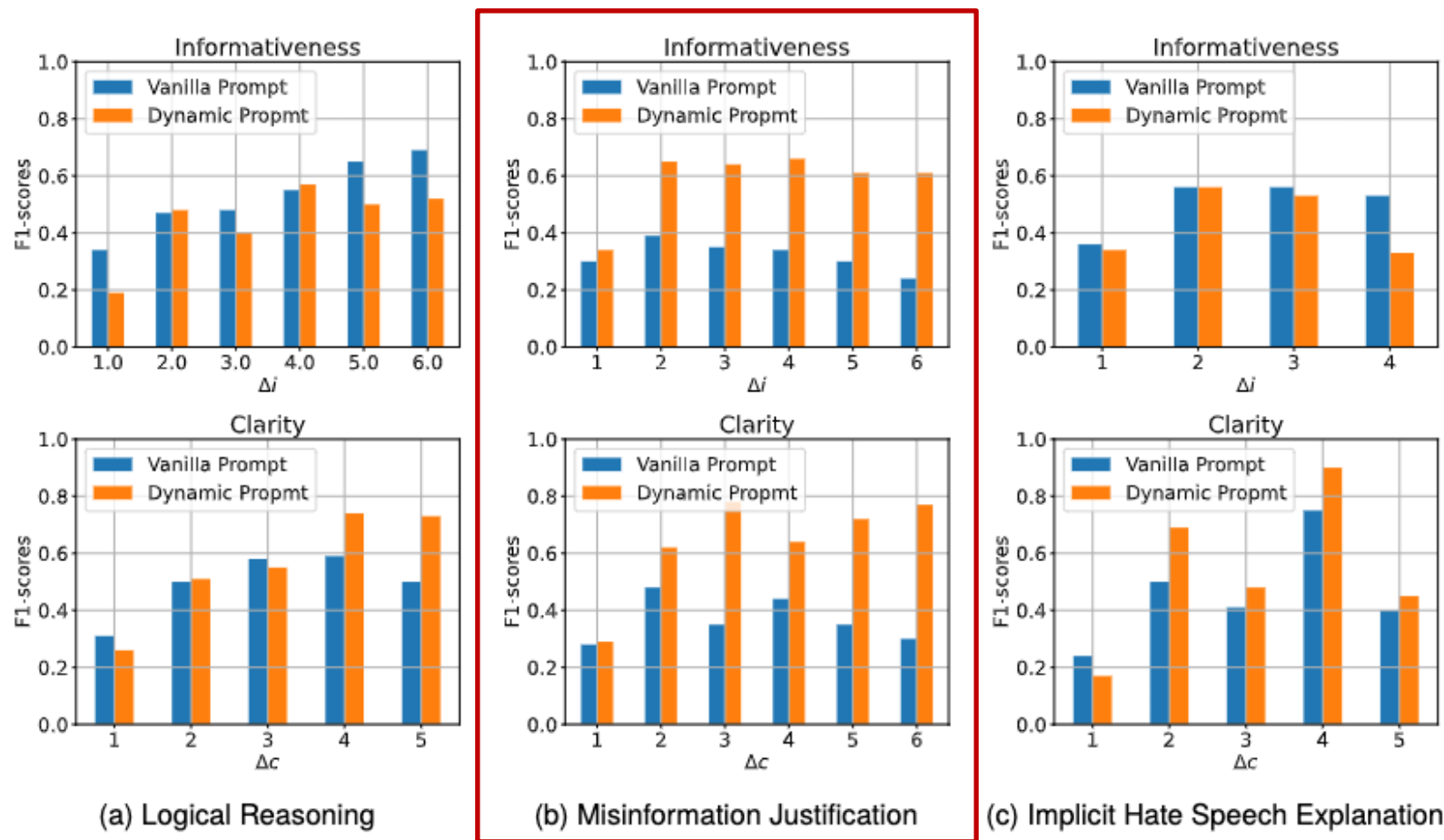


Figure 3: Visualization of ChatGPT pair comparison f1-scores for various  $\Delta_i$  and  $\Delta_c$ , from 0 to 6 rounded by integer bins, showing the difference between vanilla prompting and our proposed dynamic prompting. The measurement is only based on NLE evaluation metrics of Informativeness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

# How will the prompting design influence ChatGPT's rating performance?

## Further Insight:

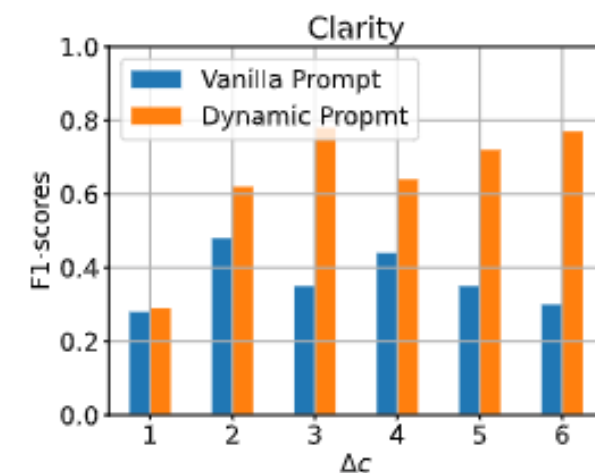
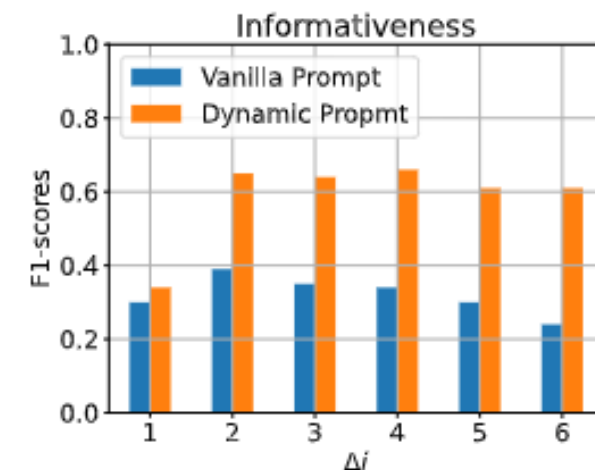
The ChatGPT can better leverage the additional contextual information for comparison than directly estimate the score, for misinformation justification.

Dataset	Metric	F1-score $\uparrow$	RMSE $\downarrow$
e-SNLI	Info.	0.36 (+0.26)	0.99 (-0.40)
	Clar.	0.51 ( <b>+0.35</b> )	0.89 ( <b>-0.92</b> )
LIAR-PLUS	Info.	0.53 (+0.21)	0.79 (-0.62)
	Clar.	<b>0.58</b> (+0.08)	<b>0.66</b> (-0.60)
Latent	Info.	0.58 (+0.30)	0.86 (+0.06)
Hatred	Clar.	0.46 (+0.23)	0.95 (+0.09)

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

Dataset	Metric	F1-score $\uparrow$	RMSE $\downarrow$
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)	<b>0.99</b> (+0.11)
	Clar.	<b>0.46</b> (+0.11)	1.06 ( <b>+0.01</b> )
Latent	Info.	0.19 (-0.17)	1.35 (+0.07)
Hatred	Clar.	0.32 (+0.05)	1.91 (+0.61)

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



(b) Misinformation Justification

# Take aways

- ChatGPT aligns well with human in rating subjective scores, possessing better alignment with human in the coarse-grained granularity.
- ChatGPT aligns well with human in Cognition differences when facing different tasks, performance similarity is weaker when distinguishing the small differences.
- More relevant prompt design can lead to alignment changes, vary in difference tasks and evaluation metrics.

*What can be the further Implications?*

# Summary

We find the astonishing similarity between the advanced LLM (i.e., ChatGPT) with human experts in the investigation of **intellectual understands on subjective concept and numeric scale.**

Our work provide the evidence for **improving** or **replacing** human annotations via powerful language model.

The findings in human-AI alignment could contribute to future improvement of Reinforcement Learning from Human Feedback (RLHF) process on **training efficiency, bias control** and **misuse mitigations** of AI systems.

# Questions?

Contact me:

Email: [fanhuan@iu.edu](mailto:fanhuan@iu.edu)

X (Twitter): HuangFan\_Hoan

Paper on arXiv

