

HOW INTELLIGENT ARE GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS IN MEDICAL PUBLICATIONS?

Salgo M.R. Elenjikamali^{a*}, Sonali Satam^a, Divya Narayan^a, Madhavi Patil^a, Sangita P. Patil^a, and Vatsal Shah^b

^aSIRO Clinpharm Pvt. Ltd., India; ^bSIRO Clinpharm USA LLC, Princeton, NJ, USA

*Presenting Author

INTRODUCTION

An all-pervading influence of generative artificial intelligence (AI) tools (GAITs) is evident in every sphere of scientific writing¹

Myriad applications: Literature-search, content generation, paraphrasing, translations, data interpretation or visualisation, figure-creation, detecting plagiarism, and reference management^{2,3,4}

These tools assist in novel idea-generation, creation of content by combining contextual comprehension and human-like responses which enhance the productivity for time-consuming preparatory activities of manuscript development^{1,2}

Widespread adoption of GAITs has the potential to revolutionise the landscape of scientific writing; though it is imperative to validate these tools

CONCLUSIONS

ChatGPT Bard



Content generation



Data interpretation and summarisation

perplexity Bard



Reference searches

• Bing, though efficient, isn't very useful as a standalone AI tool for content generation

• Thus, all 4 GAITs possess exclusive features and can assist authors/researchers to efficiently and rapidly develop high-quality publications, but require human intervention to check their veracity

METHODOLOGY

Objectives

To evaluate the selected 4 GAITs for quality and efficiency of content in medical publication

ChatGPT 3.5

Bard Experiment

perplexity Pro

Bing

SELECTED GAITs

The 4 GAITs were analysed using a blinded multi-rater (n=3) approach (score: 0-10)

Content generation

- For 3 therapeutic areas (TAs):
 - Oncology
 - Neuroscience
 - Cardiovascular Metabolism
- Using general and specific search prompts (2 each/TA)

Figures/tables summarisation

- Data interpretation and summarisation of tables and figures
- ChatGPT, Perplexity cannot analyse figures

Scientific accuracy and quality

- Assess generated content and summarised figures, tables based on outcomes

Statistical analysis

- Descriptive statistical analysis used for all outcomes
- Data outputs were presented as Mean (SD)

Outcomes

1. Accuracy

Check accuracy of factual information by verifying correctness of specific claims/data/statements against published references

2. Informativeness/comprehension

Assess whether the text includes informative novel/unique insights with logical flow

3. Human-like tone

Assess if generated content has human-like tone/naturalness

4. References

Evaluate original references from which the content is curated and verify credibility of sources.

5. Grammar/syntax

Evaluate quality of language used in the generated text, including grammar and spelling errors

6. Speed

Assess the time taken for the generated text to appear

7. Readability

Assess how engaging/readable and coherent the generated text is to the target audience

8. Reproducibility

Measure consistency in the content generated with multiple raters

9. Inter-output variability

Measure the variability of generated content basis the prompts utilised

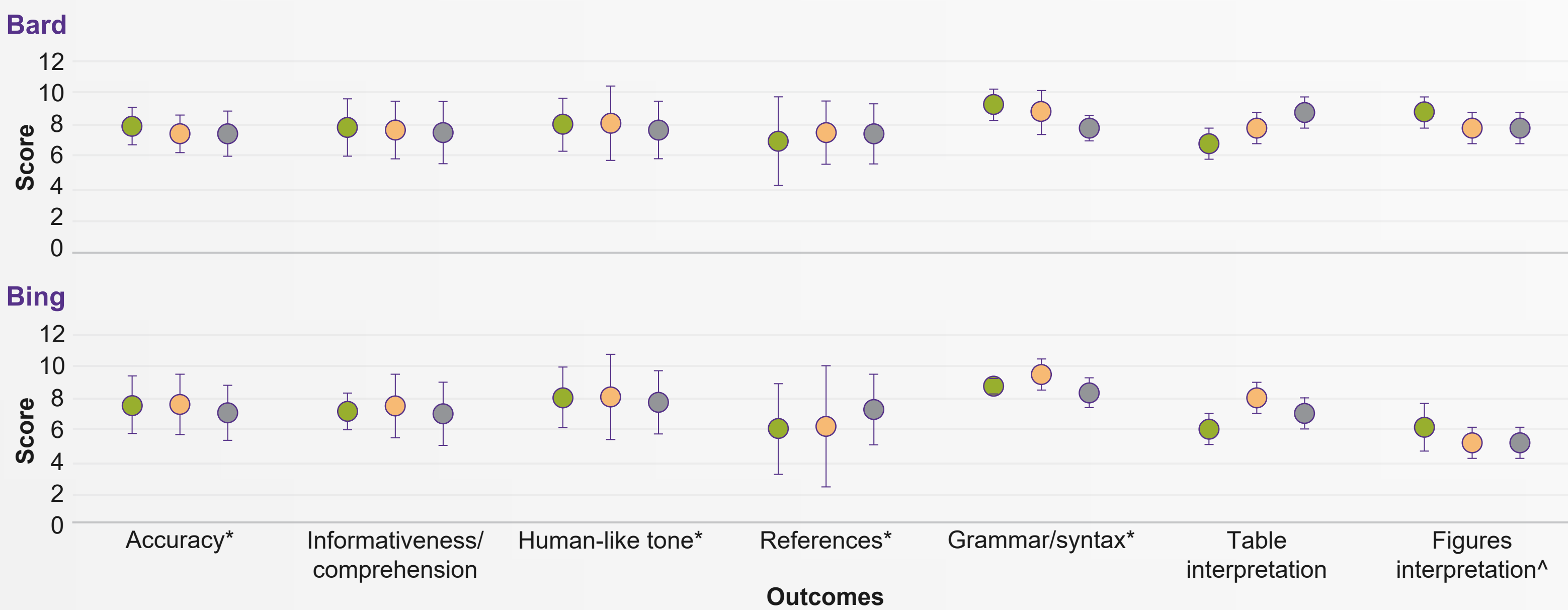
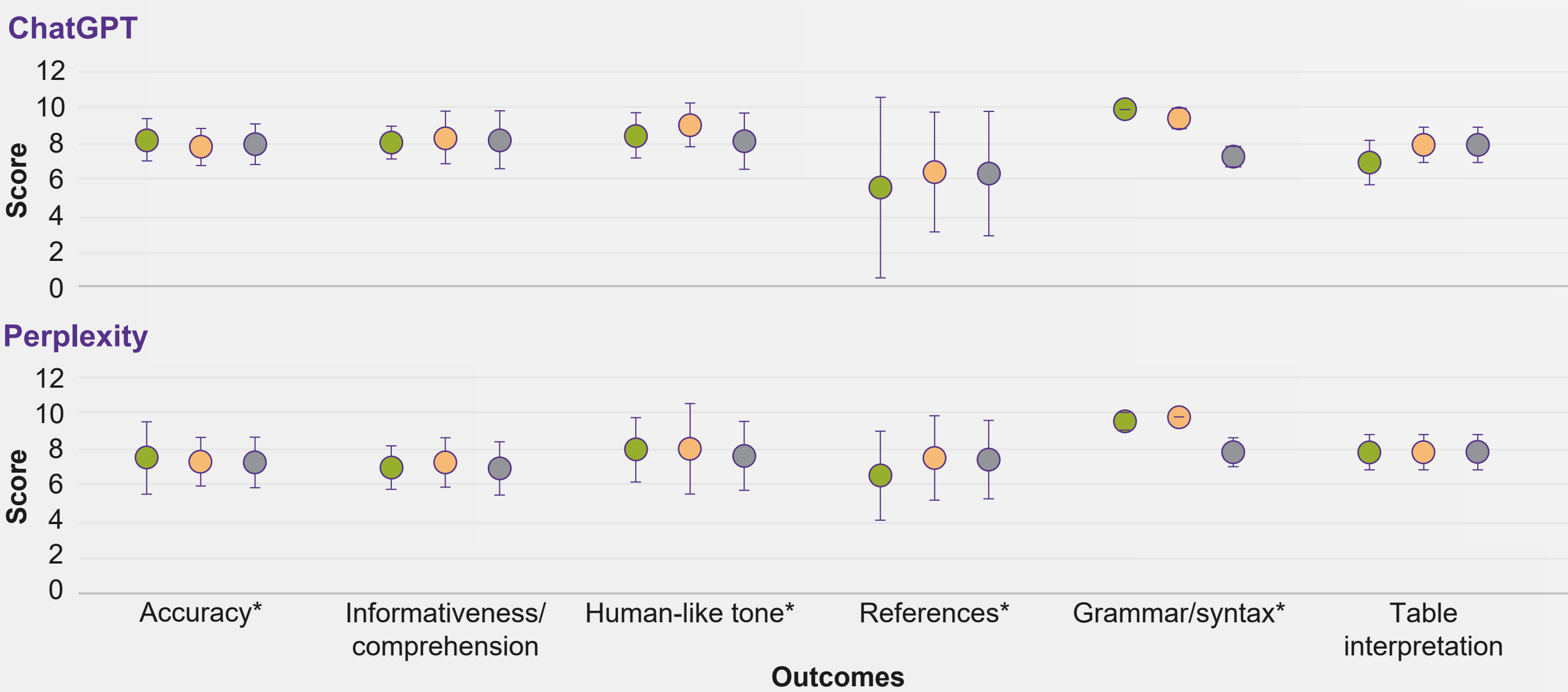
RESULTS

Comparative analysis of GAITs for publication outputs

ChatGPT and Bard generated content efficiently

Perplexity and Bard were more efficient for reference searches

Bard interpreted data from tables/figures efficiently



Note: Each analysis was performed on outputs from 3 raters and mean (±SD) values were reported.

*Accuracy, Informativeness, Human-like tone, Grammar/syntax and References were graded on a scale of 0 to 10 and were categorized as poor<5, below-average:6, average:7, good:8, excellent:9,10. *Figures cannot be uploaded to ChatGPT and Perplexity for analysis

Therapy area	Oncology				Neurosciences				Cardiovascular Metabolism			
Outcome	ChatGPT	Perplexity	Bing	Bard	ChatGPT	Perplexity	Bing	Bard	ChatGPT	Perplexity	Bing	Bard
Time (seconds, mean [SD])	25.08 (12.1)	14.50 (10.7)	28.25 (11.1)	17.25 (1.7)	23.33 (12.8)	15.25 (12)	29.50 (9.3)	18.5 (7.6)	29.00 (13.4)	16.67 (11.6)	29.00 (8.6)	19.33 (6.4)
Readability*	15.90	13.03	13.73	11.68	17.43	14.80	15.18	12.40	16.00	14.60	12.68	11.80
Reproducibility*	6.25	6.75	6.00	6.00	6.25	6.00	6.00	6.00	6.75	6.75	6.25	6.75
Inter-output variability [‡]												
Generalised	0.17	0.83	0.33	0.33	0	0.5	0	0.17	0	0.5	0.17	0.67
Specialised	0.33	0.17	0.57	0.17	0.17	0	0	0	0.17	1.17	0.67	0.57

*Readability was graded as per the Flesch Kincaid Reading tool. *Differences in mean accuracy scores of each output for generalized and specialized prompts were evaluated. *The three outputs generated for the same prompt by each GAIT were assessed for similarity.

	Key Strengths	Key Limitations
ChatGPT	Generates relevant scientific content in human-like tone • Consistent level of performance across queries in different TAs • Minimal grammar/syntax or editorial errors	No access to the real-time updates/recent information • Does not generate references voluntarily • Does not accept files/figures as input or provide any image as an output in response to queries
Bard	Provides an option to share the output • Files/figures can be uploaded as a 'query', and data from the uploaded files can be summarised • Generates recent references as it has access to 'real-time' information • Can verify the information generated with an inbuilt Google button	Does not generate references voluntarily • Tone of generated content was occasionally colloquial and non-scientific and lacked human-like tone/flow
Perplexity	Rapid response generated • Additional tabs to search images and videos for query • Provides suggestions for sequential queries • In-text references for specific statements listed along with content generated	Content sometimes lacked human-like tone, coherence, flow, readability • Files/figures can be uploaded as input for queries but requires payment
Bing	Figures can be uploaded as a query, and data from the uploaded files can be summarised • Provides suggestions for sequential queries • In-text references for specific statements listed along with the content generated • Has the option to make responses more creative, balanced, or precise	Short and crisp responses may not capture all facts for a scientific query • Listed references may not be always desirable for scientific writing • Tone of generated content was occasionally colloquial and non-scientific

References
1. Altınbaş S, Sola-Lejva A. *Reprod Biomed Online*. 2023 Jul;47(1):3-9.
2. Katja Martin. *Medical Writing*. 2023 Sep;32(3):22-27. 3. Leung TL, et al. *J Med Internet Res*. 2023 Aug 31;25:e51584. 4. GPT-4 technical report. OpenAI. 2023. Mar. <https://cdn.openai.com/papers/gpt-4.pdf>.

Acknowledgements
Ganesh Karpe and Mugdha Rokade (SIRO Clinpharm Pvt. Ltd., India) provided graphic design assistance for the development of this poster.

Disclosures
All authors are employees of SIRO Clinpharm Pvt. Ltd.



Scan the QR code for the full digital poster

The QR code is intended to provide scientific information for individual reference and the information should not be altered or reproduced in any way.