Evaluating the Use of Generative Artificial Intelligence in Academic Writing for Surgery and Urology: Instrument Validation Study

Nathaniel Hong-En Heah, Wei Tseng Tzen, Elyn Yo-Lin Tzen, Kon Voi Tay, Zhiwen Joseph Lo, Cheuk Fan Shum, Marcus Way Lunn Chow, Leynard Manuel Marrero, Phyo Aung, Gregory Kang Ee Heng, Ishara Maduka, Juefei Feng, Zhongkai Wang, Lin Yee Koong, Serene Ee Ling Tang, Cathy Po Ching Ng, Shane Wen Hui Sim, Man Hon Tang, Pravin Lingam, Shaun Wen Yang Chan, Yuan Teng Cho, Chee Wei Lee, Sadhana Chandrasekar, Wei Chong Chua

Submitted to: JMIR AI
on: November 10, 2024

Table of Contents

Evaluating the Use of Generative Artificial Intelligence in Academic Writing for Surgery and Urology: Instrument Validation Study

Nathaniel Hong-En Heah[1]; Wei Tseng Tzen[2]; Elyn Yo-Lin Tzen[2]; Kon Voi Tay[1]; Zhiwen Joseph Lo[1]; Cheuk Fan Shum[1]; Marcus Way Lunn Chow[1]; Leynard Manuel Marrero[1]; Phyo Aung[1]; Gregory Kang Ee Heng[1]; Ishara Maduka[1]; Juefei Feng[1]; Zhongkai Wang[1]; Lin Yee Koong[1]; Serene Ee Ling Tang[1]; Cathy Po Ching Ng[1]; Shane Wen Hui Sim[1]; Man Hon Tang[1]; Pravin Lingam[1]; Shaun Wen Yang Chan[1]; Yuan Teng Cho[1]; Chee Wei Lee[1]; Sadhana Chandrasekar[1]; Wei Chong Chua[1]

[1]Department of Surgery, Woodlands Health Singapore SG
[2]MOH Holdings Singapore SG

Corresponding Author:
Nathaniel Hong-En Heah
Department of Surgery, Woodlands Health
17 Woodlands Dr 17
Singapore
SG

Abstract

Background: The use of artificial intelligence (AI) has gained rapid uptake in academic, scientific, and medical writing. While it can be helpful in summarising data and providing opinions, some studies have demonstrated limitations in accuracy and potential bias.

Objective: This study aims to assess the accuracy of academic writing in the context of general surgery and urology topics and their suitability for use in scientific publications.

Methods: 20 subject matter experts provided 20 questions each for 9 General Surgery and Urology subspecialties. The questions were further classified into 4 categories: epidemiology, pathophysiology, clinical management, and surgical techniques. These questions were submitted to ChatGPT v3.5 requesting for 10 answers for each question using the following prompt “Write me 10 points on XXX, for submission to JAMA medical journal for publication, providing different reference for each point”. The primary outcomes rated by subject matter experts were: accuracy and breadth of the answers; language used in medical writing; number and accuracy of the references. Secondary outcome measured was the ability to detect AI-generated text using Giant Language-model Test Room (GLTR).

Results: There were 180 questions posed to ChatGPT, with 120 General surgery (68.6%) and 60 Urology questions (31.4%). Overall accuracy of information provided in the answers was 1560/1800 (86.7%), with good overall breadth 163/180 (90.5%) and language 167/180 (92.7%). Answers provided for General Surgery were significantly more accurate than Urology (90.0% vs 79.6%, p<0.001). Only 5.5.% of references provided were true published references, with the majority being fictitious. Subgroup analysis showed a statistical difference in accuracy between the 4 categories of epidemiology (402/405, 89.3%), pathophysiology (404/450, 89.8%), clinical management (387/450, 86.0%), and surgical techniques (370/450, 82.2%) questions (p<0.001). For the secondary outcomes, frac(p) was 0.994, indicating that almost all statements were identified to be AI generated.

Conclusions: While the information provided in ChatGPT is accurate with sufficient breadth and with appropriate medical academic language, its quotation of fictitious references limits its use for academic writing.

(JMIR Preprints 10/11/2024:68575)
DOI: https://doi.org/10.2196/preprints.68575

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Evaluating the Use of Generative Artificial Intelligence in Academic Writing for Surgery and Urology

Nathaniel HE HEAH, Wei Tseng TZEN, Elyn YL TZEN, Kon Voy TAY, Zhiwen Joseph LO, Cheuk Fan SHUM, Marcus WL CHOW, Leynard M MARRERO, Phyo AUNG, Gregory KE HENG, Ishara MADUKA, Jue Fei FENG, Zhongkai WANG, Lin Yee KOONG, Serene EL TANG, Cathy PC NG, Shane WH SIM, Man Hon TANG, Pravin LINGAM, Shaun WY CHAN, Yuan Teng CHO, Chee Wei LEE, Sadhana CHANDRASEKAR, Wei Chong CHUA
Department of Surgery, Woodlands Health, Singapore

Corresponding Author
Nathaniel HE Heah, MBBS, MRCS, FAMS
Department of Surgery, Woodlands Health
17 Woodlands Dr 17, Singapore 737628
Email: nat.heah@gmail.com

Abstract

Background: The use of artificial intelligence (AI) has gained rapid uptake in academic, scientific, and medical writing. While it can be helpful in summarising data and providing opinions, some studies have demonstrated limitations in accuracy and potential bias.

Objective: This study aims to assess the accuracy of academic writing in the context of general surgery and urology topics and their suitability for use in scientific publications.

Methods: 20 subject matter experts provided 20 questions each for 9 General Surgery and Urology subspecialties. The questions were further classified into 4 categories: epidemiology, pathophysiology, clinical management, and surgical techniques. These questions were submitted to ChatGPT v3.5 requesting for 10 answers for each question using the following prompt “Write me 10 points on XXX, for submission to JAMA medical journal for publication, providing different reference for each point”. The primary outcomes rated by subject matter experts were: accuracy and breadth of the answers; language used in medical writing; number and accuracy of the references. Secondary outcome measured was the ability to detect AI-generated text using Giant Language-model Test Room (GLTR).

Results: There were 180 questions posed to ChatGPT, with 120 General surgery (68.6%) and 60 Urology questions (31.4%). Overall accuracy of information provided in the answers was 1560/1800 (86.7%), with good overall breadth 163/180 (90.5%) and language 167/180 (92.7%). Answers provided for General Surgery were significantly more accurate than Urology (90.0% vs 79.6%, p<0.001). Only 5.5.% of references provided were true published references, with the majority being fictitious. Subgroup analysis showed a statistical difference in accuracy between the 4 categories of epidemiology (402/405, 89.3%), pathophysiology (404/450, 89.8%), clinical management (387/450, 86.0%), and surgical techniques (370/450, 82.2%) questions (p<0.001). For the secondary outcomes, frac(p) was 0.994, indicating that almost all statements were identified to be AI generated.

Conclusion: While the information provided in ChatGPT is accurate with sufficient breadth and with appropriate medical academic language, its quotation of fictitious references limits its use for academic writing.

Keywords: Artificial Intelligence, Academic Writing, ChatGPT

Introduction

Artificial intelligence (AI) is currently considered the next frontier in medical innovation, finding use in a variety of aspects in the medical industry, from the initial use in diagnostic radiology, detecting medical errors and reducing additional paperwork (such as writing discharge summaries), to recent uses in drug development, medical research editing and regulatory writing[1]. The use of AI has gained rapid uptake as a powerful assistive tool in academic, scientific, and medical writing in recent years[2,3]. In combination with Natural Language Processing (NLP), generative AI has advanced to be able to provide well-rounded answers to questions asked in natural human language[4]. Generative AI uses existing training data to generate new outputs by observing patterns and structures from the input data and leveraging generative models to generate new data that has similar characteristics. These models rely on deep learning techniques and neural networks to analyse and generate content closely mimicking human response[5].

ChatGPT is one of the largest large language models (LLM) developed by OpenAI, with more than 175 billion parameters and 45 terabytes of training data[5]. It has been and remains free to use, with more than 10 million users within 2 months of release. ChatGPT is a language model based on Generated Pre-trained Transformer (GPT) architecture, utilising a large source of data for processing language tasks and text generation. This transformer architecture trains ChatGPT through the following processes: neural network, unsupervised pre-training, and reinforced learning from human feedback (RLHF)[6].

The neural network is inspired by the human brain, and it aids the language model in analysing and classifying large sets of data into a network of data tokens emulating the structure of a human brain - with interconnecting neurons organised into layers, each connection activated accordingly with appropriate inputs to generate a response[6]. This allows the language model to recognise the relationship between every word, and thus allows it to recognise the meaning of a word or phrase in context.

The LLM undergoes unsupervised pre-training through analysing a large corpus of data to recognise and generate patterns in language, allowing it to predict the next word in a text, given the sequence and arrangement of words within given text[7]. This process is unsupervised, meaning no inputs or prompts are used in each language task, and a large data set is directly fed into the language model, without any human feedback as to whether an output is appropriate[5].

This process is optimised through RLHF. RLHF allows fine tuning of the language model policy with human intervention, through (1) “supervised fine tuning” - demonstration writing of a desired output by a human annotator, (2) development of a reward model of desired output through ranking by human labellers to rank each output generated based on human preference, and (3) optimisation of the language model policy through reinforcement learning against the reward model. After the policy has been optimised, a different input is then tested against the reward model to achieve proximal policy optimisation (PPO)[5].

So far, generative AI has shown to be useful in summarising or scouring information. However, the quality and reliability of this data can vary depending on the topic and data source[8]. Several commentaries have acknowledged the usefulness of AI to generate text based on prompts, but also emphasised the necessity of experts to review this information for accuracy and clarity before publication[9]. Although early iterations of AI generated text were easily identifiable, generative AI has since improved significantly in its ability to generate human-sounding responses; and researchers continue to develop various methods and tools to distinguish between AI-generated and human-written text[10,11]. There are also ethical challenges that this new frontier faces, such as paywalls, transparency of machine learning sources, as well as attribution, copyright and plagiarism issues that have yet to be addressed in full[12].

To date there is still no quantified error margin for these texts, specifically in the context of scientific writing, despite there being commentaries that have highlighted specific scenarios where ChatGPT fabricates data[13]. Therefore, in this study, we aim to evaluate the performance of AI-generated medical writing on general surgery (GS) and urologic topics in the context of academic writing, and to compare its content accuracy and its ability to mimic human written content.

Methods

Recruitment

20 subject matter experts from the Department of Surgery at a tertiary hospital were asked to provide 20 questions from 9 GS and Urology subspecialties (Appendix 1). These questions were classified into epidemiology, pathophysiology, clinical management and surgical techniques. The questions were submitted to ChatGPT v3.5 using the following prompt: “Write me 10 points on XXXX, for submission to JAMA medical journal for publication, providing different reference for each point”.

Evaluation Outcomes

The primary outcomes were to evaluate the accuracy of the answers, number of references, latest year of reference, the accuracy of each reference, breadth of the answers and the medical writing language. The secondary outcome was the ability of the Giant Language model Test Room (GLTR) to detect the presence of generated text in the answers[14].

Statistical Analysis

The generated answers were evaluated by the subject matter experts within the specialties of GS and Urology and references were checked against existing academic databases. Overall breadth of answers as well as medical language was assessed as Good, Moderate or Poor. The GLTR result that was reported was frac(p), which is a measure of how confident the model is in its prediction that a word in a position within an AI-generated text will actually appear. Statistical significance was measured with the Chi-Square test and ANOVA test. All statistical analysis was performed on GraphPad Prism10 (Dotmatics, MA, USA).

Results

A total of 180 questions were submitted to ChatGPT v3.5, comprising a total of 120 GS (68.6%) and 60 Urology questions (31.4%). ChatGPT was then asked to generate 10 statements for each question. Overall, subject matter experts assessed the information provided in the answers to be accurate in 1560/1800 (86.7%) of all responses. The majority of responses were also shown to have good overall breadth (90.5%) and language used for medical writing (92.7%).

Table 1. Accuracy of information, breadth of answers, language of medical writing and accuracy of references of all responses.

		Overall
Accuracy of information	Yes	1560/1800 (86.7)
	No	126/1800 (7.0)
	Partial	114/1800 (6.2)
Breadth	Good	163/180 (90.5)
	Moderate	16/180 (8.9)
	Low	1/180 (0.6)
Language	Good	167/180 (92.7)
	Moderate	10/180 (5.6)
	Low	3/180 (1.7)
Accuracy of reference	Yes	99/1800 (5.5)
	No	1521/1800 (84.5)
	Partial	178/1800 (9.9)

The majority of references quoted were from journals (1653/1671, 98.9%), with the remaining coming from textbooks (12/1671, 0.72%) and websites (6/1671, 0.36%). Significantly, only 99/1800 (5.5%) of all references provided were accurate, with the majority of references fabricated. On further scrutiny, references quoted had reasonable titles relating to the questions asked, and were quoted from journals that existed. Furthermore, many of the quoted authors of these fabricated journal articles had written articles on the relevant topic previously, but had not actually written any of the articles that were referenced in the ChatGPT responses. For those articles that existed, 41/1671 (2.5%) came from journals with an impact factor (IF) of 4.99 or less, while 129/1671 (7.72%) came from journals with IF > 5.

When analysing the different categories of surgical techniques, pathophysiology, epidemiology and clinical management, there was found to be a statistical difference between these 4 groups in the area of accuracy of information ($p<0.001$) and accuracy of references ($p<0.001$). The category with the lowest accuracy was surgical techniques (370/450, 82.2%) while the highest was pathophysiology (404/450, 89.8%), a difference of 7.6%. The category with the lowest number of accurate references was surgical techniques (7/259, 2.7%) while the category with the highest accuracy was pathophysiology (36/323, 11.1%). There were also no differences noticed between breadth of answers ($p=0.182$) or language use.

Table 2. Accuracy of information, breadth of answers, language of medical writing and accuracy of references between categories of answers.

		Surgical Techniques	Pathophysiology	Epidemiology	Clinical Management	p-value
Accuracy of information	Yes	370/450 (82.2)	404/450 (89.8)	402/450 (89.3)	387/450 (86)	
	No	54/450 (12)	20/450 (4.4)	28/450 (6.2)	24/450 (5.3)	
	Partial	26/4550 (5.8)	26/450 (5.8)	20/450 (4.4)	39/450 (8.7)	p<0.001
Breadth	Good	39/45 (86.7)	42/45 (93.3)	41/45 (91)	40/45 (88.9)	
	Moderate	6/45 (13.30	3/45 (6.7)	3/45 (6.7)	5/45 (11.1)	
	Low	0/45 (0)	0/45 (0)	1/45 (2.2)	0/45 (0)	p=0.956

Language	Good	40/45 (88.9)	41/45 (91.1)	44/45 (97.8)	42/45 (93.3)	
	Moderate	3/45 (6.7)	4/45 (8.9)	1/45 (2.2)	3/45 (6.7)	
	Low	2/45 (4.4)	0/45 (0)	0/45 (0)	0/45 (0)	p=0.866
Accuracy of reference	Yes	7/259 (2.7)	36/323 (11.1)	11/334 (3.3)	13/311 (4.2)	
	No	240/259 (92.7)	230/323 (71.2)	299/334 (89.5)	268/311 (86.2)	
	Partial	12/259 (4.6)	57/323 (17.7)	2/334 (7.2)	30/311 (9.6)	p<0.001

On further subgroup analysis, it was found that there was a significant difference between GS and Urology responses when it came to accuracy of the information (p<0.001) and references (p<0.001) in the responses. There were no differences noted between both the language used in the responses (p=0.590), or the breadth of the responses (p=0.182).

Table 3. Accuracy of information, breadth of answers, language of medical writing and accuracy of references between General Surgery and Urology.

		GS	Urology	p-value
Accuracy of information	Yes	1080/1200 (90)	478 (79.6)	
	No	72/1200 (6)	55/600 (9.2)	
	Partial	48/1200 (4)	67/600 (11.2)	p<0.001
Breadth	Good	111/120 (92.5)	51/60 (85.5)	
	Moderate	8/120 (6.7)	9/60 (14.5)	
	Low	1/120 (0.8)	0/60 (0)	p=0.182
Language	Good	109/120 (90.8)	58/60 (96.7%)	
	Moderate	8/120 (6.7)	2/60 (3.3%)	
	Low	3/120 (2.5)	0/60 (0)	p=0.590
Accuracy of reference	Yes	60/872 (6.9)	7/354 (1.8)	
	No	716/872 (82.1)	321/354 (90.7)	
	Partial	95/872 (10.9)	26/354 (7.3)	p<0.001

Further subgroup analysis was performed between the various subspecialties within GS and Urology. The subspecialty groups represented within Urology included Oncology, Endourology and Functional Urology while in GS the subspecialty groups represented were Acute Care Surgery (ACS), Colorectal surgery, Endocrine surgery, Trauma, Upper Gastrointestinal surgery and Vascular surgery.

Within Urology, it was found that there was a statistical difference in the accuracy of information (p<0.001) between Oncology (129/200, 64.5%), Endourology (149/200, 74.5%) and Functional urology (192/200, 96%). The accuracy of the references (p=0.838), breadth of response (p=0.501) as well as language (p=0.751) showed no statistical difference.

Table 4. Accuracy of information, breadth of answers, language of medical writing and accuracy of references within Urology

		Uro-Oncology	Uro-Endourology	Uro-Functional	p-value
Accuracy of information	Yes	129/200 (64.5)	149/200 (74.5)	192/200 (96)	
	No	29/200 (14.5)	28/200 (14)	1/200 (0.5)	
	Partial	42/200 (21)	23/200 (11.5)	7/200 (3.5)	p<0.001
Breadth	Good	16/20 (80)	15/20 (75)	20/20 (100)	
	Moderate	4/20 (20)	5/20 (25)	0/20 (0)	
	Low	0/20 (0)	0/20 (0)	0/20 (0)	p=0.501
Language	Good	17/20 (85)	20/20 (100)	20/20 (100)	
	Moderate	3/20 (15)	0/20 (0)	0/20 (0)	
	Low	0/20 (0)	0/20 (0)	0/20 (0)	p=0.751
Accuracy of	Yes	3/93 (3.2)	2/163 (1.2)	2/98 (2.0)	p=0.838

reference	No	83/93 (89.2)	150/163 (92.0)	88/98 (89.8)	
	Partial	7/93 (7.5)	11/163 (6.7)	8/98 (8.2)	

Among GS subspecialties, there was also noted to be a statistical difference between the accuracy of information (p<0.001) and accuracy of the references (p<0.001), while no difference was found between the breadth of the answers (p=0.679) and the language used (p=0.542). The subspecialty with the highest accuracy of information was Endocrine surgery (196/200, 98%), while the lowest accuracy was Trauma surgery (173/200, 86.5%). When analysing the accuracy of references, the most accurate subspecialty was Trauma surgery (25/188, 13.3%) while the least accurate was Endocrine surgery (4/73, 5.5%).

Table 5. Accuracy of information, breadth of answers, language of medical writing and accuracy of references within GS subspecialties.

		Acute Care Surgery	**Colorectal**	**Endocrine**	**Trauma**	**UGI**	**Vascular**	**p-value**
Accuracy of information	Yes	175/200 (87.5)	172/200 (86)	196/200 (98)	173/200 (86.5)	180/200 (90)	184/200 (92)	
	No	25/200 (12.5)	28/200 (14)	4/200 (2)	10/200 (5)	3/200 (1.5)	2/200 (1)	
	Partial	0/200 (0)	0/200 (0)	0/200 (0)	17/200 (8.5)	17/200 (8.5)	14/200 (7)	p<0.001
Breadth	Good	17/20 (85)	18/20 (90)	20/20 (100)	18/20 (90)	20/20 (100)	18/20 (90)	
	Moderate	2/20 (10)	2/20 (10)	0/20 (0)	2/20 (10)	0/20 (0)	2/20 (10)	
	Low	1/20 (5)	0/20 (0)	0/20 (0)	0/20 (0)	0/20 (0)	0/20 (0)	p=0.679
Language	Good	17/20 (85)	19/20 (95)	20/20 (100)	18/20 (90)	20/20 (100)	15/20 (75)	
	Moderate	0/20	1/20 (5)	0/20 (0)	2/20 (10)	0/20 (0)	5/20 (25)	
	Low	3/20 (15)	0/20 (0)	0/20 (0)	0/20 (0)	0/20 (0)	0/20 (0)	p=0.542
Accuracy of reference	Yes	14/143 (9.8)	7/114 (6.1)	4/73 (5.5)	25/188 (13.3)	11/190 (5.8)	10/164 (6.1)	
	No	118/143 (82.5)	101/114 (88.6)	63/73 (86.3)	124/188 (66.0)	179/190 (94.2)	132/164 (80.5)	
	Partial	11/143 (7.7)	6/114 (5.3)	6/73 (8.2)	39/188 (20.7)	0/190 (0)	22/164 (13.4)	p<0.001

The secondary outcome measured was the ability of existing models to distinguish between human-written text or AI-generated responses. The closer to 1 the Frac(p) is, the more likely that a specific text or response is worded in a way that the model expects AI to generate a response. If the model is less confident, frac(p) will be closer to 0. When analysing the responses with the GLTR model, it was found that the overall Frac(p) was 0.993, which means that most responses were likely to be AI-generated. There was no significant difference between the specialties (p=0.313) and the categories of questions (p=0.831) but there was a statistical difference noted between the subspecialty responses (p<0.005). In this case, the lowest frac(p) value was 0.961 (Upper GI surgery), which is still almost 1.

Table 6. Mean Frac(p)

	Frac(p)	**p-value**
Overall	0.993686	
General Surgery	0.991858	
Urology	0.997673	P=0.313
Type of question		
Surgical Techniques	0.997156	
Pathophysiology	0.991778	
Epidemiology	0.99495	P=0.831

Clinical Management	0.991	
Subspecialty		
Acute Care Surgery	1	
Colorectal	0.9899	
Endocrine	1	
Trauma	1	
Upper GI	0.96125	
Vascular	1	
Uro-oncology	1	
Uro-Endourology	1	
Uro-Functional	0.9936	P<0.005

Discussion

Principal Results

It is evident here that ChatGPT has excellent generative capabilities. The AI model is able to provide responses with broad and relevant vocabulary, and uses appropriate language for academic writing. Overall, the content of the answers is assessed to be accurate, with 86.7% of responses deemed acceptable by clinicians. This is testament to the development of LLMs with excellent natural processing abilities. However, this study aims to look at the ability of AI to provide responses suitable for submission to a medical journal, with the content supported by references. The results show that the references are all of very poor quality, quoting articles that are completely fabricated with only 5.5% of all responses having references that were authentic. During the review of the references, it was observed that many of the articles quoted have realistic sounding titles, and were quoted from existing journal articles. Furthermore, the authors quoted often had previously published articles on related topics. This presents a significant challenge for ChatGPT users, as it can be difficult to determine if a quoted reference is fabricated or not. A less discerning user may be fooled by these references, as the information provided in the responses are generally accurate. This greatly reduces its usefulness in the context of academic writing.

The astonishingly high proportion of fabricated references is likely a result of the inherent nature of generative AI - the generation of human-like responses based on analysis of word sequences, context, and its understanding of language patterns. Generative AI like ChatGPT are trained language models to mimic human writing and speech, without the interpretation or understanding of the content analysed[15]. Fabricated references sound convincing - with reasonable titles and authors - which is exactly what ChatGPT is designed to do. ChatGPT-v3.5 appears not to have been pre-trained in the aspect of accurate reference citation.

Although there was a difference noticed between the information accuracy of GS (90%) and Urology (79.6%) responses, the accuracy of the responses was still high. These differences were seen again in further sub-group analyses between the various subspecialties. As these were completed by single clinicians in their role as subject matter experts, it is possible that the differences may be related to the tolerances of what a clinician may deem to be ‘acceptable’. On the other hand, the high proportion of fabricated references, together with the differences in accuracy of information both in GS, Urology, and its subsets, may be explained by the datasets used to train ChatGPT. ChatGPT utilises an immense plethora of information including Common Crawl (a collection of text from billions of web texts), published books, scientific journals and even Wikipedia[5]. Its ability to differentiate and utilise relevant, scientific, and academically acclaimed research from non-academic data may be questionable, given its inability to cite accurate references. These differences may be exaggerated in subspecialty and niche topics (such as surgical techniques) where professional and academic sources may be scarce, in which case non-academic resources may be the main source of its responses. In these cases, the use of search engines that provide multiple resources, such as Google, may be superior compared to generative AI that combines its analysis of academic and non-academic data to generate its response[16].

When comparing the accuracy of references generated by ChatGPT between the various specialties, it was noted that there were some statistically significant differences. It is telling, however, when we look at the raw data and the percentages, that practically, the statistical significance is likely due to the exceedingly small numbers of accurate references.

Although the breadth and language of the responses were considered ‘Good’ when analysed by human assessors, the GLTR model was still able to detect that the text was AI-generated, with the overall Frac(p) at 0.994. While ChatGPT has been demonstrated to bypass traditional plagiarism models and even human medical researchers, the GLTR model uses a probability model to evaluate if a text is likely to be AI-generated by comparing the text with existing models and determining the probability of each word that was selected to be part of the sentence. The higher the proportion of high probability words highlighted in each text, the more likely a text is labelled as AI generated[14]. These tools remain an important part of attributing credit to

the correct authors and ensuring that academic writing remains the domain of human authors. Nonetheless, there are plenty of scenarios where generative AI can be useful in academic writing, such as guiding new university students in writing academic papers or helping those whose first language is not English to navigate the challenges of writing in a foreign tongue[2,17]. Generative AI has been shown to produce eloquent writing, aid in streamlining searches for research articles, and even act as a brainstorming intermediary for topic and research question selection[18, 19, 20].

It is notable, however, that GLTR was developed with access to GPT-2 language model by OpenAI, and thus likely already recognises the text generation patterns of ChatGPT as "AI generated". Separate studies show that GLTR can detect AI generated texts with the accuracy of just 72%, although still considerably better compared to an abysmal 54% by human counterparts[14].

Limitations

One of the limitations of this study is that only GS and Urology were assessed. While results for both were similar, further investigation in other specialities, such as medical specialties or other surgical specialties like orthopaedics or otorhinolaryngology, may elucidate some differences. There was also only a single clinician that provided the assessment of these responses. Recruitment of more reviewers may magnify or reduce differences seen. Finally, the ChatGPT version used was version 3.5, which, with the release of ChatGPT v4.0 by OpenAI, is no longer the latest version of ChatGPT. While ChatGPT v4.0 utilises an equivalent size of data, the up-scaling of its deep learning has allowed it to demonstrate human levels of academic excellence. For example, ChatGPT v4.0 has demonstrated its ability to score around the top 10% of a stimulated bar exam and perform near or at the passing threshold for the US Medical Licensing Examination[21, 22]. However, ChatGPT v4.0 is still not readily available for free to much of the public, and requires a subscription, limiting its market penetration. This study also did not include the analysis of other NLP models, such as Bidirectional Encoder Representation from Transformers (BERT), and its biomedical counterparts BioBERT and PubMedBERT[23]. Additionally, the emergence of BioGPT, a domain-specific transformer developed by Ruo et al. pre-trained on 15 million PubMed texts, shows promise in generating fluent biomedical text superior to its BERT counterparts that tended to excel in biomedical text mining[24].

Conclusion

While the information provided in ChatGPT is accurate with sufficient breadth and with appropriate medical academic language, its quotation of fictitious references limits its use for academic writing. This study conclusively shows that generative AI, in the form of ChatGPT, is not sufficiently sophisticated to produce scientific writing with references of adequate quality. With more than 90% of responses containing invented yet convincing references, users must exercise caution and discrimination in utilising ChatGPT in scientific writing. This may be attributed to its inherent nature as a language generating model, and its vast dataset that includes a substantial proportion of non-academic resources. The emergence of newer versions and models that pre-trains with specific biomedical texts and up-scaling of deep learning may eventually result in a sophisticated model with a complex neural network and a rich scientific dataset capable of assisting in accurate and citable medical academic writing[23]. Until then, we must take great caution in relying on generative AI for medical academic writing, for they lack the specialised understanding and knowledge required[25].

Acknowledgments

Nathaniel HE HEAH: writing of manuscript and statistical analyses
Wei Tseng TZEN: editing of manuscript
Elyn YL TZEN: editing of manuscript
Kon Voy TAY, Zhiwen Joseph LO, Cheuk Fan SHUM, Marcus WL CHOW, Leynard M MARRERO, Phyo AUNG, Gregory KE HENG, Ishara MADUKA, Jue Fei FENG, Zhongkai WANG, Lin Yee KOONG, Serene EL TANG, Cathy PC NG, Shane WH SIM, Man Hon TANG, Pravin LINGAM, Shaun WY CHAN, Yuan Teng CHO, Chee Wei LEE, Sadhana CHANDRASEKAR, Wei Chong CHUA: subject matter experts in General Surgery and Urology, assessment of ChatGPT responses

Abbreviations

ACS: acute care surgery
AI: artificial intelligence
BERT: Bidirectional Encoder Representation from Transformers
GLTR: Giant Language model Test Room
GPT: Generated Pre-trained Transformer
GS: general surgery
IF: impact factor
JAMA: The Journal of the American Medical Association
LLM: large language model
NLP: Natural Language Processing
PPO: proximal policy optimisation
RLHF: reinforced learning from human feedback
GI: upper gastrointestinal surgery

Conflicts of Interest

None declared.

References

1. Parisis N. Medical writing in the era of artificial intelligence. *Med Writ*. 2019;28(4):4-9.
2. Schmohl T, Watanabe A, Fröhlich N, Herzberg D. How Artificial Intelligence Can Improve the Academic Writing of Students. *Conf Proceedings Futur Educ 2020*. June 2020. https://conference.pixel-online.net/library_scheda.php?id_abs=4769. Accessed January 24, 2024.
3. Malik AR, Pratiwi Y, Andajani K, et al. Exploring Artificial Intelligence in Academic Essay: Higher Education Student's Perspective. *Int J Educ Res Open*. 2023;5:100296. doi:10.1016/J.IJEDRO.2023.100296
4. Biswas S. ChatGPT and the Future of Medical Writing. *https://doi.org/101148/radiol223312*. 2023;307(2). doi:10.1148/RADIOL.223312
5. Cheng Y. Study of ChatGPT and its Comparison with Other Mainstream Large LanguageModels. International Journal of Knowledge and Language Processing [Internet]. 2022 Dec;13:1–12. Available from: https://www.google.com/search?q=Study+of+ChatGPT+and+its+Comparison+with+Other+Mainstream+Large+LanguageModels&oq=Study+of+ChatGPT+and+its+Comparison+with+Other+Mainstream+Large+LanguageModels&gs_lcrp=EgZjaHJvbWUyBggAEEUYOdIBBzY2M2owajGoAgCwAgA&sourceid=chrome&ie=UTF-8
6. Wu T, He S, Liu J, Sun S, Liu K, Han QL, et al. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. IEEE/CAA Journal of Automatica Sinica [Internet]. 2023 May;10(5):1122–36. Available from: https://ieeexplore.ieee.org/abstract/document/10113601
7. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 2022 Jan 31;3(1):1–23.
8. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *https://doi.org/101148/radiol230424*. April 2023. doi:10.1148/RADIOL.230424
9. Kacena MA, Lilian ·, Plotkin I, Fehrenbacher JC. The Use of Artificial Intelligence in Writing Scientific Review Articles. *Curr Osteoporos Reports 2024*. January 2024:1-7. doi:10.1007/S11914-023-00852-0
10. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiol*. 2023. doi:10.1007/S00256-023-04340-5
11. Heikkilä M. How to spot AI-generated text | MIT Technology Review. MIT Technology Review. https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/. Published 2022. Accessed January 24, 2024.
12. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Heal*. 2023;5(3):e105-e106. doi:10.1016/S2589-7500(23)00019-5
13. Zheng H, Zhan H. ChatGPT in Scientific Writing: A Cautionary Tale. *Am J Med*. 2023. doi:10.1016/j.amjmed.2023.02.011
14. Gehrmann S, Strobelt H, Rush AM. GLTR: Statistical detection and visualization of generated text. *ACL 2019 - 57th Annu Meet Assoc Comput Linguist Proc Syst Demonstr*. 2019:111-116. doi:10.18653/V1/P19-3019
15. Abdullah M, Madain A, Jararweh Y. ChatGPT: Fundamentals, Applications and Social Impacts. 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2022 Nov 29;1–8.
16. Shen OY, Pratap JS, Li X, Chen NC, Bhashyam AR. How Does ChatGPT Use Source Information Compared With Google? A Text Network Analysis of Online Health Information. Clinical Orthopaedics and Related Research. 2024 Mar 1;482(4):578–88.
17. Zulfa S, Sari Dewi R, Nuruddin Hidayat D, Hamid F, Defianty M. The Use of AI and Technology Tools in Developing Students' English Academic Writing Skills. *Annu Int Conf Educ*. 2023;1:47-63.

https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/proceedings/article/view/1811/624. Accessed January 24, 2024.
18. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croatian Medical Journal. 2023 Feb;64(1):1–3.
19. Else H. Abstracts written by ChatGPT fool scientists. Nature. 2023 Jan 12;613(7944).
20. Gordijn B, Have H ten. ChatGPT: Evolution or revolution? Medicine, Health Care and Philosophy. 2023 Jan 19;26(1).
21. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Medical Education [Internet]. 2023 Feb 8;9(9):e45312. Available from: https://mededu.jmir.org/2023/1/e45312/PDF
22. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation. 2023 Apr;185:109732.
23. Sharaf S, S AV. An Analysis on Large Language Models in Healthcare: A Case Study of BioBERT [Internet]. arXiv.org. 2023 [cited 2024 Oct 22]. Available from: https://arxiv.org/abs/2310.07282
24. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics. 2022 Sep 24;23(6).
25. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence [Internet]. 2023 May 4;6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10192861/

Supplementary Files

Multimedia Appendixes

General Surgery and Urology questions posed to ChatGPT and answers with corresponding references by ChatGPT.
URL: http://asset.jmir.pub/assets/3301f6d5e7dff481a7223a2e89c8e6fc.doc