Research Article

Pamukkale Medical Journal
doi:https://dx.doi.org/10.31362/patd.1487575

Artificial intelligence meets medical expertise: evaluating GPT-4's proficiency in generating medical article abstracts

Yapay zeka tıbbi uzmanlıkla buluşuyor: GPT-4'ün tıbbi makale özetleri oluşturmadaki yeterliliğinin değerlendirilmesi

Ergin Sağtaş, Furkan Ufuk, Hakkı Peker, Ahmet Baki Yağcı

Posted date:21.05.2024 Acceptance date:03.06.2024

Abstract

Purpose: The advent of large language models like GPT-4 has opened new possibilities in natural language processing, with potential applications in medical literature. This study assesses GPT-4's ability to generate medical abstracts. It compares their quality to original abstracts written by human authors, aiming to understand the effectiveness of artificial intelligence in replicating complex, professional writing tasks.

Materials and methods: A total of 250 original research articles from five prominent radiology journals published between 2021 and 2023 were selected. The body of these articles, excluding the abstracts, was fed into GPT-4, which then generated new abstracts. Three experienced radiologists blindly and independently evaluated all 500 abstracts using a five-point Likert scale for quality and understandability. Statistical analysis included mean score comparison inter-rater reliability using Fleiss' Kappa and Bland-Altman plots to assess agreement levels between raters.

Results: Analysis revealed no significant difference in the mean scores between original and GPT-4 generated abstracts. The inter-rater reliability yielded kappa values indicating moderate to substantial agreement: 0.497 between Observers 1 and 2, 0.753 between Observers 1 and 3, and 0.645 between Observers 2 and 3. Bland-Altman analysis showed a slight systematic bias but was within acceptable limits of agreement.

Conclusion: The study demonstrates that GPT-4 can generate medical abstracts with a quality comparable to those written by human experts. This suggests a promising role for artificial intelligence in facilitating the abstract writing process and improving its quality.

Key words: Artificial intelligence, ChatGPT, radiology, diagnosis, abstracts.

Sagtas E, Ufuk F, Peker H, Yagci AB. Artificial intelligence meets medical expertise: evaluating GPT-4's proficiency in generating medical article abstracts. Pam Med J 2024;17:756-762.

Öz

Amaç: GPT-4 gibi büyük dil modellerinin ortaya çıkışı, tıbbi literatürdeki potansiyel uygulamalarla birlikte doğal dil işlemede yeni olanaklar sağlamıştır. Bu çalışma GPT-4'ün tıbbi makale özetleri oluşturma yeteneğini değerlendirmektedir. Çalışma yapay zekanın karmaşık, profesyonel yazma görevlerini kopyalamadaki etkinliğini anlamayı amaçlamakta ve kalitelerini insan yazarlar tarafından yazılan orijinal özetlerle karşılaştırmaktadır.

Gereç ve yöntem: 2021-2023 yılları arasında yayınlanan beş önde gelen radyoloji dergisinden toplam 250 orijinal araştırma makalesi seçildi. Bu makalelerin tamamı, özetler hariç, GPT-4'e yüklendi ve daha sonra GPT-4 tarafından yeni özetler oluşturuldu. Üç deneyimli radyolog, kalite ve anlaşılabilirlik açısından beşli Likert ölçeği kullanarak 500 özetin tamamını kör ve bağımsız bir şekilde değerlendirdi. İstatistiksel analizde, değerlendiriciler arasındaki güvenilirliği ölçmek için Fleiss' Kappa testi ve değerlendiriciler arasındaki uyum düzeylerini değerlendirmek için Bland-Altman grafikleri kullanıldı.

Bulgular: Analiz, orijinal ve GPT-4 ile oluşturulan özetler arasında ortalama puanlar açısından anlamlı bir fark olmadığını ortaya koymuştur. Değerlendiriciler arası güvenilirlik açısından, orta ile önemli düzeyde uyuma işaret eden kappa değerleri bulunmuştur; değerler Gözlemci 1 ve 2 arasında 0.497, Gözlemci 1 ve 3 arasında 0.753 ve Gözlemci 2 ve 3 arasında 0.645 idi. Bland-Altman analizi hafif bir sistematik sapma göstermiş ancak kabul edilebilir uyum sınırları içinde kalmıştır.

Sonuç: Çalışma, GPT-4'ün insan uzmanlar tarafından yazılanlarla karşılaştırılabilir kalitede tıbbi özetler oluşturabildiğini göstermektedir. Yapay zeka kullanımı özet yazma sürecini kolaylaştırma ve kalitesini artırma konusunda önemli katkılar sağlayabilir.

Ergin Sağtaş, Assoc. Prof. Department of Radiology, Faculty of Medicine, Pamukkale University, Denizli, Türkiye, e-mail: sagtasergin@yahoo.com (https://orcid.org/0000-0001-6723-6593) (Corresponding Author)

Furkan Ufuk, Assoc. Prof. Department of Radiology, Faculty of Medicine, Pamukkale University, Denizli, Türkiye, e-mail: furkan.ufuk@hotmail.com (https://orcid.org/0000-0002-8614-5387)

Hakkı Peker, M.D. Department of Radiology, Faculty of Medicine, Pamukkale University, Denizli, Türkiye, e-mail: hakkipeker95@gmail.com (https://orcid.org0000-0002-9604-7529)

Ahmet Baki Yağcı, Prof. Department of Radiology, Faculty of Medicine, Pamukkale University, Denizli, Türkiye, e-mail: bakiyagci@yahoo.com (https://orcid.org/0000-0001-7544-5731)

Anahtar kelimeler: Yapay zeka, ChatGPT, radyoloji, tanı, özet.

Sağtaş E, Ufuk F, Peker H, Yağcı AB. Yapay zeka tıbbi uzmanlıkla buluşuyor: GPT-4'ün tıbbi makale özetleri oluşturmadaki yeterliliğinin değerlendirilmesi. Pam Tıp Derg 2024;17:756-762.

Introduction

Recent advancements in natural language processing have culminated in the creation of sophisticated large language models (LLMs) like GPT-4, which have demonstrated proficiency in producing high-quality text. GPT-4, in particular, has garnered significant interest for its capacity to generate text that is both coherent and richly informative across a diverse array of subjects [1-4]. LLMs offer educational support to medical students by enhancing their understanding with insightful explanations and demonstrating deductive reasoning [5, 6]. Patients also benefit from LLMs as they provide accurate information on various health conditions and offer emotional support, empowering patients and caregivers to navigate health challenges more effectively [7]. Moreover, LLMs can be used as a writing assistant in medical articles [8-10].

Abstracts in medical articles hold paramount importance as they serve as concise summaries that encapsulate the essential elements of a study, such as the objectives, methodology, results, and conclusions [11]. They function as a pivotal reference, enabling readers, including healthcare professionals and researchers, to swiftly discern the relevance and applicability of the study to their respective interests or fields. Abstracts facilitate quick decision-making by providing an accessible overview, which is especially crucial in the fast-paced medical environment where timely information is essential. They also enhance the visibility and accessibility of research by acting as a screening tool, allowing for efficient navigation through databases and journals and helping identify the most pertinent articles without delving into the full texts [11, 12]. Additionally, they play a crucial role in academic gatherings such as conferences, where they serve as a brief synopsis of the research, aiding participants in identifying sessions of interest. Thus, abstracts are instrumental in disseminating knowledge, fostering scientific communication, and facilitating informed decisions in medical practice and research [13]. GPT-4 can generate abstracts of medical articles, and the quality of the generated abstracts depends on various factors, such as the complexity of the content and the quality of the input provided. While GPT-4 is a valuable tool in assisting human authors, the ability and quality of abstract generation in radiology articles of GPT-4 have not been investigated yet. Herein, we aimed to assess the effectiveness of GPT-4 in generating research article abstracts and examine the quality of these abstracts.

Materials and methods

A reviewer (H.P.) collected a total of 250 research articles that were published between 2021 and 2023 in the five radiology journals (*Radiology, European Radiology, American Journal of Roentgenology, Japanese Journal of Radiology and Diagnostic and Interventional Radiology*). Fifty consecutive articles from each journal and sub-specialty (Abdominal, Breast, Cardiothoracic, Neuro, and Musculoskeletal radiology) were collected. The reviewer uploaded the text of the 250 articles to GPT-4, excluding the abstract section, and the abstracts were regenerated by GPT-4. The prompt fed to the GPT-4 were as follows:

1. *For articles in Radiology:* Generate an abstract for this article with a maximum word count of 300, using these subheadings: Background, Purpose, Materials and Methods, Results, and Conclusion.

2. *For articles in European Radiology:* Generate an abstract for this article with a maximum word count of 250, using these subheadings: Objective, Materials and methods, Results, Conclusions.

3. *For articles in American Journal of Roentgenology:* Generate an abstract for this article with a maximum word count of 350, using these subheadings: Background, Objective, Methods, Results, Conclusion, and Clinical Impact.

4. *For articles in Diagnostic and Interventional Radiology:* Generate an abstract for this article with a maximum word count of 400, using these subheadings: Purpose, Methods, Results, and Conclusion.

5. *For articles in Japanese Journal of Radiology:* Generate an abstract for this article with a maximum word count of 300, using these subheadings: Purpose, Materials and Methods, Results, and Conclusion.

Then the reviewer (H.P.) created a document including 250 original abstracts and 250 abstracts generated by GPT-4 in random order.

Three experienced academic radiologists with 8 (F.U.), 21 (E.S.), and 22 (A.B.Y.) years of experience in radiology independently evaluated the 500 abstracts using a five-point Likert scale about the quality and understandability of the abstract. The scoring in this Likert scale ranges from “Very poor” to “Very good”. A score of 1 represents a “Very poor” quality, indicating the lowest level of quality in the evaluation. A score of 2 corresponds to “Poor” quality, showing a level slightly better but still below average. A neutral or average quality is represented by a score of 3, labeled as “Fair”, indicating a middle ground in the quality assessment. A score of 4 corresponds to “Good” quality, indicating an above-average level of quality. Finally, the highest quality level is signified by a score of 5, labeled as “Very good”, representing the optimum level of quality in this scale. The observers conducted their evaluations without knowledge of whether the abstracts were originals or generated by GPT-4, ensuring that they were blind to the origin of each abstract to maintain objectivity in the assessment process.

Permission was obtained from Pamukkale University Non-Interventional Clinical Research Ethics Committee for the study.

Descriptive statistics including mean, median, standard deviation, and variance were calculated to summarize and describe the main aspects of the dataset and to give a comprehensive overview of the ratings. The Shapiro-Wilk Test was used to ascertain whether the dataset followed a normal distribution, guiding the selection between parametric and non-parametric tests. Independent samples t test (depending on the normality of the data) was conducted to compare the scores of the original abstracts against those generated by GPT-4, helping to identify if there were significant differences in quality perceptions. The Fleiss’ Kappa test, utilized to evaluate inter-rater reliability among the three experienced radiologists, yielded values that indicated the extent of agreement, with kappa (K) ranges typically interpreted as follows: below 0.20 signifying poor agreement, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.00 indicating almost perfect agreement [14]. To assess the concordance between the quality scores assigned to the original and GPT-4 generated abstracts, a Bland-Altman plot analysis was conducted, providing a visual representation of the agreement between observers and highlighting any systematic differences or anomalies. Statistical analyses were executed utilizing MedCalc version 20 (MedCalc Software) and SPSS version 23 (IBM), with a p value of less than 0.05 designated as the threshold for statistical significance.

Results

Three observers, in a blind and independent assessment, evaluated a total of 500 abstracts from 250 research articles, with 250 being the original versions and the remaining 250 re-generated using GPT-4. The analysis revealed no significant differences in the mean scores between the original and the GPT-4 generated abstracts across all observers, as detailed in Table 1. Furthermore, when comparing scores based on the journal and subspecialty, no significant differences were found. The p-values, according to the journal, were 0.384, 0.368, and 0.446 for Observers 1, 2, and 3, respectively. Regarding subspecialty, the P-values were 0.929, 0.610, and 0.871 for Observers 1, 2, and 3, correspondingly.

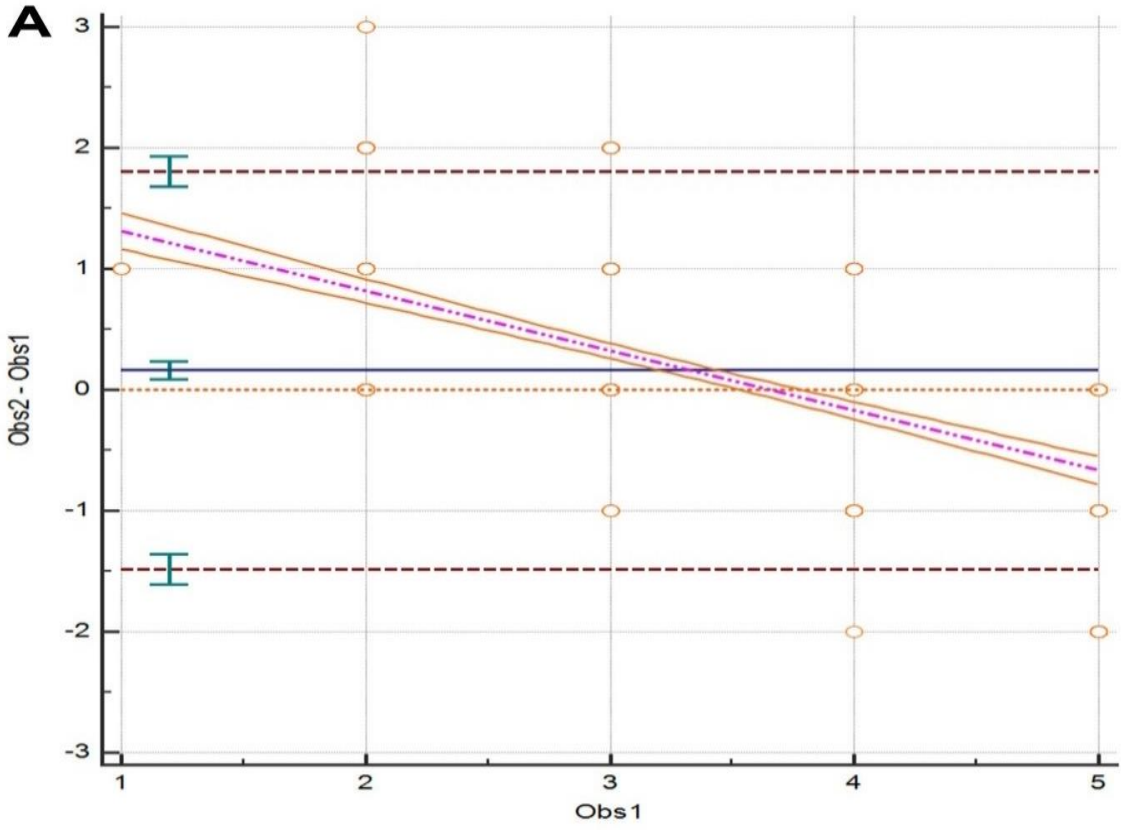
The assessments between Observer 1 and Observer 2 exhibited moderate agreement (κ=0.497) with a 95% confidence interval (CI) ranging from 0.442 to 0.552. Between Observer 1 and Observer 3, there was a substantial agreement (κ=0.753) with a 95% CI of 0.708 to 0.798. Similarly, a substantial agreement was noted between Observer 2 and Observer 3 (κ=0.645), with the 95% CI extending from 0.592 to 0.699.

Table 1. Comparative evaluation of abstract quality scores

Reference		Number of abstracts	Score (Mean±SD)	*p* value	t value
Observer 1	***Original***	250	3.32±0.98	0.989	-0.11
	GPT-4	250	3.33±1.03		
Observer 2	***Original***	250	3.36±0.85	0.107	-2.76
	GPT-4	250	3.57±0.85		
Observer 3	***Original***	250	3.4±0.91	0.867	-0.12
	GPT-4	250	3.41±1		

The p-values and t-values presented in the table represent the results of independent samples t-tests conducted to compare the means between the Original and GPT-4 groups

The Bland-Altman analysis was conducted to assess the agreement between the evaluations made by Observers 1, 2, and 3 (Figure 1). Systematic differences indicated by the mean differences were 0.1617 (95% CI: 0.08804 to 0.2353) for Observer 2 and 0.07984 (95% CI: 0.02648 to 0.1332) for Observer 3, respectively. These values suggest a small bias between Observer 1 and the other two observers. Limits of agreement, which define the range in which 95% of differences between observations by Observer 1 and the other observers lie, were calculated. For Observer 2, the limits of agreement ranged from -1.4826 (95% CI: -1.6086 to -1.3567) to 1.8060 (95% CI: 1.6801 to 1.9319), while for Observer 3, the range was -1.1118 (95% CI: -1.2030 to -1.0205) to 1.2714 (95% CI: 1.1802 to 1.3627).



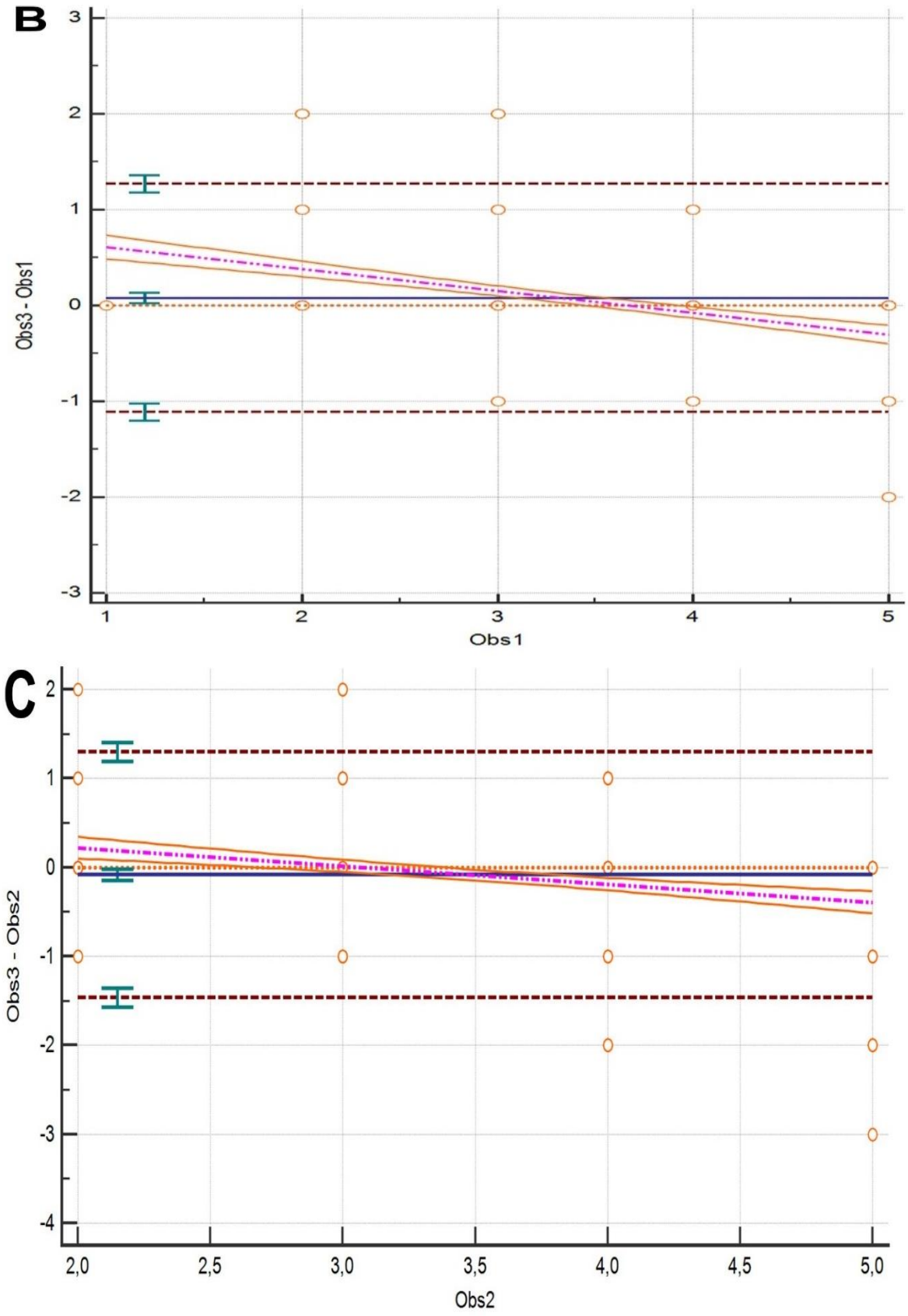


Figure 1. Bland-Altman plots for inter-observer agreement. A) Observer 1 and 2, B) Observer 1 and 3, C) Observer 2 and 3

Discussion

This study demonstrates that GPT-4 can create abstracts comparable in quality to their original counterparts, a finding reinforced by the negligible differences in mean scores assigned by three observers to both the original and the GPT-4-generated abstracts. It reveals that GPT-4 can adeptly undertake tasks typically reserved for skilled professionals, such as creating research article abstracts. Incorporating GPT-4 into the abstract writing process could positively influence the quality.

Recently, Jeblick et al. [15] and Li et al. [16] investigated the effectiveness of ChatGPT in simplifying radiology reports for better understanding and showed that ChatGPT regenerated reports in a way that was easily understood. In these studies, the authors showed that although the data were generally considered accurate and safe when evaluated by radiologists in terms of accuracy, completeness and safety, there were some errors and omissions that could mislead patients [15, 16]. While the present study found high-quality output indistinguishable from human-generated abstracts, the previous studies highlight a need for caution due to inaccuracies that could lead to patient harm. These results emphasize the necessity for further development and human oversight of LLMs within clinical practice.

There are only a few studies in the literature that evaluated the capabilities of LLMs, including GPT-4. In the present study on GPT-4's performance in abstract generation, GPT-4 produced work on par with human experts regarding quality, suggesting a high level of linguistic competence and understanding. Similarly, Ueda et al. [17] found GPT-4 capable of formulating differential and final diagnoses, highlighting its potential utility as a diagnostic aid. This is in line with the present study, where GPT-4 demonstrated the ability to synthesize and communicate complex medical information accurately. Fink et al. [18] also observed GPT-4's superior performance over ChatGPT in extracting and labeling data from oncologic CT reports. This suggests that GPT-4 has advanced text-processing abilities that can be precisely tuned to the subtleties of medical information extraction. Sun et al. [19] further extend the conversation by examining how GPT-4's generated impressions compare with human radiologists' work. While radiologists were favored for their detailed and accurate reports, non-radiologist physicians found GPT-4's outputs more straightforward and less likely to contribute to clinical missteps [19]. Comparing these studies reveals both the promise and the nuanced performance of GPT-4. While GPT-4 can replicate professional-level writing and data interpretation, it may not yet match the deep clinical understanding that comes with human expertise, as noted in Sun et al. [19] study. These findings collectively highlight the potential of GPT-4 as a supportive tool rather than a replacement for human professionals in medical settings.

This study has several limitations. Firstly, the assessment of abstract quality is inherently subjective, and despite the use of experienced radiologists as evaluators, their judgments may not fully represent the broader academic or clinical community. Secondly, the choice of articles and the prompts provided to GPT-4 could also influence the quality of the generated abstracts, potentially limiting the applicability of the findings to scenarios where such careful selection and prompting are not feasible. Lastly, the study only evaluated the abstracts based on quality and understandability without assessing other critical aspects such as accuracy of content, relevance, and the inclusion of key findings. Despite these limitations, this study boasts several notable strengths, including its methodologically sound approach, characterized by a rigorous blind and independent review process conducted by experienced radiologists using a well-established evaluation scale. Additionally, the study is pioneering in its exploration of AI's role in medical writing, aligning with contemporary technological trends and providing relevant insights for the application of LLMs in medical research and education.

In conclusion, the results reveal that LLMs can produce abstracts of a quality that is statistically indistinguishable from those written by human authors, as judged by experienced radiologists. The moderate to substantial agreement between observers and the slight systematic differences suggest that while GPT-4's capabilities are promising, there is a discernible variance in human evaluations of abstract quality. The negligible biases and

proportional differences in scores emphasize the potential of LLMs for assisting with medical writing tasks.

Conflict of interest: No conflict of interest was declared by the authors.

References

1. Elkassem AA, Smith AD. Potential Use Cases for ChatGPT in Radiology. AJR 2023;221:373-376. https://doi.org/10.2214/AJR.23.29198
2. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307:e230163. https://doi.org/10.1148/radiol.230163
3. Ufuk F. The role and limitations of large language models such as ChatGPT in clinical settings and medical journalism. Radiology 2023;307:e230276. https://doi.org/10.1148/radiol.230276
4. Sevgi UT, Erol G, Doğruel Y, Sönmez OF, Tubbs RS, Güngor A. The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. Neurosurg Rev 2023;46:86(e1-11). https://doi.org/10.1007/s10143-023-01998-2
5. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307:e230582. https://doi.org/10.1148/radiol.230582
6. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2023;30:80-90. https://doi.org/10.4274/dir.2023.232417
7. Amin K, Khosla P, Doshi R, Chheang S, Forman HP. Artificial intelligence to improve patient understanding of radiology reports. Yale J Biol Med 2023;96:407-417. https://doi.org/10.59249/NKOY5498
8. Ghim JL, Ahn S. Transforming clinical trials: the emerging roles of large language models. Transl Clin Pharmacol 2023;31:131-138. https://doi.org/10.12793/tcp.2023.31.e16
9. Tippareddy C, Jiang S, Bera K, Ramaiya N. Radiology reading room for the future: harnessing the power of large language models like ChatGPT. Curr Probl Diagn Radiol 2023;1-6. https://doi.org/10.1067/j.cpradiol.2023.08.018
10. Currie GM. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?. Semin Nucl Med 2023;53:719-730. https://doi.org/10.1053/j.semnuclmed.2023.04.008
11. Gastel B, Day RA. How to write and publish a scientific paper. 9th ed. Greenwood, USA: Bloomsbury Publishing, 2022.
12. Atzen SL, Bluemke DA. How to write the perfect abstract for radiology. Radiology 2022;305:498-501. https://doi.org/10.1148/radiol.229012
13. Woolston C. Words matter: jargon alienates readers. Nature 2020;579:309. https://doi.org/10.1038/d41586-020-00580-w
14. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm 2013;9:330-338. https://doi.org/10.1016/j.sapharm.2012.04.004
15. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2023:(e1-9). https://doi.org/10.1007/s00330-023-10213-1
16. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging 2023;101:137-141. https://doi.org/10.1016/j.clinimag.2023.06.008
17. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology 2023;308:e231040. https://doi.org/10.1148/radiol.231040
18. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology 2023;308:e231362. https://doi.org/10.1148/radiol.231362
19. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. Radiology 2023;307:e231259(e1-4). https://doi.org/10.1148/radiol.231259

Ethics committee approval: Permission was obtained from Pamukkale University Non-Interventional Clinical Research Ethics Committee for the study (approval date: 06.04.2023, and approval number: E-60116787-020-353871).

Authors' contributions to the article

E.S. and F.U. constructed the main idea and hypothesis of the study. E.S., F.U. and A.B.Y. developed the theory and arranged/edited the material and method section. H.P. and E.S. have evaluated the data in the Results section. Discussion section of the article written by E.S., F.U. and A.B. reviewed, corrected and approved. In addition, all authors discussed the entire study and approved the final version.