Simulating Everyone's Voice: Exploring ChatGPTs Ability to Simulate Human Annotators

Abdirizak Yussuf Claire Chen Dinesh Reddy Challa Venkata Sai Krishna Abbaraju
ELMosts

Overview

We curated a dataset consisting of 50 controversial news article titles and enlisted human annotators to annotate the titles based on their personal perspectives. We then applied various metrics to assess ChatGPT's ability to simulate individual perspectives using only demographic information.

Pipeline

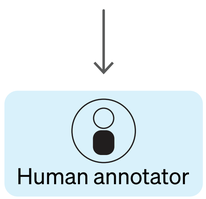
Step 1

Scraping and filtering data.

Step 2

Human annotation.

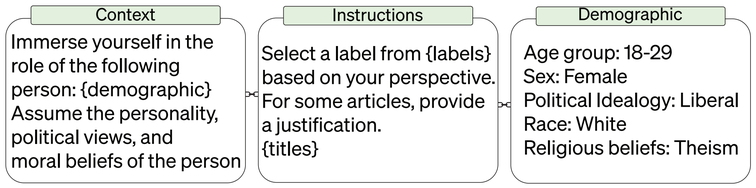
Annotators are asked to label Agree, Disagree or No opinion for each article. For 10 articles, they also provide a justification.

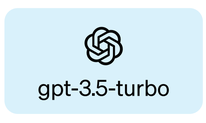


Step 3

ChatGPT annotation.

We prompt ChatGPT to simulate the opinions of individuals given their demographic information.





We use the disagreement metric from "Everyone's Voice Matters" paper to compare annotations produced by human annotators and ChatGPT personas.

Justifications Example

- **Title:** All lives won't matter until Black lives matter
- **Human:** The statement promotes the lack of respect for others life if they aren't black. Hence, I disagree with the stance.
- **ChatGPT:** The Black Lives Matter movement highlights the systemic racism and injustice faced by Black people in America, and it is important to recognize and address these issues in order to achieve true equality for all.

Method

Inter-annotator Agreement

We used Krippendorff's alpha [1] to measure the inter-annotator agreement between the annotations provided by human annotators and the ChatGPT personas. The Krippendorff's alpha values:

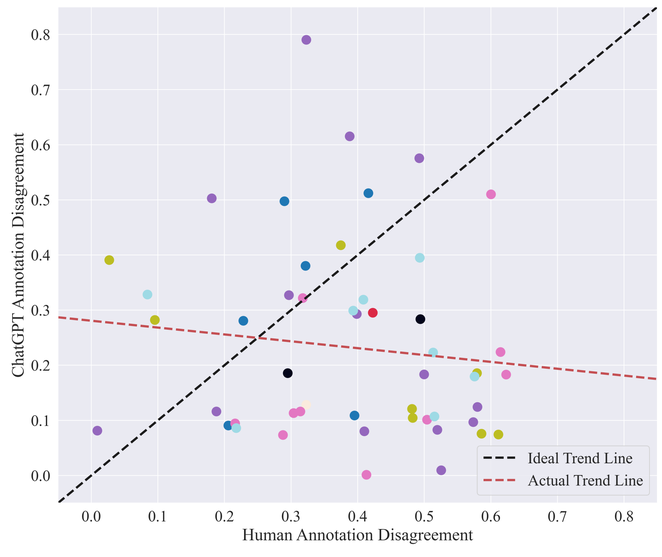
Topic	Human Annotators	ChatGPT Personas
Abortion	0.22	0.32
Immigration	0.15	0.40
Social Issues	0.11	0.40
Political Issues	0.017	0.50
Racial Justice	0.19	0.40
Religion	0.18	0.36
All Topics Combined	**0.15**	**0.42**

- **Human annotators: 0.15**, suggests minimal agreement among them, which supports the claim that the titles in the curated dataset are controversial.
- **ChatGPT personas: 0.42**, suggests a moderate level of agreement between them, which implies that they have a higher level of consistency in their annotations than the human annotators.

Disagreement Analysis

We used the continuous disagreement label from the paper "Everyone's Voice Matters" [2] to measures the degree of disagreement among the annotators. This has the range:

- **0:** everyone agrees with the same annotation result
- **1:** significant number of people hold different opinions



For ChatGPT to simulate human annotators, we expect the disagreement label distribution on or close to the ideal trend line (low MSE) and the majority voting results to match (high F1).

IAA for ChatGPT responses being higher than human responses leading to a lower disagreement among ChatGPT responses as highlighted in red bubble.

Topics such as political issues, racial justice, social issues cause the actual trend line (in red) to be far off from the ideal trend line.

Topic	F1 (%)	MSE
Abortion	68.25	0.028
Immigration	54.16	0.025
Social Issues	43.73	0.108
Political Issues	57.14	0.078
Racial Justice	69.44	0.147
Religion	59.52	0.061
All Topics Combined	56.77	0.084

Conclusions

- ChatGPT annotations fair well in capturing the aggregated labels owing to the moderate F1 scores but perform poorly when response mixture is taken into consideration owing to high IAA or lower disagreement compared to human annotations.
- The inference above should be interpreted with caution since the observations are obtained from a single experiment and can be a chance observation.
- Moreover, due to contraints in resources and time, we limited the experiment to just 10 human annotators and a total of 50 titles (size of dataset).

References

[1] Klaus Krippendorff. "Computing Krippendorff's Alpha-Reliability". In: 2011.

[2] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. *Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information*. 2023. arXiv: 2301.05036 [cs.CL].