NeuroCure
Cluster of Excellence

DRIVERS JOURNAL CLUB

Use of ChatGPT to Explore Gender and Geographic Disparities in Scientific Peer Review

P. Sebo

Journal of Medical Internet Research, Short Paper, 2024

Charité Universitätsmedizin Berlin
Helmholtz Research for Grand Challenges
Leibniz Association

Funded within the Excellence Strategy by the **German Federal** and **Berlin State Governments** through:

DFG Deutsche Forschungsgemeinschaft
EXC 2049 / project number: 390688087

Context

Earlier DRIVERS discussion:

Koch, S. (2020). Responsible research, inequality in science and epistemic injustice: an attempt to open up thinking about inclusiveness in the context of RI/RRI.

- With focus on gender and geography, „…how to make science internally more inclusive and tackle barriers that prevent marginalised scholars from participating in knowledge production and societal meaning-making.”

Today’s DRIVERS discussion:

Sebo, P. (2024). Use of ChatGPT to Explore Gender and Geographic Disparities in Scientific Peer Review

Introduction

Background: Increasing focus on transparency and accountability has led some journals publish peer review reports

Study focus: Use of ChatGPT to analyse peer review reports for possible gender or geographic bias, by scoring for:

- Sentiment – opinions, attitudes, emotions categorized as +, -, 0
- Politeness - rude, neutral, polite language

Why: To shed light on potential of AI to mitigate biases and promote fairness in scholarly communication

Introduction

Assumption/Assertion: ChatGPT is accurate, with advantages over other methods*:

- Understands natural language, picks up on nuances, context-specific sentiments
- Adapts to diverse domains without training
- AI-driven sentiment analysis is efficient, scalable, consistent, objective (devoid of human biases and preconceptions)

*based on Verharen, J.P.H., ChatGPT identifies gender disparities in scientific peer review. *Elife,* 2023

Methods

How: 12 papers from 9 high-impact general medical journals were analyzed

Journal	2022 impact factor, n	Papers (n=108), n (%)	Reviews (n=291), n (%)	Sentiment score, median (IQR)	Politeness score, median (IQR)
BMJ	107.7	12 (11.1)	51 (17.5)	68 (53-78)	73 (60-80)
PLOS Medicine	15.8	12 (11.1)	41 (14.1)	60 (50-73)	70 (62-78)
BMC Medicine	9.3	12 (11.1)	29 (10.0)	63 (48-70)	73 (60-78)
Journal of Clinical Medicine	3.9	12 (11.1)	31 (10.7)	43 (17-70)	50 (25-67)
Diagnostics	3.6	12 (11.1)	28 (9.6)	37 (3-60)	57 (28.5-68.5)
Journal of Personalized Medicine	3.4	12 (11.1)	27 (9.3)	57 (17-70)	68 (27-82)
BMJ Open	2.9	12 (11.1)	27 (9.3)	63 (40-72)	60 (45-70)
BMC Primary Care	2.9	12 (11.1)	25 (8.6)	57 (10-67)	55 (30-70)
Medicina	2.6	12 (11.1)	32 (11.0)	48.5 (20-65)	57 (43.5-66)

-5 most represented countries: UK (14), Germany (13), US (10), China(9), Italy (6)
-3 main regions of affiliation: Western Europe (56), Asia (16), North America (15)

Methods

Prompt for ChatGPT:

Below you will find a scientific peer review. Can you score this peer review on the sentiment, on a scale from –100 (negative) to 0 (neutral) to 100 (positive), and politeness of language use, on a scale of –100 (rude) to 0 (neutral) to 100 (polite)?

Scores stratified by:

- journal
- gender (based on first, last name)
- affiliation (N. America, Europe, Pacific /Asia / Latin America, Middle East, Africa)

Results

Variable	Papers (n=108), n (%)	Reviews (n=291), n (%)	Sentiment score, median (IQR)	Crude *P* value[a]	Adjusted *P* value[b]	Politeness score, median (IQR)	Crude *P* value[a]	Adjusted *P* value[b]
First authors' gender				.49	.48		.37	.68
Woman	47 (43.5)	127 (43.6)	58 (33-72)			65 (47-77)		
Man	61 (56.5)	164 (56.4)	57.5 (27-70)			63 (48.5-73)		
Last authors' gender				.52	.88		.74	.86
Woman	33 (30.6)	91 (31.3)	57 (33-68)			63 (50-75)		
Man	75 (69.4)	200 (68.7)	60 (27.5-72)			63 (45-75)		
First authors' affiliation				.001	.02[c]		<.001	.02[d]
North America, Europe, and Pacific	82 (75.9)	220 (75.6)	60 (33-72)			67 (53-77)		
Asia	16 (14.8)	43 (14.8)	62 (40-70)			65 (50-73)		
Middle East, Latin America, and Africa	10 (9.3)	28 (9.6)	27 (–3 to 55)			43.5 (17.5-57)		

Discussion

- The author found no gender disparity, against expectations

  Explanation: DEI efforts in scholarly community may have been more effective in mitigating gender disparities than other forms of bias

- Results showed regional disparities, aligning with prior research highlighting challenges facing researchers from Global South

  Explanation: must be understood in terms of sociocultural, economic, institutional contexts – limited funding opportunities, language barriers, cultural differences

  → However… there may be other explanations for discrepancies in politeness

Examples from original data – supplemental material

Boston

I was initially cautious about accepting this review as the study is based in the US. Having read the paper though I can see how the results may help inform health equality work in the UK as well.

I think the researchers have followed a robust process to focus in on those differences that cannot be explained away by system factors. However, I was disappointed to see that there has been little involvement of the public (beyond the researchers' own experiences) in shaping or implementing this study.

Reviewers all native English speakers

Mexico

The research method is good, what about the inclusion and exclusion criteria along with how you choose the sample and how about the ethical test that you carried out? Please explain.

The results and discussion are quite good, add your opinion to the discussion which is supported by theory.

Conclusions must answer the research objectives.

Reviewers all non-native English speakers

Limitations

Identified by author:

- …
- Manual scoring not performed for comparison
- Uncertainty about gender and geographic distributions of rejected papers

Possible additional limitations:

- Does not take into account regional affiliation of reviewers
- No information on criteria used to measure sentiment and politeness
- No explanation for regional groupings (e.g. N. America, Europe, Pacific vs. Asia)
- AI objectivity not questioned
- Author cites himself 6 times in a list of 28 references

Questions

- This paper showed a discrepancy in language sentiment and politeness based on country affiliation, but did it convincingly show bias?

- Is measuring sentiment and politeness a good metric for capturing biases in peer review?

- All the papers analyzed here were actually accepted by the journals. Can disparities in language sentiment and politeness be considered bias/discrimination if they do not lead to rejection of the papers?

- Is there nevertheless a negative impact of sentiment/politeness discrepancy on authors? Is it a matter of “chilly climate”?

Thank you!