applied sciences

MDPI

Article

Automated Assessment of Reporting Completeness in Orthodontic Research Using LLMs: An Observational Study

Fahad Alharbi * and Saeed Asiri

Department of Pediatric Dentistry, College of Dentistry, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia; s.asiri@psau.edu.sa
* Correspondence: fahad.alharbi@psau.edu.sa

Abstract: This study evaluated the usability of Large Language Models (LLMs), specifically ChatGPT, in assessing the completeness of reporting in orthodontic research abstracts. We focused on two key areas: randomized controlled trials (RCTs) and systematic reviews, using the CONSORT-A and PRISMA guidelines for evaluation. Twenty RCTs and twenty systematic reviews published between 2018 and 2022 in leading orthodontic journals were analyzed. The results indicated that ChatGPT achieved perfect agreement with human reviewers on several fundamental reporting items; however, significant discrepancies were noted in more complex areas, such as randomization and eligibility criteria. These findings suggest that while LLMs can enhance the efficiency of literature appraisal, they should be used in conjunction with human expertise to ensure a comprehensive evaluation. This study underscores the need for further refinement of LLMs to improve their performance in assessing research quality in orthodontics and other fields.

Keywords: ChatGPT; CONSORT-A; orthodontic; artificial intelligence; natural language models; systematic reviews; randomized controlled trials; evidence-based medicine

Citation: Alharbi, F.; Asiri, S. Automated Assessment of Reporting Completeness in Orthodontic Research Using LLMs: An Observational Study. *Appl. Sci.* **2024**, *14*, 10323. https://doi.org/10.3390/app142210323

Academic Editors: Rosanna Guarnieri, Vincenzo D'Antò and Ersilia Barbato

Received: 24 September 2024
Revised: 4 November 2024
Accepted: 8 November 2024
Published: 10 November 2024

1. Introduction

The recent development of large language models (LLMs) has sparked widespread interest owing to their potential impact across many disciplines. These models, powered by massive transformer-based architectures and trained on vast datasets of text and code, have demonstrated remarkable capabilities in natural language processing (NLG). A prime example is the language model ChatGPT, which possesses an extensive training dataset comprising websites, books, and the entirety of Wikipedia [1,2]. This allows ChatGPT to perform a wide range of tasks, from question answering to creative writing [1].

Rapid progress in LLMs has generated significant interest in their potential applications across various fields. From revolutionizing natural language processing to automating content creation and powering knowledge-intensive applications, LLMs are poised to significantly impact diverse industries [2]. The medical field, in particular, presents fertile ground for LLM integration. Patients, caregivers, healthcare providers, researchers, academics, and legal professionals all benefit from these technologies.

The emerging literature suggests that large language models (LLMs) demonstrate potential for various applications within the healthcare domain, including triage, medical information translation, clinical research support, workflow optimization, medical education, and patient consultations. However, their current capabilities and effectiveness in clinical settings remain under investigation. Further studies are needed to validate their practical utility and limitations [2].

Preliminary research indicates that ChatGPT has the potential to be applied in healthcare settings, such as guiding patients to relevant departments, facilitating communication across language barriers, and supporting scientific writing through literature-review assistance and article summarization [3–5]. LLMs also have the potential for clinical applications,

ranging from diagnostic support and treatment recommendations to generating discharge summaries and medical notes and even extracting structured data from these documents [6–17]. Beyond clinical care, LLMs can be utilized for effective patient education [18–20].

Beyond clinical care, LLMs can be utilized for effective patient education [21–23]. LLMs can also be valuable assets in research projects, such as systematic reviews, where they can aid in tasks such as preparing Boolean search terms, screening abstracts, classifying articles, and performing content analysis [24,25]. While research suggests promising potential for LLMs in extracting explicitly stated information from articles, their performance appears less satisfactory for tasks requiring deeper analysis [24,26,27].

Numerous studies in orthodontics have analyzed the information provided by LLMs, particularly ChatGPT, on orthodontic topics [19,20,27–30]. Findings regarding the accuracy of ChatGPT responses vary. While some studies utilizing earlier versions found unreliable responses, others using updated models reported a higher accuracy.

Recent research has also assessed the ability of ChatGPT 3.5 and 4.0 to generate PICO-based queries in orthodontics, suggesting their effectiveness in producing research questions, keywords, and Boolean queries tailored to specific research objectives under appropriate settings [31]. Furthermore, a recent study compared ChatGPT's proficiency in assessing the reporting quality of randomized controlled trial abstracts according to the CONSORT-A statement against human reviewers. This study suggests LLMs like ChatGPT can potentially automate medical literature appraisal, aiding in the identification of accurately reported research [32].

Although LLMs hold significant promise, it is important to recognize that their performance across different applications can be inconsistent. LLM performance can vary from falling below human standards to achieving results comparable to human capabilities [25,26,31,33]. This study aims to assess the usability of LLMs, particularly ChatGPT, in assessing the completeness of reporting in two key areas of orthodontic research: abstracts of randomized controlled trials (RCTs) and abstracts of systematic reviews. Additionally, a secondary objective of this investigation was to identify areas where the assessments performed by LLMs differ from those conducted by human reviewers for both RCT and systematic review abstracts.

2. Materials and Methods

2.1. Sample Selection

This cross-sectional observational study investigated the quality of reporting in abstracts of randomized controlled trials (RCTs) and systematic reviews published in four leading orthodontic journals: (1) *American Journal of Orthodontics and Dentofacial Orthopedics* (AJO-DO), (2) *Journal of Orthodontics* (JO), (3) *European Journal of Orthodontics* (EJO), and (4) *The Angle Orthodontist* (AO). The timeframe included publications published between 2018 and 2022.

2.2. Identification of Relevant Articles

To identify relevant articles, a search was conducted for publications containing keywords indicative of randomized controlled trials (RCTs) or systematic reviews in the title or abstract. These keywords included "systematic review", "meta-analysis", "randomized controlled trial", "assigned", "prospective", or "comparative." Articles containing at least one of these keywords were retrieved for a full-text review to confirm that they truly represented a systematic review or RCT.

2.3. Selection of Studies

Following the initial search, a random sample of 20 RCTs and 20 systematic reviews was selected for further analysis. This resulted in a balanced representation, with each of the four journals contributing five publications on RCTs and five publications on systematic reviews.

2.4. Quality Assessment of RCT Abstracts

Two independent reviewers (F.A. and S.A.) assessed the reporting quality of RCT abstracts in duplicate using the CONSORT for the Abstract checklist. Reviewers directly referred to the full CONSORT guidelines and associated explanations for clarification. Disagreements were resolved through discussion until consensus was reached. Each checklist item received a score of "Yes" if reported, "No" if not reported, or "NA" if not applicable. Items were marked "NA" if the study design precluded reporting, such as blinding patients in studies comparing untreated controls to intervention groups [16]. Following individual scoring, the total score for each RCT abstract was calculated and converted to percentage using the following formula:

$$\text{Total Score} = (\text{Total "Yes" Items}/[19 - \text{Total "NA" Items}]) \times 100$$

2.5. Quality Assessment of Systematic Review Abstracts

Similar to the RCT abstracts, two independent reviewers (F.A. and S.A.) evaluated the reporting quality of the systematic review abstracts in duplicate using a checklist aligned with the PRISMA guidelines for abstract reporting. The reviewers consulted the full PRISMA guidelines and provided explanations during the assessment. Disagreements were resolved through discussion until consensus was reached. Each item was scored as "Yes" if reported, "No" if not reported, or "NA" if not applicable because of the specific review design (e.g., no meta-analysis conducted). Items marked "NA" were excluded from the analysis, and the denominator in the calculation was adjusted accordingly. The total score and percentage were computed for each systematic review abstract using the following formula:

$$\text{Total Score} = (\text{Total "Yes" Items}/[12 - \text{Total "NA" Items}]) \times 100$$

2.6. Prompt Design and Model Instructions

This study utilized the large language model GPT-3.5 (OpenAI, San Francisco, CA, USA), which is currently offered at no cost. We employed a common prompt engineering strategy known as in-context expert impersonation to enhance model performance. System prompts were initiated by introducing GPT-3.5 as an "expert in systematic reviews" for PRISMA guidelines and an "expert in clinical trial design" for CONSORT-A guidelines [33,34]. Subsequently, the model was instructed to list all items from the CONSORT checklist for reporting RCT abstracts. Following a concise one-sentence task explanation, "briefing information" was provided in the form of the included systematic review or RCT abstract, manually copied from PubMed (Supplementary File S1). This information served as the basis for the model analysis.

2.7. Chain-Of-Thought Prompting and Quality Control

To ensure the quality of the LLM output, we employed chain-of-thought prompting in the final user prompt [33]. This technique requires LLMs to perform three consecutive steps for each item on the reporting checklist. First, the model had to identify and extract relevant quotes from a full-text publication that directly addressed a specific item. Second, it was necessary to explain the rationale behind the chosen quotes and how they supported their assessment. Finally, the model assigned a bracketed rating for each item: "[Yes]" if reported, "[No]" if not reported, or "[NA]" if not applicable owing to the study design (e.g., blinding not possible or no meta-analysis conducted). Importantly, all three steps were included within a single prompt, and we refrained from providing feedback on the initial response to minimize bias. Because LLMs can generate inaccurate information, major deviations from instructions (e.g., missing ratings or hallucinating information) necessitated at least three manual prompt repetitions. Minor deviations, such as incorrect labeling or wording of the response, were corrected through a single manual intervention. We quantified both minor and major deviations for further analysis, and the completeness of the generated list

was verified by a single author (F.A.). Subsequently, the model was instructed to utilize this list for assessing the reporting of each item using a standardized response format: "Yes" if reported, "No" if not reported, or "NA" if not applicable due to the specific review design (e.g., no meta-analysis conducted).

2.8. Consistency and Data Management

To ensure consistency throughout the study, a single researcher (F.A.) formulated and submitted all prompts to language models. The text from the abstracts was then pasted into ChatGPT version 3.5 on 30 May 2024. A custom Excel (Microsoft Excel for Mac (v 16.90.2)) spreadsheet was used to facilitate data collection and scoring. The two authors independently assessed the accuracy of the collected responses by directly referencing the full CONSORT guidelines and the associated explanations for clarification. Disagreements were resolved through discussion until a consensus was reached.

2.9. Statistical Analysis

Frequencies and percentages were calculated to describe the distribution of ratings across all the categories. The ratings were treated as categorical variables and presented as absolute numbers and proportions of the cases. Comparisons were made using chi-square or Fisher's exact tests, as appropriate. All statistical analyses were performed using R software (R statistical software version 2.4.6.26) [35,36].

3. Results

3.1. Quality Assessment of RCT Abstracts

Twenty randomized controlled trials (RCTs) were independently evaluated by human raters and ChatGPT 3.5 using a 17-item checklist. The human raters and ChatGPT achieved perfect agreement in their assessment of the six checklist items (title identification, author details, trial design, intended intervention for each group, objectives, and conclusions). Near-complete agreement was observed for four additional items (eligibility criteria, clearly defined outcomes, outcome results for each group, and funding information). The alignment between human and ChatGPT ratings was lower for the remaining seven items, with statistically significant discrepancies identified for two items: randomization and recruitment details. Table 1 summarizes the characteristics of the human and ChatGPT ratings along with the findings from Fisher's exact test.

Table 1. The characteristics of human and ChatGPT ratings along with findings from the Fisher's exact.

Variable	N	ChatGPT, N = 20	Human, N = 20	*p*-Value
1. Title identification of the study as randomized	40			
Reported		20 (100%)	20 (100%)	
2. Authors contact details for the corresponding author	40			
Reported		20 (100%)	20 (100%)	
3. Trial design	40			0.11
Reported		20 (100%)	16 (80%)	
Not reported		0 (0%)	4 (20%)	
4. Participants' eligibility criteria for participants and the settings	40			>0.9
Reported		20 (100%)	19 (95%)	
Not reported		0 (0%)	1 (5.0%)	
5. Interventions intended for each group	40			
Reported		20 (100%)	20 (100%)	

Table 1. *Cont.*

Variable	N	ChatGPT, N = 20	Human, N = 20	*p*-Value
6. Objective	40			
Reported		20 (100%)	20 (100%)	
7. Outcome clearly defined	40			0.11
Reported		20 (100%)	16 (80%)	
Not reported		0 (0%)	4 (20%)	
8. Randomization/how participants were allocated to interventions	40			0.001
Reported		20 (100%)	11 (55%)	
Not reported		0 (0%)	9 (45%)	
9. Blinding	40			0.7
Reported		8 (40%)	9 (45%)	
Not reported		12 (60%)	11 (55%)	
10. Number of participants randomized to each group	40			0.5
Reported		16 (80%)	14 (70%)	
Not reported		4 (20%)	6 (30%)	
11. Recruitment trial status and period or duration	40			<0.001
Reported		20 (100%)	0 (0%)	
Not applicable		0 (0%)	20 (100%)	
12. Number of participants analyzed in each group	40			0.2
Reported		15 (75%)	11 (55%)	
Not reported		5 (25%)	9 (45%)	
13. Outcome result for each group and estimated effect size	40			>0.9
Reported		20 (100%)	19 (95%)	
Not reported		0 (0%)	1 (5.0%)	
14. Harms/important adverse events or side effects	40			0.082
Reported		8 (40%)	3 (15%)	
Not reported		11 (55%)	17 (85%)	
Not applicable		1 (5.0%)	0 (0%)	
15. Conclusions/general interpretation of the results	40			
Reported		20 (100%)	20 (100%)	
16. Trial registration	40		21 (100%)	0.5
Reported		9 (45%)	11 (55%)	
Not reported		11 (55%)	9 (45%)	
17. Funding	40			>0.9
Reported		4 (20%)	3 (15%)	
Not reported		16 (80%)	17 (85%)	

3.2. Quality Assessment of Systematic Review Abstracts

Twenty systematic reviews were independently assessed by human raters and the ChatGPT using a 12-item checklist. The human and ChatGPT ratings achieved perfect agreement for three items: identifying the report as a systematic review, objectives, and interpretation. Near-complete agreement was observed for information sources, the included studies, synthesis of results, and limitations of the evidence. The alignment

scores were lower for the remaining five items, with a statistically significant discrepancy identified for the eligibility criteria. Table 2 summarizes the characteristics of the human and ChatGPT ratings along with the findings from Fisher's exact test.

Table 2. The characteristics of human and ChatGPT ratings along with findings from the Fisher's exact test.

Variable	N	ChatGPT, N = 20	Human, N = 20	*p*-Value
1. Identify the report as a systematic review	40			
Reported		20 (100%)	20 (100%)	
2. Objectives	40			
Reported		20 (100%)	20 (100%)	
3. Eligibility criteria	40			0.028
Reported		18 (90%)	12 (60%)	
Not reported		2 (10%)	8 (40%)	
4. Information sources	40			0.11
Reported		20 (100%)	16 (80%)	
Not reported		0 (0%)	4 (20%)	
5. Risk of bias	40			>0.9
Reported		15 (75%)	15 (75%)	
Not reported		5 (25%)	5 (25%)	
6. Methods of synthesis results	40			0.14
Reported		17 (85%)	13 (65%)	
Not reported		3 (15%)	7 (35%)	
7. Included studies	40			>0.9
Reported		20 (100%)	19 (95%)	
Not reported		0 (0%)	1 (5.0%)	
8. Synthesis of results	40			0.2
Reported		18 (90%)	14 (70%)	
Not reported		2 (10%)	6 (30%)	
9. Limitation of evidence	40			>0.9
Reported		17 (85%)	17 (85%)	
Not reported		3 (15%)	3 (15%)	
10. Interpretation	40			
Reported		20 (100%)	20 (100%)	
11. Funding	40			0.091
Reported		6 (30%)	1 (5.0%)	
Not reported		14 (70%)	19 (95%)	
12. Registration	40			0.7
Reported		9 (45%)	8 (40%)	
Not reported		11 (55%)	12 (60%)	

4. Discussion

The increasing integration of LLMs, particularly ChatGPT, into the evaluation of reporting quality in orthodontic research highlights both the potential and limitations of AI in clinical settings. Our study aimed to assess the completeness of reporting in the abstracts of randomized controlled trials (RCTs) and systematic reviews using the CONSORT-A and PRISMA guidelines. The results indicated that while ChatGPT demonstrated a commendable ability to assess certain reporting elements, discrepancies remained when compared to human reviewers.

The findings revealed that ChatGPT achieved perfect agreement with human raters on several key items, such as trial design and objectives for RCTs and the identification of systematic reviews and their objectives for systematic reviews. This suggests that LLMs can effectively recognize and evaluate fundamental aspects of research reporting, which is consistent with previous studies that have highlighted the utility of AI in healthcare settings for tasks such as triage and diagnostic support [1,3]. However, the lower alignment of critical items, such as randomization details and eligibility criteria, highlights the limitations of LLMs in understanding delicate reporting requirements.

Several factors may contribute to the observed discrepancies between ChatGPT and human reviewers when assessing reporting quality. ChatGPT's training data likely lack sufficient examples of high-quality, detailed reporting in orthodontic research, leading to gaps in its understanding of field-specific requirements. While the AI possesses broad knowledge, it may lack the deep, specialized understanding that human experts in orthodontics have, particularly regarding methodological nuances. LLMs may struggle to grasp the full context of a study, including the rationale behind certain methodological choices, which human experts can infer based on their experience and field knowledge. Additionally, reporting guidelines like CONSORT and PRISMA are periodically updated, and ChatGPT's training data might not reflect the most current versions, leading to evaluation discrepancies. Unlike human reviewers, ChatGPT cannot seek clarification or additional information when faced with ambiguous reporting, potentially resulting in misinterpretations. Moreover, discrepancies may stem from the inherent complexity of clinical trial designs and the subtleties involved in systematic reviews, which require a deep understanding of methodological precision that current LLMs may not fully grasp [2]. These factors collectively contribute to the AI's limitations in assessing complex aspects of research reporting in orthodontics.

Moreover, the variability in performance across different reporting items underscores the need for the continued refinement of LLMs. While they can assist in automating the assessment process, reliance solely on these models can lead to oversights in critical reporting areas. This aligns with prior research indicating that LLMs, including ChatGPT, can perform comparably to human doctors in certain diagnostic scenarios but may struggle with more complex clinical decision making [4,11].

As the field of orthodontics continues to evolve, the integration of LLMs into research practices could enhance the efficiency of literature reviews and improve reporting standards. However, it is essential to approach this integration with caution to ensure that human expertise remains a cornerstone of the evaluation process. Future studies should focus on refining the prompts and training data used for LLMs, potentially incorporating feedback loops that allow for continuous learning and improvement [26].

Additionally, the potential of LLMs extends beyond evaluation; they could also facilitate improved reporting practices by guiding researchers in adhering to established guidelines, such as CONSORT and PRISMA. Automating parts of the literature review process can enhance the efficiency of research synthesis and promote higher standards of reporting [22,24]. However, as our study indicates, integration of LLMs should be approached with caution. Human expertise remains crucial, especially when interpreting complex data and ensuring comprehensive evaluation of research quality.

This study has several limitations. First, the sample size, while sufficient for a preliminary investigation, may not be large enough to draw definitive conclusions. A larger sample

size would provide greater statistical power and allow for more robust validation of LLM performance. Second, the evaluation of LLM performance relied on human raters, which inherently introduces subjectivity into the process. To mitigate this, two experienced reviewers independently assessed the articles, referring to the full CONSORT and PRISMA guidelines to ensure consistency. However, it is acknowledged that complete objectivity is difficult to achieve in such evaluations

5. Conclusions

In conclusion, this study highlights the potential of LLMs, particularly ChatGPT, for assessing the quality of reporting in orthodontic research. While the model demonstrated proficiency in evaluating certain aspects of RCT and systematic review abstracts, significant discrepancies were observed in more complex reporting items. These findings suggest that, while LLMs can serve as valuable tools in the research process, they should not replace human expertise. Instead, a hybrid approach that combines the strengths of LLMs with human oversight may yield the best outcomes for enhancing the quality of scientific reporting in orthodontics and beyond. As LLM technology continues to advance, ongoing research is necessary to optimize its application in clinical and academic settings.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app142210323/s1, File S1: List of RCTs and systematic reviews included in the study sample.

Author Contributions: F.A. contributed to conceptualization, study design, data collection, and analysis; S.A. contributed to data collection and conceptualization. Both authors contributed to drafting the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors gratefully acknowledge the generous support of Prince Sattam Bin Abdulaziz University, which facilitated the conduct of this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kalla, D.; Smith, N.; Samaah, F.; Kuraku, S. Study and Analysis of ChatGPT and Its Impact on Different Fields of Study. *Int. J. Innov. Sci. Res. Technol.* **2023**, *8*, 827–833. [CrossRef]
2. Li, J.; Dada, A.; Puladi, B.; Kleesiek, J.; Egger, J. ChatGPT in Healthcare: A Taxonomy and Systematic Review. *Comput. Methods Programs Biomed.* **2024**, *245*, 108013. [CrossRef] [PubMed]
3. Baker, A.; Perov, Y.; Middleton, K.; Baxter, J.; Mullarkey, D.; Sangar, D.; Butt, M.; DoRosario, A.; Johri, S. A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. *Front. Artif. Intell.* **2020**, *3*, 543405. [CrossRef] [PubMed]
4. Paslı, S.; Şahin, A.S.; Beşer, M.F.; Topçuoğlu, H.; Yadigaroğlu, M.; İmamoğlu, M. Assessing the Precision of Artificial Intelligence in Emergency Department Triage Decisions: Insights from a Study with ChatGPT. *Am. J. Emerg. Med.* **2024**, *78*, 170–175. [CrossRef]
5. Grimm, D.R.; Lee, Y.-J.; Hu, K.; Liu, L.; Garcia, O.; Balakrishnan, K.; Ayoub, N.F. The Utility of ChatGPT as a Generative Medical Translator. *Eur. Arch. Oto-Rhino-Laryngol.* **2024**, *281*, 6161–6165. [CrossRef] [PubMed]
6. Caruccio, L.; Cirillo, S.; Polese, G.; Solimando, G.; Sundaramurthy, S.; Tortora, G. Can ChatGPT Provide Intelligent Diagnoses? A Comparative Study between Predictive Models and ChatGPT to Define a New Medical Diagnostic Bot. *Expert Syst. Appl.* **2024**, *235*, 121186. [CrossRef]
7. Delsoz, M.; Madadi, Y.; Raja, H.; Munir, W.M.; Tamm, B.; Mehravaran, S.; Soleimani, M.; Djalilian, A.; Yousefi, S. Performance of ChatGPT in Diagnosis of Corneal Eye Diseases. *Cornea* **2024**, *43*, 664–670. [CrossRef]
8. Horiuchi, D.; Tatekawa, H.; Shimono, T.; Walston, S.L.; Takita, H.; Matsushita, S.; Oura, T.; Mitsuyama, Y.; Miki, Y.; Ueda, D. Accuracy of ChatGPT Generated Diagnosis from Patient's Medical History and Imaging Findings in Neuroradiology Cases. *Neuroradiology* **2024**, *66*, 73–79. [CrossRef]

9. Kozel, G.; Gurses, M.E.; Gecici, N.N.; Gökalp, E.; Bahadir, S.; Merenzon, M.A.; Shah, A.H.; Komotar, R.J.; Ivan, M.E. ChatGPT on Brain Tumors: An Examination of Artificial Intelligence/Machine Learning's Ability to Provide Diagnoses and Treatment Plans for Neuro-Oncology Cases. *Clin. Neurol. Neurosurg.* **2024**, *239*, 108238. [CrossRef]
10. Mayo-Yáñez, M.; González-Torres, L.; Saibene, A.M.; Allevi, F.; Vaira, L.A.; Maniaci, A.; Chiesa-Estomba, C.M.; Lechien, J.R. Application of ChatGPT as a Support Tool in the Diagnosis and Management of Acute Bacterial Tonsillitis. *Health Technol.* **2024**, *14*, 773–779. [CrossRef]
11. Oon, M.L.; Syn, N.L.; Tan, C.L.; Tan, K.B.; Ng, S.B. Bridging Bytes and Biopsies: A Comparative Analysis of ChatGPT and Histopathologists in Pathology Diagnosis and Collaborative Potential. *Histopathology* **2024**, *84*, 601–613. [CrossRef] [PubMed]
12. Panwar, P.; Gupta, S. A Review: Exploring the Role of ChatGPT in the Diagnosis and Treatment of Oral Pathologies. *Oral Oncol. Rep.* **2024**, *10*, 100225. [CrossRef]
13. Sandmann, S.; Riepenhausen, S.; Plagwitz, L.; Varghese, J. Systematic Analysis of ChatGPT, Google Search, and Llama 2 for Clinical Decision Support Tasks. *Nat. Commun.* **2024**, *15*, 2050. [CrossRef] [PubMed]
14. Shojaei, M. ChatGPT and Artificial Intelligence in Medical Endocrine System and Interventions. *Eurasian J. Chem. Med. Petrol. Res.* **2024**, *3*, 197–209.
15. Singh, S.; Djalilian, A.; Ali, M.J. ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes. *Semin Ophthalmol.* **2023**, *38*, 503–507. [CrossRef]
16. Kernberg, A.; Gold, J.A.; Mohan, V. Using ChatGPT-4 to Create Structured Medical Notes from Audio Recordings of Physician-Patient Encounters: Comparative Study. *J. Med. Internet Res.* **2024**, *26*, e54419. [CrossRef]
17. Huang, J.; Yang, D.M.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L.J.; et al. A Critical Assessment of Using ChatGPT for Extracting Structured Data from Clinical Notes. *npj Digit. Med.* **2024**, *7*, 106. [CrossRef]
18. Johnson, S.B.; King, A.J.; Warner, E.L.; Aneja, S.; Kann, B.H.; Bylund, C.L. Using ChatGPT to Evaluate Cancer Myths and Misconceptions: Artificial Intelligence and Cancer Information. *JNCI Cancer Spectr.* **2023**, *7*, pkad015. [CrossRef]
19. Hatia, A.; Doldo, T.; Parrini, S.; Chisci, E.; Cipriani, L.; Montagna, L.; Lagana, G.; Guenza, G.; Agosta, E.; Vinjolli, F.; et al. Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. *J. Clin. Med.* **2024**, *13*, 735. [CrossRef]
20. Abu Arqub, S.; Al-Moghrabi, D.; Allareddy, V.; Upadhyay, M.; Vaid, N.; Yadav, S. Content Analysis of AI-Generated (ChatGPT) Responses Concerning Orthodontic Clear Aligners. *Angle Orthodontist.* **2024**, *94*, 263–272. [CrossRef]
21. Ollivier, M.; Pareek, A.; Dahmen, J.; Kayaalp, M.E.; Winkler, P.W.; Hirschmann, M.T.; Karlsson, J. A Deeper Dive into ChatGPT: History, Use, and Future Perspectives for Orthopaedic Research. Knee Surg. *Sports Traumatol. Arthrosc.* **2023**, *31*, 1190–1192. [CrossRef] [PubMed]
22. Salvagno, M.; Taccone, F.S.; Gerli, A.G. Can Artificial Intelligence Help for Scientific Writing? Crit. *Care* **2023**, *27*, 75. [CrossRef]
23. Biswas, S. ChatGPT and the Future of Medical Writing. *Radiol. Soc. North Am.* **2023**, *307*, e223312. [CrossRef] [PubMed]
24. Alshami, A.; Elsayed, M.; Ali, E.; Eltoukhy, A.E.; Zayed, T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* **2023**, *11*, 351. [CrossRef]
25. Wang, S.; Scells, H.; Koopman, B.; Zuccon, G. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023. [CrossRef]
26. Mahmoudi, H.; Chang, D.; Lee, H.; Ghaffarzadegan, N.; Jalali, M.S. A Critical Assessment of Large Language Models for Systematic Reviews: Utilizing ChatGPT for Complex Data Extraction. *SSRN* **2024**. [CrossRef]
27. Kılınç, D.D.; Mansız, D. Examination of the Reliability and Readability of Chatbot Generative Pretrained Transformer's (ChatGPT) Responses to Questions about Orthodontics and the Evolution of These Responses in an Updated Version. *Am. J. Orthod. Dentofac. Orthop.* **2024**, *165*, 546–555. [CrossRef]
28. Campbell, D.J.; Estephan, L.E.; Mastrolonardo, E.V.; Amin, D.R.; Huntley, C.T.; Boon, M.S. Evaluating ChatGPT Responses on Obstructive Sleep Apnea for Patient Education. *J. Clin. Sleep Med.* **2023**, *19*, 1989–1995. [CrossRef]
29. Daraqel, B.; Wafaie, K.; Mohammed, H.; Cao, L.; Mheissen, S.; Liu, Y.; Zheng, L. The Performance of Artificial Intelligence Models in Generating Responses to General Orthodontic Questions: ChatGPT vs. Google Bard. *Am. J. Orthod. Dentofac. Orthop.* **2024**, *165*, 652–662. [CrossRef]
30. Makrygiannakis, M.A.; Giannakopoulos, K.; Kaklamanos, E.G. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models in Orthodontics: A Comparative Study of ChatGPT, Google Bard, and Microsoft Bing. *Eur. J. Orthod.* **2024**, *46*, cjae017. [CrossRef]
31. Demir, G.B.; Süküt, Y.; Duran, G.S.; Topsakal, K.G.; Görgülü, S. Enhancing Systematic Reviews in Orthodontics: A Comparative Examination of GPT-3.5 and GPT-4 for Generating PICO-Based Queries with Tailored Prompts and Configurations. *Eur. J. Orthod.* **2024**, *46*, cjae011. [CrossRef]
32. Roberts, R.H.; Ali, S.R.; Hutchings, H.A.; Dobbs, T.D.; Whitaker, I.S. Comparative Study of ChatGPT and Human Evaluators on the Assessment of Medical Literature According to Recognized Reporting Standards. *BMJ Health Care Inform.* **2023**, *30*, e100830. [CrossRef] [PubMed]
33. Woelfle, T.; Hirt, J.; Janiaud, P.; Kappos, L.; Ioannidis, J.; Hemkens, L.G. Benchmarking Human-AI Collaboration for Common Evidence Appraisal Tools. *J. Clin. Epidemiol.* **2024**, *175*, 111533. [CrossRef] [PubMed]

34. Salewski, L.; Alaniz, S.; Rio-Torto, I.; Schulz, E.; Akata, Z. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. *arXiv* **2024**, arXiv:2305.14930. [CrossRef]
35. R Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. [CrossRef]
36. Sjoberg, D.D.; Whiting, K.; Curry, M.; Lavery, J.A.; Larmarange, J. Reproducible Summary Tables with the gtsummary Package. *R J.* **2021**, *13*, 570–580. [CrossRef]