RESEARCH ARTICLE

From Algorithms to Academia: An Endeavor to Benchmark AI-Generated Scientific Papers against Human Standards

Jackson Woodrow, BS; Nour Nassour, MD; John Y. Kwon, MD; Soheil Ashkani-Esfahani, MD; Mitchel Harris, MD

Research performed at Foot & Ankle Research and Innovation Laboratory (FARIL), Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Received: 17 June 2024 *Accepted: 18 September 2024*

Abstract

Objectives: The aim of this study is to quantitatively investigate the accuracy of text generated by AI large language models while comparing their readability and likelihood of being accepted to a scientific compared to human-authored papers on the same topics.

Methods: The study consisted of two papers written by ChatGPT, two papers written by Assistant by scite, and two papers written by humans. A total of six independent reviewers were blinded to the authorship of each paper and assigned a grade to each subsection on a scale of 1 to 4. Additionally, each reviewer was asked to guess if the paper was written by a human or AI and explain their reasoning. The study authors also graded each AI-generated paper based on factual accuracy of the claims and citations.

Results: The human-written calcaneus fracture paper received the highest score of a 3.70/4, followed by Assistant-written calcaneus fracture paper (3.02/4), human-written ankle osteoarthritis paper (2.98/4), ChatGPT calcaneus fracture (2.89/4), ChatGPT Ankle Osteoarthritis (2.87/4), and Assistant Ankle Osteoarthritis (2.78/4). The human calcaneus fracture paper received a statistically significant higher rating than the ChatGPT calcaneus fracture paper ($P = 0.028$) and the Assistant calcaneus fracture paper ($P = 0.043$). The ChatGPT osteoarthritis review showed 100% factual accuracy, the ChatGPT calcaneus fracture review was 97.46% factually accurate, the Assistant calcaneus fracture was 95.56% accurate, and the Assistant ankle osteoarthritis was 94.98% accurate. Regarding citations, the ChatGPT ankle osteoarthritis paper was 90% accurate, the ChatGPT calcaneus fracture was 69.23% accurate, the Assistant ankle osteoarthritis was 35.14% accurate, and the Assistant calcaneus fracture was 39.68% accurate.

Conclusion: Through this paper we emphasize that while AI holds the promise of enhancing knowledge sharing, it must be used responsibly and in conjunction with comprehensive fact-checking procedures to maintain the integrity of the scientific discourse.

Level of evidence: III

Keywords: Artificial intelligence, ChatGPT, Large language models, Natural language processing, Prompt engineering

Introduction

The advent of Artificial Intelligence (AI) has brought about a paradigm shift in the way large data sets are processed in the fields of finance, transportation, drug development, and computer science.[1] However, with the development of large language models (LLMs), the underlying AI models now have the ability to read and generate human language in a way that is sometimes indistinguishable from human-produced text.[2] By far, the most popular LLM currently on the market is the third-

Corresponding Author: Jackson Woodrow, Foot & Ankle Research and Innovation Lab (FARIL), Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Email: jwoodrow@mgh.harvard.edu

THE ONLINE VERSION OF THIS ARTICLE ABJS.MUMS.AC.IR

generation pre-trained transformer (ChatGPT-3, Chat Generative Pre-trained Transformer, OpenAI Limited Partnership, San Francisco, CA, USA) model, which was released on November 30, 2022 by OpenAI and quickly reached over 100 million active monthly users in just 2 months.[3] This was followed shortly by the release of GPT-4 on March 13, 2023, which is a paid version of the free ChatGPT-3 that allows for faster response generation, image processing, expanded response word counts, and increased accuracy.[4]

In the field of orthopaedic surgery, AI is starting to be implemented in both academic and clinical medicine. AI models are beginning to read radiographs to predict treatment outcomes for patients with knee osteoarthritis as well as for preoperative planning to assist surgeons in selecting the proper hardware for spinal deformity correction procedures.[5,6] Aside from these advancements with patient care, one area of great concern within the academic community has been the potential abuse of AI through plagiarism.[7] Many AI plagiarism detectors have since been developed, but a study by Gao *et al.* Found that only 68% of ChatGPT generated abstracts were detected as AI-generated and 14% of human-written abstracts were incorrectly identified as being written by AI.[8] However, with the rapidly evolving AI advancements, there is no promise that these software programs will continue to be effective in the future.

Beyond plagiarism, there is also the risk of factual inaccuracies, or "*hallucinations*" produced by LLMs.[9,10] ChatGPT is extremely well-versed at producing text, but it has no intrinsic ability to discern what is true and what is false. OpenAI is aware of this occurrence, and according to them, "there have been instances where advanced AI systems, such as generative models, have been found to produce *hallucinations*, particularly when trained on large amounts of unsupervised data".[9] At the time of this study, ChatGPT was trained on a data set that is current through September 2021, although it has since been updated to more a more current date.[11] In the field of science where discoveries are continuously made, amended, and refuted, ChatGPT may not always provide the best nor most up-to-date factual evidence.[11] A study by Alkaissi and McFarlane demonstrated the *hallucination* phenomenon when they asked ChatGPT to write a short essay on the liver involvement in late-onset Pompe disease (LOPD).[9] ChatGPT followed instructions and provided background about the liver's role in LOPD, although this has never been described in the literature and has only been reported in the infantile form. This report showed that ChatGPT can be a very helpful tool in producing research papers, but it cannot yet replace an expert reviewer who strictly adheres to the scientific process. The aim of this study is to quantitatively investigate the accuracy of the text generated by LLMs as well as compare their readability and likelihood of being accepted to a scientific journal compared to human-authored papers on the same topics.

Materials and Methods

This study employed a double-blind review process to evaluate the quality and discernibility of AI-generated scientific papers to those written and published by human authors. The study consisted of two papers written by ChatGPT-4® with the ScholarAI® plugin, two papers written by Assistant by scite® (scite, scite.ai, Brooklyn, NY, USA), and two papers written by humans that were published in peer-reviewed scientific journals.

Paper Selection and Generation

Human-authored review papers on management of osteoarthritis of the ankle and management of calcaneus fractures were selected from peer-reviewed journals. The human papers consisted of "Management of displaced intra-articular calcaneal fractures; current concept review and treatment algorithm" by Salameh *et al.*[12] and "Ankle osteoarthritis: comprehensive review and treatment algorithm proposal" by Herrera-Pérez *et al.*[13] The ankle osteoarthritis paper was chosen for a more broad topic in the field of foot and ankle surgery while the intra-articular calcaneal fracture paper was chosen as a more specialized discipline within foot and ankle surgery to assess the capability of the LLMs to generate text on a more niche subject. The text was copied into a word processing program with only the formatting being changed so all of the papers would have similar structure as as part of the blinding process. Both human papers were published after September 2021, which is the training database cutoff for ChatGPT, thus theoretically limiting AI's ability to reference our control papers.

AI Paper Generation

Identical prompts were given to both ChatGPT and Assistant by scite.ai. The prompts were written in an attempt to balance clear directions for text generation while also not being overly specific and hindering the creativity of the AI software. In order to increase standardization, the AI prompts were based on the structure of the human-authored review papers. All prompts are listed in the [Appendix Figures 1A and 2A]. After generation of a response to each prompt, the text was copied and pasted into a document to generate the manuscript. The content and references of the manuscript were unchanged while only the formatting was modified.

Reviewers

A total of six independent reviewers participated in the process. Three of the reviewers were orthopaedic surgery attending physicians while the other three were orthopaedic foot and ankle surgery fellows. All orthopaedic surgery attendings previously served as a peer-reviewed for an orthopaedic scientific peer-reviewed journal and each had at least 7 years of experience practicing as a board-certified orthopaedic surgeon.

Review Process

Each reviewer was tasked with evaluating each paper in its entirety, as well as each subsection individually. Reviewers were blinded to the authorship of each paper and were excluded if they believed they had read or discussed the paper previously. Reviewers assigned a grade to each subsection on a scale of 1 to 4 with 4 representing accept without revisions, 3 representing accept with minor revisions, 2 representing accept with major revisions, or 1 repressing reject. Additionally, each reviewer was asked to guess if the paper was written by a human or AI and explain their reasoning. Reviewers did not know how many of the

papers were AI generated or which AI algorithms were employed. Responses were only identified by reviewer level of training (i.e. attending vs fellow).

The study authors also graded each AI-generated paper based on factual accuracy of the claims and citations. The study authors were initially blinded to the authorship of the paper and individually graded each sentence and citation as either true or false. Veracity of sentences were verified by referencing previously peer-reviewed literature and UpToDate (Wolters Kluwer, Waltham, MA). Scores for each sentence and citation were compared and any discrepancies were discussed until there was a consensus. For a citation to be considered accurate it must 1) reference a previously published journal paper and 2) contain factually correct information mentioned in said paper. Claims requiring but missing a citation were graded as incorrect.

Data Analysis

To calculate the overall score of a paper, the grades of each subsection were averaged to determine the paper's overall score. This was done for each reviewer. Then, the reviewers' scores were averaged to give a final paper score. This two-tiered averaging process ensured that both the detail of each subsection and the consensus among reviewers contributed to the final score. The difference of mean scores for all of the papers was analyzed using a Wilcoxon Signed-Rank Test with a significance level of $p < 0.05$. The mean scores for each paper were also broken down by if the reviewer was an attending or fellow. The accuracy of the reviewers' guesses about the authorship were also calculated. Narrative feedback about each paper was also collected and analyzed thematically to identify common patterns, observations, and trends within the writing.

Results

Reviewers' Scores

The human-written calcaneus fracture paper received the highest score of a 3.70 out of a possible 4, followed by Assistant by scite-written calcaneus fracture paper with a 3.02 out of a possible 4. Next was the human-written ankle osteoarthritis paper (2.98/4), followed by the ChatGPT calcaneus fracture (2.89/4), ChatGPT Ankle Osteoarthritis (2.87/4), and Assistant by scite Ankle Osteoarthritis (2.78/4) [Table 1]. When compared to papers written on the same topic, the human calcaneus fracture paper received a statistically significant higher rating than the ChatGPT calcaneus fracture paper ($P = 0.028$) and the Assistant by scite calcaneus fracture paper ($P = 0.043$) [Table 1]. The ChatGPT and Assistant by scite calcaneus fracture papers showed no significant difference in their overall ratings ($P = 0.138$) [Table 1]. For the ankle osteoarthritis papers, the human paper received the highest average rating but did not differ significantly from the ChatGPT ankle osteoarthritis paper ($P = 0.917$) or the Assistant by scite ankle osteoarthritis paper ($P = 0.753$). The Assistant by scite and ChatGPT ankle osteoarthritis papers also showed no significant difference ($P = 0.916$). Orthopedic foot and ankle fellows tended to rate papers higher compared to orthopedic attendings, however this difference was not statistically significant [Table 1].

Table 1. Average scores given by the reviewers to each paper. Average scores for each paper consisted of three orthopedic attendings and three orthopedic foot and ankle fellows

Paper	Average Score /4
Human Ankle Osteoarthritis	
Overall	2.98
Attending	2.81
Fellow	3.14
ChatGPT Ankle Osteoarthritis	
Overall	2.87
Attending	2.73
Fellow	3.00
Assistant by scite Ankle Osteoarthritis	
Overall	2.78
Attending	2.77
Fellow	2.80
Human Calcaneus Fracture	
Overall	3.70*
Attending	3.48
Fellow	3.91
ChatGPT Calcaneus Fracture	
Overall	2.89
Attending	2.58
Fellow	3.21

Table 1. Continued	
Assistant by scite Calcaneus Fracture	
Overall	3.02
Attending	2.58
Fellow	3.45

*The human-authored calcaneus fracture paper was rated statistically significantly higher than both the ChatGPT calcaneus fracture (p=0.028) and Assistant by scite calcaneus fracture papers (p=0.043). There was no significant difference between the ankle osteoarthritis papers and between the ChatGPT calcaneus fracture and Assistant by scite calcaneus fracture papers. There was also no statistically significant difference in ratings between attendings and fellows

Sentence and Citation Accuracy

All AI-generated papers displayed an incredibly high level of factual accuracy but struggled with citation accuracy. The ChatGPT ankle osteoarthritis showed no factual inaccuracies while the ChatGPT calcaneus fracture review was 97.46% factually accurate, the Assistant by scite calcaneus fracture was 95.56% accurate, and the Assistant by scite ankle osteoarthritis was 94.98% accurate [Table 2]. Regarding citations, the ChatGPT ankle osteoarthritis paper was 90% accurate which included 4 statements missing citations [Table 2]. The ChatGPT calcaneus fracture had 69.23% correct citations which included 8 statements missing citations [Table 2]. The Assistant by scite ankle osteoarthritis had 35.14% correct citations which included 26 statements missing citations, and the Assistant by scite calcaneus fracture had 39.68% correct citations which included 12 statements missing citations [Table 2].

Table 2. Sentence and citation accuracy for AI-generated papers

Paper	Sentence Accuracy (%)	Citation Accuracy (%)	Number of Incorrect Citations	Number of Missing Citations
Assistant by scite Osteoarthritis	94.98	35.14	24	26
ChatGPT Osteoarthritis	100	90	3	4
Assistant by scite Calcaneus Fracture	95.56	39.68	38	12
ChatGPT Calcaneus Fracture	97.46	69.23	8	8

Authorship Guess

Of the papers written by an AI software, the reviewers correctly guessed it was written by AI 70.83% of the time while the reviewers correctly guessed the authorship of human generated papers 83.3% of the time [Figure 1]. The attendings were more likely to correctly identify an AI-authored paper (75% accuracy) while fellows were more likely to correctly identify a human-authored paper (100% accuracy).

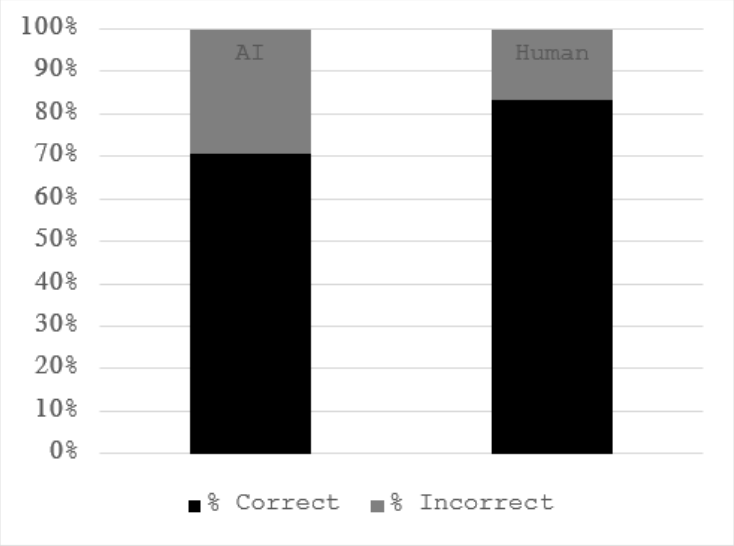

Figure 1. Reviewers' guesses of paper authorship for human vs AI. Reviewers were slightly better at correctly guessing human-authored papers (83.3%) compared to AI-authored papers (70.8%)

Narrative Comments

Some of the common narrative comments about the AI-generated papers were that they contained repetitive language, awkward sentence structure, and tended to include unrelated topics [Table 3]. Even if a topic was unrelated, none of the reviewers, however, commented on any noticeable hallucinations or factual inaccuracies. However, for this project, reviewers were not asked to comment on citation accuracy since that was calculated by the study authors. The coherence also tended to be rated as stronger for the ChatGPT produced papers compared to the Assistant by scite papers.

Table 3. Narrative comments given by the reviewers to each paper. Authors were asked to give narrative comments about their overall impression of the paper as well as guess the authorship of the paper

Paper	Narrative Comments	Authorship Guess
Human Ankle Osteoarthritis	Abstract written as bullet points which is unusual, several run-on sentences, at times poor sentence structure, use of unusual non-scientific words such as "peculiarities", certain subsections completely cursory	*AI*
Human Ankle Osteoarthritis	This study was straight to the point, involving specific focus that a foot and ankle surgeon can benefit from	*Human*
Human Ankle Osteoarthritis	The paper is well organized and has a good flow in terms of reading. Information is well organized, and it is not repetitive. I would suggest that the authors move the are discussing radiologic changes in different modalities to a new section for radiographic changes and not keep it on the clinical evaluation section. I also suggest that the section of treatment for non-preserving joint procedures is subdivided in joint movement preserving procedures and joint movement sacrificing procedures	*Human*
Human Ankle Osteoarthritis	This paper does a good job in providing a basic overview of ankle arthritis and management for someone who does not have a strong background in the field. Some sections of the paper are better than others. There paper delves into great depth regarding surgical management, but does not provide enough detail under the intra-articular therapies. The bullet-points listed in the abstract does not flow well with the remainder of the paper. There is also too much factual information in the abstract	*Human*
Human Ankle Osteoarthritis	The paper discusses in depth with supporting literature about causes, clinical evaluation and various treatment options. The paper is easy to follow and has a very high readability. However, there are few statements treamtent options discussed without supporting literature. e.g. Reducing levels of blood cholesterol and increasing the intake of foods rich in vitamin K, which plays an important role in the mineralization of bones and cartilage, are also beneficial for OA." In addition, it should be highlighted for the readers that there is level 4/5 studies are described in literature for some of the joint preserving osteotomies. Moreover, the total talus replacement in isolation is not indicated in isolation for advanced ankle OA. Finally, i would recommend discussing outcomes of recent studies comparing arthroscopic ankle fusion vs. open ankle fusion vs total ankle replacement	*Human*
Human Ankle Osteoarthritis	A very comprehensive review of ankle arthritis and management options with good references backing up content. However, in the introduction the authors stated the onset of postraumatic ankle arthritis ranges from 18-44. I question this and the study cited did not back this up. Overall this paper had good flow and told a story in a cohesive manner	*Human*
ChatGPT Ankle Osteoarthritis	Better flow than some other papers, some unusual text such as layman's term for describing OA, section on injections incomplete (no discussion of cortisone)	*AI*
ChatGPT Ankle Osteoarthritis	Coherence, fluency is really good, easy to read but content is vague. It is not adding much to the common knowledge. "A study by Sun et al. found that HA injections reduced pain and improved function in patients with ankle OA2. However, the authors noted that more high-quality studies are needed to confirm these findings" this is another thing, nearly all articles end with this statement	*AI*
ChatGPT Ankle Osteoarthritis	The review is comprehensive and complete developing the sections in an appropriate manner. However, the discussion requires better flow. Some of the closing statements of the sections need to be less repetitive. The listing of some of the findings of the literature can be summarized in a more effective manner	*Human*

Table 3. Continued

ChatGPT Ankle Osteoarthritis	The paper briefly discusses the treatment options without diving into details. In addition, there is no comparison between various measurements strategies. Finally, intra-articular therapies are very poorly discussed.	*AI*
ChatGPT Ankle Osteoarthritis	Comprehensive, but concise review on the topic. There are more options than just Hyaluronic acid for intra articular injection; authors should have discussed the other common options such as PRP and cortisone	*AI*
ChatGPT Ankle Osteoarthritis	This article is a basic-level review. It is almost like reading a summary of facts. Even the conclusion is a summary of the statements previously mentioned. The section on intra-articular injections states that hyaluronic acid injections are commonly used for ankle arthritis when this is not true. This sections also fails to mention corticosteroid injections	*AI*
Assistant by scite Ankle Osteoarthritis	Well written and organized, succinct, good transitions, good summary of various topics, writing style is coherent with great readability	*Human*
Assistant by scite Ankle Osteoarthritis	While the language in this scientific paper is well-written, it appears that the content lacks specificity for the intended audience, namely foot and ankle surgeons. The flow of ideas is somewhat overwhelming, and the material could benefit from a more focused and detailed approach to make it more relevant and useful for our target readership	*Human*
Assistant by scite Ankle Osteoarthritis	This manuscript is a review of the epidemiology, pathophysiology and treatment of OA of the ankle joint. Despite being well organize, the flow of the manuscript is not appropriate with multiple areas of repeated content. The writing style and structure are repetitive which makes the paper hard to read and follow	*AI*
Assistant by scite Ankle Osteoarthritis	This is an overall well-written article that is concise and factual. It would be great for a review article on ankle OA for JAAOS. There are some weaknesses; table 1 description states that the classification involves distinguishing joint sparing and joint sacrificing procedures, when there is no mention of that in the table. The writing style	*AI*
Assistant by scite Ankle Osteoarthritis	Paper is written in a generic language without highlighting the clinical and surgical options in detail including indications and contraindications for each treatment options as well as their short term and long term which have previously been published in literature. In addition, conclusion is too lengthy. Overall this article does not add or summarize the existing literature and will have poor readability	*AI*
Assistant by scite Ankle Osteoarthritis	This paper offers a comprehensive review of ankle arthritis and treatment options available. The authors' description of potential genetic bases for ankle osteoarthritis did not seem rooted in evidence that directly linked these genes to ankle osteoarthritis. There was evidence to support conservative management modalities described. This paper was overall coherent and had good readability	*AI*
Human Intra-articular Calcaneus Fracture	Well organized, well written, good transition sentences, good detail and thoroughness	*Human*
Human Intra-articular Calcaneus Fracture	Paper is presented well with fluency and straight to the point approach	*Human*
Human Intra-articular Calcaneus Fracture	Whe paper is a very good summary of the diagnosis, treatment and prognosis of displaced and intra-articular fractures of the calcaenous. The paper it well written and comprehensive. However, there are some areas where the flow of the manuscript can be improved and the information discussed less repetitive. I recommended that the term calcareous it's used instead of calcaneum. There ares some areas where the term "in conclusion" is used in a receptive manner at the conclusion of sections. I suggest using diverse concluding sentences such as "in summary", " to conclude" etc	*AI*
Human Intra-articular Calcaneus Fracture	Overall, well-written paper. Great, easy to read style with a lot of pertinent information. Great flow to the articles	*Human*
Human Intra-articular Calcaneus Fracture	The article discusses all treatment options including surgical approaches and implant choices in detail. The article is well written and has high readability	*Human*

Table 3. Continued

Human Intra-articular Calcaneus Fracture	Very well written paper with great flow with appropriate citations to back up content. My one critic would be the authors provide evidence that suggests the sinus tarsi approach may be superior to extensile lateral approach, however, they seem to move away from this in their conclusion	*Human*
ChatGPT Intra-articular Calcaneus Fracture	Abstract poorly organized, introduction section has poor flow, classifications section incomplete and inadequate, certain subsections have text not relevant to that section	*AI*
ChatGPT Intra-articular Calcaneus Fracture	Nice explanations but again they are not specific mostly indicating common knowledge	*AI*
ChatGPT Intra-articular Calcaneus Fracture	The manuscript is well written and comprehensive with updated and well discussed literature related to treatment, diagnosis and complications. However, the flow of the manuscript needs to be improved along with reduction of repetitive closing lines and each section such as "in conclusion". recommend the use of alternative sentences such as "in summary" or "to conclude"	*AI*
ChatGPT Intra-articular Calcaneus Fracture	The article is well written and the content is appropriate. The readability is high. However, there is a line "A study by Boutsiadis et al. suggested that the best implant choice for coracoid graft fixation during the Latarjet procedure depends on patients' morphometric considerations.22" should be removed	*Human*
ChatGPT Intra-articular Calcaneus Fracture	Good paper and review, however, section on primary fusion was more about implant choice. Not much in evidence presented on outcomes of primary fusion as a surgical option	*Human*
ChatGPT Intra-articular Calcaneus Fracture	The article is over simplified. It does not go into the expected depth of an article on comprehensive management of calcaneus fractures. The last paragraph in the abstract section only consists of one section. The introduction is very brief. The classification section would benefit from illustrations/images. There are multiple areas throughout the paper where paragraphs consists of only 1 or 2 sentences. The bone graft sections mentions coracoid graft for Latarjet procedure which is not relevant to the topic	*AI*
Assistant by scite Intra-Articular Calcaneus Fracture	Awkward sentence structure at times, unusual formatting, unusual inclusion of certain topics (for ex. ethical considerations), certain subsections written in a very cursory manner	*AI*
Assistant by scite Intra-Articular Calcaneus Fracture	This paper was hard to understand its origin. But it had some general information that would not be included in human-originated reviews. Also There were so many bullet point explanations which is commonly preferred method for AI	*AI*
Assistant by scite Intra-Articular Calcaneus Fracture	The paper is a comprehensive review of fractures of the calcaneus. However, many sections are not well developed and limit the discussion to a simple outline of the problem/controversy (surgical techniques, use of allograft, type of implant. They authors should remove the section about ethics in the abstract. Please develop sections such as classification systems. Literature needs to be discussed in more detail	*AI*
Assistant by scite Intra-Articular Calcaneus Fracture	The article is well written but there are limited studies that have been discussed in detail. The overall readability is high	*AI*
Assistant by scite Intra-Articular Calcaneus Fracture	Good overview of the topic. Authors did not follow through to address ethical considerations which was one of the goals of the paper. The authors briefly mentioned dual incision approach an option in operative fixation, but provided no evidence on the efficacy of this approach. Primary fusion section seemed disjointed with loss of subtalar motion and adjacent joint degeneration both listed as potential advantages of primary fusion	*Human*
Assistant by scite Intra-Articular Calcaneus Fracture	The paper does a good job in providing a comprehensive review. The section titled "highlights" could be omitted from the abstract. Reference numbers 8 and 9 do not seem to be appropriate for the topic. The Introduction could benefit from having the second paragraph as the first so there is better flow. Overall, the paper is well-written and suitable for review article in JAAOS or JB JS	*Human*

Discussion

AI's role in scientific writing has grown significantly with the introduction of advanced language generation models, becoming increasingly helpful by enhancing research writing and transforming how scholars engage, offering powerful tools that shape the way knowledge is shared and explored within academic communities. This study aims to compare peer reviewed published human-authored papers with papers entirely generated by two language generation models: ChatGPT and Assistant by scite, evaluating whether experts in the field can discern between the two categories.

The strength of this work lies in its presentation of innovation metrics to measure the quality of a paper by

using sentence and citation accuracies. This work also sheds light on a subject that holds great relevance in contemporary scientific writing. In an era where AI-based writing tools are increasingly accessible, this research provides valuable insights into the quality of such writing. It also delves into the utility, reliability, and accuracy of AI-based writing tools, while also exploring the potential for improvement in this domain.

The comparative analysis conducted among ChatGPT, Assistant by scite, and human-produced scientific content yielded insightful results in terms of sentence and citation accuracy. In fact, ChatGPT demonstrated impressive proficiency in formulating factually accurate sentences while showing mixed results on correctly referencing and utilizing citations. For ankle osteoarthritis papers, the ratings for the AI software were not statistically significant from the human-generated papers, meaning that the true authorship of the ankle osteoarthritis papers was not able to be determined. This is further supported by the finding where Assistant by scite ankle osteoarthritis paper was incorrectly labeled as human produced by two out of the six reviewers. On the contrary, for the calcaneus fracture papers, the human produced manuscript was rated statistically significantly higher than the two AI produced papers. Furthermore, five out of six reviewers correctly identified the AI-produced calcaneus fracture papers. There may be a few explanations for this finding. Osteoarthritis is a much broader subject with numerous papers written on the subject while intraarticular calcaneus fractures is a much more subspecialty based topic and may have less literature written on the subject. This may have offered the AI models more information on ankle osteoarthritis to pull from while leaving it to speak in less specific terms for the calcaneus fracture papers. Further studies on how AI deals with common versus rare topics is needed.

The quality of the text produced by ChatGPT's comes from the extensive and continuous training and validation of the AI software on diverse and granular data, including peer-reviewed scientific literature.[14] A study by Kacena *et al.* that aimed to evaluate the helpfulness of ChatGPT in writing scientific articles found that while the use of AI in scientific writing decreased the amount of time spent writing, it required extensive fact and citation checks and had a higher potential for plagiarism.[15] In fact, in our AI generated papers, there were instances where factual inaccuracies were noted; while ChatGPT seems to be adept at generating syntactically correct and contextually appropriate sentences, it seems unable to verify the authenticity of its claims.[15,16] Similarly, Assistant by scite showed high levels of factual accuracy with its generated text while having poor citation accuracy. While all of the citations referenced authentic scientific papers from peer-reviewed journals, sometimes the content was tangential at best. For instance, when discussing the initial evaluation of an ankle fracture by a medical professional, Assistant by scite listed that it is necessary to take a thorough history and identify the duration, progression, precipitating factors, and previous injuries. While this is an undisputed part of the diagnostic approach, the paper it referenced aimed to summarize the clinical findings of children with a chylothorax. It appears as if some generative algorithms search references based on a few key words without assessing the entire context of the article.

While the AI-generated manuscripts demonstrated excellent levels of factual accuracy, the citation accuracy may place the submitting author in danger of plagiarism and academic dishonesty. The two main citation errors the AI algorithms made were 1) omitting a citation when one would be required and 2) the referenced paper did not contain the claim intended to be cited. While the prior could be grounds for plagiarism, the latter could potentially be more detrimental by distributing false information while citing it as fact. In a similar paper on the evaluation of AI in scientific writing, Salvagno *et al.* discussed how AI software such a ChatGPT can be useful tools to assist researchers as they comprehend information faster than their human counterparts, it cannot generate fresh ideas, it has the ability to arrange, rephrase, and structure an author's thoughts, making it very similar to human writing.[17] There is a contention that the implementation of chatbots will enable writers to allocate more of their time to exploring other facets of their work, delving into significant areas that can yield more substantial result.[18] In general, chatbots like ChatGPT are considered valuable auxiliary tools for researchers, however, it is crucial to exercise caution and recognize that they should not be seen as replacements, nor should they be used in such a manner.[16-18] The use of such automated tools gives rise to several questions, including ethical considerations, factual accuracy, and the preservation of the "human touch."

Like most modern pieces of technology, AI is arguably a neutral tool that can be used for positive or negative purposes. Several studies discuss the ethical usability of AI chatbots, stating that while they are efficient and helpful tools, they should be used with caution not only from the context of plagiarism but also from the perspective of research ethics.[17,18] In this study, we found that generative AI programs showed impressive sentence accuracy and was able to craft an entire scientific journal paper in under 30 minutes. Researchers can leverage the speed of AI to reduce the time spent searching current literature and formulating an outline for the paper. Also, drafts can be uploaded to AI programs to check for spelling and grammar errors, sentence structure, and areas where conciseness can be improved. Although the use of AI has many positives, there are potential areas of abuse. As highlighted in our results, both ChatGPT and Assistant by scite exhibited levels of citation accuracy that would not be acceptable for any journal. Without demonstrating high levels of citation accuracy, misinformation could be introduced into the literature, underscoring the need for a stringent peer review process. The use of AI in scientific writing also brings up complex ethical challenges, particularly regarding authorship and accountability. The type of paper for this study was deliberately chosen to be a review paper to test the AI's ability to reference previous work without having to generate any original data. However, to advance the scientific field forward, research studies that generate original data have to be performed. While ChatGPT and other AI chat programs may be able to generate review articles and reference previous work, its implementation into the broader space of scientific writing is still limited and is not a substitute for human authorship in all areas.

Limitation

While this study is currently the only one to our knowledge that assessed the accuracy of both sentences and citations as well as subjective scores from journal reviewers, it does have limitations. First, the text produced from AI programs is highly variable. The responses are highly dependent on the specificity given in the prompt, and furthermore, one prompt will yield a variety of different outputs. While we tried to keep our process as standardized as possible, there is no guarantee that the results will be reproducible. By definition, the AI algorithms use machine learning to continuously revise and update how they respond to prompts, which further decreases the reproducibility of this study. A second limitation is that some articles are found behind paywalls. While most of these can be accessed through institution logins, it is unknown if the AI programs have access to them since their training database is kept confidential. Furthermore, at the time of data collection, ChatGPT was not trained on anything uploaded to the internet after September 2021, so its information may be outdated in rapidly-evolving fields. The ScholarAI plugin for ChatGPT allowed us to access articles published after the cutoff date, but introducing third-party plugins further increases the variability of results. Finally, the sentence style structure of the AI programs was very formulaic and may have introduced a potential form of bias for the reviewers. The AI text followed a very predictable format with most sections ending in "In conclusion," which may have led reviewers to prescribe lower ratings to the AI papers. Larger studies on more diverse topics may be warranted to further investigate the generative capacity of AI programs.

Conclusion

The realm of AI is an evolving field that warrants careful regulation. While it can serve as a valuable tool for facilitating the sharing of scientific discoveries and knowledge, its application should be underpinned by rigorous verification and meticulous fact-checking processes. To this end, stringent oversight and robust verification mechanisms are essential, in fact, these measures help safeguard against potential misinterpretation, bias, or misinformation that can sometimes result from the unregulated use of AI in scientific communication. With models that are based on prompts, such as the large language models used in this study, developing appropriate prompts to guide the model is essential. In essence, while AI holds the promise of enhancing knowledge sharing and it is continuously improving, it must be used responsibly and in conjunction with comprehensive fact-checking procedures to maintain the integrity of the scientific discourse.

Acknowledgement

We would like to thank Santiago Lozano-Caldaron, MD, PhD, Lercan Aslan, MD, Ayesha Yahya, DO, Bernard Burgesson, MD, and Rohan Bhimani, MD, MBA for their participation in the review process of this study.

Authors Contribution: Authors who conceived and designed the analysis: JW, NN, JK, SAE, MH/ Authors who collected the data: JW, NN/ Authors who contributed data or analysis tools: JW, NN/ Authors who performed the analysis: JW, NN/ Authors who wrote the paper: JW, NN, JK, SAE, MH

Declaration of Conflict of Interest: The authors do NOT have any potential conflicts of interest for this manuscript.

Declaration of Funding: The authors received NO financial support for the preparation, research, authorship, and publication of this manuscript.

Declaration of Ethical Approval for Study: This study did not involve human subjects and did not require ethical approval.

Declaration of Informed Consent: There is no information (names, initials, hospital identification numbers, or photographs) in the submitted manuscript that can be used to identify patients.

Jackson Woodrow BS [1]
Nour Nassour MD [1]
John Y. Kwon MD [1]
Soheil Ashkani-Esfahani MD [1]
Mitchel Harris MD [1]

1 Foot & Ankle Research and Innovation Lab (FARIL), Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

References

1. Zinkula J, Mok A. ChatGPT may be coming for our jobs. Here are the 10 roles that AI is most likely to replace. Business Insider. Available at: https://www.businessinsider.com/chatgpt-jobs-at-risk-replacement-artificial-intelligence-ai-labor-trends-2023-02. 2024.
2. Oremus W. Analysis | Google's AI passed a famous test — and showed how the test is broken. Washington Post. Available at: https://www.washingtonpost.com/technology/2022/06/17/google-ai-lamda-turing-test/. 2022.
3. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. Available at: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. 2023.
4. Edwards, B. OpenAI's GPT-4 exhibits "human-level performance" on professional benchmarks. ARS Technica. Available at: https://arstechnica.com/information-technology/2023/03/openai-announces-gpt-4-its-next-generation-ai-language-model/. 2023.
5. Ramazanian T, Fu S, Sohn S, Taunton MJ, Kremers HM. Prediction Models for Knee Osteoarthritis: Review of Current Models and Future Directions. Arch Bone Jt Surg. 2023; 11(1):1-11. doi: 10.22038/ABJS.2022.58485.2897.
6. Abedi R, Fatouraee N, Bostanshirin M, Arjmand N, Ghandhari H. Prediction of Fusion Rod Curvature Angles in Posterior Scoliosis Correction Using Artificial Intelligence. Arch Bone Jt Surg. 2024; 12(7):494-505. doi: 10.22038/ABJS.2024.76701.3545.

7. Dehouche N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). Ethics Sci Environ Polit. 2021; 21:17-23. doi:10.3354/esep00195.
8. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. BioRxiv. 2022 Dec 27:2022-12. doi:10.1101/2022.12.23.521610.
9. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023; 15(2):e35179. doi:10.7759/cureus.35179.
10. Athaluri SA, Manthena SV, Kesapragada VKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. 2023;15(4):e37432. doi: 10.7759/cureus.37432.
11. Pequeño A. Major ChatGPT update: AI program no longer restricted to September 2021 knowledge cutoff. Forbes. Available at: https://www.forbes.com.au/news/innovation/chatgpt-removes-september-2021-knowledge-cutoff/. 2023.
12. Salameh M, Al-Hashki L, Al-Juboori S, Rayyan R, Hantouly A, Blankenhorn B. Management of displaced intra-articular calcaneal fractures; current concept review and treatment algorithm. Eur J Orthop Surg Traumatol. 2023; 33(4):779-785. doi:10.1007/s00590-022-03264-5.
13. Herrera-Pérez M, Valderrabano V, Godoy-Santos AL, de César Netto C, González-Martín D, Tejero S. Ankle osteoarthritis: comprehensive review and treatment algorithm proposal. EFORT Open Rev. 2022; 7(7):448-459. doi:10.1530/EOR-21-0117.
14. Ramponi, M. How ChatGPT actually works. AssemblyAI. Available at: https://www.assemblyai.com/blog/how-chatgpt-actually-works/. 2022.
15. Kacena MA, Plotkin LI, Fehrenbacher JC. The Use of Artificial Intelligence in Writing Scientific Review Articles. Curr Osteoporos Rep. 2024; 22(1):115-121. doi:10.1007/s11914-023-00852-0.
16. Kitamura FC. ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment. Radiology. 2023; 307(2):e230171. doi:10.1148/radiol.230171.
17. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care. 2023; 27(1):75. doi:10.1186/s13054-023-04380-2.
18. Lee JY. The Use of Artificial Intelligence in Writing Scientific Review Articles. Curr Osteoporos Rep. 2024; 22(1):115-121. doi: 10.1007/s11914-023-00852-0.

Appendix

Appendix Figure 1A. Ankle Osteoarthritis Prompts

- Write a scientific journal review article on ankle osteoarthritis covering background, pathophysiology, clinical evaluation, classification, conservative treatment, patient education, diet, physical measures, footwear modifications, orthotics and insoles, pharmacological treatment, intra-articular therapies (hyaluronic acid, corticosteroids, PRP, stem cells), surgical treatment joint-reserving procedures, joint sacrificing surgical treatments, total ankle arthroplasty, arthrodesis, talar body replacement, and bipolar allograft. Write one section of the paper at a time based on my prompts. List and cite all sources in AMA style. Begin by writing the introduction by using the keywords ankle osteoarthritis, conservative treatment, and surgical management.
- Now write the pathophysiology section.
- Now write the clinical evaluation section.
- Now write the ankle osteoarthritis classification section. Now write a table summarizing the different classifications.
- Now write the section on conservative treatment.
- Now write a section on joint-preserving surgical procedures.
- Now write a section on joint-sacrificing surgical procedures.
- Now write a decision tree algorithm summarizing the conservative, joint-preserving, and joint-sacrificing procedures.
- Now write a section on intra-articular therapies.
- Now write the conclusion section.
- Now write the abstract.
- Give a title for the paper.
- Give me 3 to 5 highlights no more than 80 characters in length.
- Make a figure legend for the classification table you made.

Make a figure legend for the decision tree you made

Appendix Figure 2A. Intra-articular Calcaneus Fracture Prompts
• Write a scientific journal review article on the management of displaced intra-articular calcaneal fractures covering background, classification, operative versus non-operative management, extensile lateral approach versus sinus tarsi approach, minimally invasive and arthroscopic-assisted surgery, primary fusion, implant choice, bone grafting, and the conclusion. Write one subsection of the paper at a time based on my prompts. List and cite all sources in AMA style. Begin by writing just the introduction by using the keywords calcaneum, fracture, intra-articular, management, surgical. • Now write a section on the classification. • Now write a section on operative versus non-operative management. • Now write a section on operative management. • Now write a section on extensile lateral approach versus sinus tarsi approach. • Now write a section on minimally invasive and arthroscopic-assisted surgery. • Now write the section on primary fusion. • Now write a section on implant choice. • Now write a section on bone grafting. • Now write a conclusion section. • Now write an abstract. • Now give a title for the paper. • Give me 3 to 5 highlights no more than 80 characters in length. • Now give a reference section listing all of the references you mentioned above. • Create a table summarizing the types of fractures and the possible treatment options for each fracture.