


BMJ Open Application of generative language models to orthopaedic practice

Jessica Caterson,¹ Olivia Ambler,² Nicholas Cereceda-Monteoliva,³ Matthew Horner,^{4,5} Andrew Jones,⁶ Arwel Tomos Poacher ^{4,7}

To cite: Caterson J, Ambler O, Cereceda-Monteoliva N, *et al.* Application of generative language models to orthopaedic practice. *BMJ Open* 2024;**14**:e076484. doi:10.1136/bmjopen-2023-076484

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2023-076484>).

Received 08 June 2023

Accepted 08 January 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹London School of Hygiene & Tropical Medicine, London, UK

²Plastic Surgery, Morriston Hospital, Swansea, UK

³Guy's and St Thomas' Hospitals NHS Trust, London, UK

⁴Trauma Department, University Hospital of Wales, Cardiff, UK

⁵Trauma and Orthopaedic Surgery, University Hospital of Wales, Cardiff, UK

⁶Orthopaedic Surgery, University Hospital of Wales, Cardiff, UK

⁷School of Biosciences, Cardiff University, Cardiff, UK

Correspondence to

Dr Arwel Tomos Poacher; drarwelpoacher@gmail.com

ABSTRACT

Objective To explore whether large language models (LLMs) Generated Pre-trained Transformer (GPT)-3 and ChatGPT can write clinical letters and predict management plans for common orthopaedic scenarios.

Design Fifteen scenarios were generated and ChatGPT and GPT-3 prompted to write clinical letters and separately generate management plans for identical scenarios with plans removed.

Main outcome measures Letters were assessed for readability using the Readable Tool. Accuracy of letters and management plans were assessed by three independent orthopaedic surgery clinicians.

Results Both models generated complete letters for all scenarios after single prompting. Readability was compared using Flesch-Kincaid Grade Level (ChatGPT: 8.77 (SD 0.918); GPT-3: 8.47 (SD 0.982)), Flesch Readability Ease (ChatGPT: 58.2 (SD 4.00); GPT-3: 59.3 (SD 6.98)), Simple Measure of Gobbledygook (SMOG) Index (ChatGPT: 11.6 (SD 0.755); GPT-3: 11.4 (SD 1.01)), and reach (ChatGPT: 81.2%; GPT-3: 80.3%). ChatGPT produced more accurate letters (8.7/10 (SD 0.60) vs 7.3/10 (SD 1.41), $p=0.024$) and management plans (7.9/10 (SD 0.63) vs 6.8/10 (SD 1.06), $p<0.001$) than GPT-3. However, both LLMs sometimes omitted key information or added additional guidance which was at worst inaccurate.

Conclusions This study shows that LLMs are effective for generation of clinical letters. With little prompting, they are readable and mostly accurate. However, they are not consistent, and include inappropriate omissions or insertions. Furthermore, management plans produced by LLMs are generic but often accurate. In the future, a healthcare specific language model trained on accurate and secure data could provide an excellent tool for increasing the efficiency of clinicians through summarisation of large volumes of data into a single clinical letter.

INTRODUCTION

Accurate and readable clinical letters are an essential part of orthopaedic practice. The British Orthopaedic Association (BOA) has issued guidance about writing clinical letters, advising the inclusion of likely diagnosis, investigations and management plan including any risk/benefits of such plan.¹

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This is the first study of its kind evaluating the use of language models in orthopaedic practice.
- ⇒ We have evaluated the use of a variety of language models and made comparisons between them a novel advancement in the utilisation of artificial intelligence models in clinical practice.
- ⇒ This study quantified readability and accuracy through the use of the validated 'Readable' tool rather than just using Likert scales.
- ⇒ This study is limited by the variation of clinical management provided by the language models which was mitigated using specialist consultants' agreement on a management plan for each case.

However, effective documentation can be time-consuming, and with rising caseloads and increasing workplace pressures, this guidance is often not adhered to. One study of fracture clinical letters reported only 26% contained relevant information,² and a later re-audit found an improvement only to 48%.³ With a record 730 000 people waiting on trauma and orthopaedic waiting lists as of March 2022, the burden of orthopaedic documentation, is only set to grow larger.⁴

One possible solution to the problem of increasing demand, is part-automation of the writing process using artificial intelligence (AI) technology. Natural Language Processing (NLP) offers a solution. NLP is a broad term encompassing multiple methods including text auto-completion and summarisation of large quantities of text.⁵ In November 2022, the public release of ChatGPT by OpenAI was seen as part of a greater paradigm shift in the capabilities of this technology.⁶

Chat-Generated Pre-trained Transformer (ChatGPT) is a supervised learning model (GPT-4, previously GPT-3), reinforced by human feedback.⁶ This model is designed to generate text responses by predicting the next word or sequence of words from

the input it receives.⁶ GPT-3 and GPT-4 have widespread functionality, from text summarisation and generation, to answering questions. Additionally, ChatGPT additionally has human-led fine-tuning for certain tasks or domains⁶ such as to have conversations with the user, as a chatbot or virtual assistant. Both are trained on a large data set accessed from a diverse range of sources from the internet.

The potential applications of ChatGPT in medical settings are vast and continue to be explored in academia, medical education and clinical practice. ChatGPT has been used to identify new systematic review prompts and write scientific papers and case reports.^{7 8} In education its use as a study tool to explain complex concepts, design scenarios for teaching and create multiple choice questions has been investigated.^{9–11} Researchers have also asked ChatGPT to give antimicrobial advice,¹² select appropriate imaging resources for breast pathology presentations¹³ and generate documentation such as discharge summaries and clinical letters in simulated clinical settings.^{14 15}

However, a frequent concern has been its accuracy, including responses that contain oversimplified, incomplete or falsified information, that could result in incorrect medical advice.^{8 11 12} For medical research, education and clinical practice, such inaccuracies could have damaging consequences.

The aim of this study is to explore the use of large language models (LLMs) in text summarisation and generation in common orthopaedic clinic scenarios, using GPT-3 and ChatGPT as examples which could be applied in this context.

METHODS

Four elective and 11 fracture clinical scenarios were simulated by authors ATP, MH and AJ (table 1). Core details, such as basic patient demographics, mechanism of injury, relevant medical history and social history, investigations, examinations and a management plan were composed in clinical note format (see online supplemental appendix 1 for full set of prompts and responses).

The first response for each scenario was collected to avoid selection bias, as responses are not consistent for repeated identical prompts. Each response was then assessed for readability and accuracy.

Readability was assessed with the online tool ‘Readable’.¹⁶ This tool, which has been used previously in similar readability studies,^{17 18} provides validated metrics such as Flesch-Kincaid Grade Level¹⁹, Flesch Reading Ease,¹⁹ Simple Measure of Gobbledygook (SMOG) Index²⁰ and reach.²¹

Accuracy was assessed by three independent senior orthopaedic clinicians using a Likert scale rating from 0 to 10; 0 indicating a completely inaccurate letter, and 10 being completely accurate. Outputs were also analysed for general tone, and any omissions and insertions noted by the clinicians and authors JC and ATP. An additional analysis was conducted assessing the ability of ChatGPT and GPT-3 to create appropriate management plans for each case. For each case, the following prompt was used: ‘Write an appropriate management plan for the following patient seen in an orthopaedic clinic based on the information provided:’. Again, management plans were rated for accuracy by three independent senior orthopaedic clinicians using the same Likert scale.

Table 1 Common orthopaedic scenarios selected as prompts for clinical letters, covering both elective and fracture clinical settings

Category	Case vignette
Elective case	Hip arthritis – total hip replacement. Knee arthritis – non-operative treatment. Carpal tunnel syndrome – surgical decompression. Lumbar disc prolapse – nerve root block.
Fracture clinic	Distal radial fracture – non-operative. Clavicle fracture – non-operative. Proximal humerus fracture – surgical fixation. Olecranon fracture – surgical fixation. Midshaft ulna fracture – non-operative. Biceps rupture – for further imaging. Buckle fracture – discharge. Quadriceps rupture – for further imaging. MCL rupture – non-operative. Weber A ankle fracture – discharge. Bimalleolar ankle fracture – surgical fixation.

ChatGPT and GPT-3 were then prompted to write clinical letters with the following prompts, for elective and fracture clinical cases, respectively: ‘Write an outpatient clinic letter for the following patient to the patient and their GP:’ and ‘Write a letter to the patient and their GP about the following:’.

GP, general practitioner; GPT, Generated Pre-trained Transformer; MCL, medial collateral ligament.

Readability and accuracy of the ChatGPT and GPT-3 responses were each compared with the original prompt using the paired t-test for each of the listed metrics. Statistical significance was deemed as $p < 0.05$. All statistical analysis was conducted using R (V.4.2.2).²²

Public and patient involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research given the nature of this study. Public and patient involvement was characterised as ‘none’.

RESULTS

ChatGPT and GPT-3 both created complete clinical letters with a single prompt (see online supplemental appendix 1 for all responses). Without any more specification than above, the letters included blanks to fill in the patient’s name and clinician responsible for composing the letter. ChatGPT also successfully wrote separate letters for the patient and general practitioner for single fracture clinic prompts.

For readability, the mean Flesch-Kincaid Grade Level was 8.77 (SD 0.918) and 8.47 (SD 0.982) for ChatGPT and GPT-3, respectively. This metric describes the approximate US school grade that would be expected to understand the letter, which in both prompts is equivalent to the reading level expected of children aged 14–15. These scores equate to mean Flesch Readability Ease scores of 58.2 (SD 4.00) and 59.3 (SD 6.98), respectively. SMOG Index, which is a measure of the average number of years of education needed for someone to understand a piece of text, was greater for ChatGPT letters than GPT-3 letters, with mean index scores of 11.6 (SD 0.755) and 11.4 (SD 1.01), respectively. GPT-3 letters also had a higher mean reach than ChatGPT-3 letters (81.2% vs 80.3%). Comparison of scores using paired Student’s t-test showed no statistically significant differences for any readability metric assessed (table 2).

The quality of the written content from both ChatGPT and GPT-3 was inconsistent. In some cases, the letters summarised all content well, with good inference of some relevant information. For example, it was able to infer that some occupations were relevant as their work could be impacted by their injury. However, in other cases,

relevant pieces of information were omitted, notably medical histories in elective cases. Both language models consistently added content unprompted. In some cases, this was relevant and accurate, for example, instructing a patient to follow *nil by mouth* instructions (assuming they had been told these), and counselling another on the risks of steroid use for muscular tendon rupture. However, sometimes these statements were inappropriate, such as describing that a cast will be removed after one week, when this would depend on a follow-up X-ray, and comments such as ‘Thank you for choosing our hospital’ and ‘return to my office’, which are not applicable in a UK setting. Both models were also inconsistent in handling abbreviations used in the prompts, for example ChatGPT described a ‘Web A’ fracture identical to the prompt, whereas GPT-3 changed this to ‘Weber A’. Conversely, ChatGPT changed ‘PMHx’ to medical history, whereas in some cases GPT-3 did not. In some cases when abbreviations were unabbreviated inappropriately, for example, a ‘high BMI’ was written as ‘we noted that you have a large body habitus’, which may be considered inappropriate in a patient’s letter.

The management plans provided were generic and sometimes inaccurate. The mean accuracy score for ChatGPT from three independent reviewers was 7.9/10 (SD 0.63) across all prompts, which was statistically significantly greater than for GPT-3 (mean accuracy 6.8/10 (SD 1.06, $p < 0.001$)) (figure 1B).

Both ChatGPT and GPT-3 were methodical in formation of management plans, consistently including points for pain management, physical therapy and lifestyle advice, as well as general advice of management for the orthopaedic injuries. However, specific management of the orthopaedic injury was generally unfocussed, often listing the most common conservative and surgical options, but failing to recognise the management the prompt was clearly pointing to. For example, a severe case of carpal tunnel syndrome requiring urgent decompression was first advised to try a splint. There were also notable inaccuracies, particularly duration of cast application, and in some cases outright incorrect suggestions, such as ‘[providing the] patient with written instructions on how to change the cast’. However, in other

Table 2 Mean readability scores for ChatGPT and GPT-3 derived responses (SD=SD deviation) and p value for paired t-test comparing mean difference. Statistical significance equates to p value < 0.05

Metric	Mean ChatGPT response (SD)	Mean GPT-3 response (SD)	P value (significance)
Flesch-Kincaid Grade Level	8.77 (0.918)	8.47 (0.982)	0.4023 (NS)
Flesch Readability Ease	58.2 (4.00)	59.3 (6.98)	0.4861 (NS)
SMOG Index	11.6 (0.755)	11.4 (1.01)	0.5870 (NS)
General Public Reach (%)	80.3 (4.20)	81.2 (5.78)	0.6218 (NS)

ChatGPT had a mean accuracy of 8.7/10 (SD 0.60) according to independent ratings from three senior orthopaedic clinicians. This compared with a mean accuracy of 7.3/10 (SD 1.41) for GPT-3 generated clinical letters. This difference was statistically significant ($p = 0.024$) (figure 1A). GPT, Generated Pre-trained Transformer; NS, not significant; SMOG, Simple Measure of Gobbledegoose.

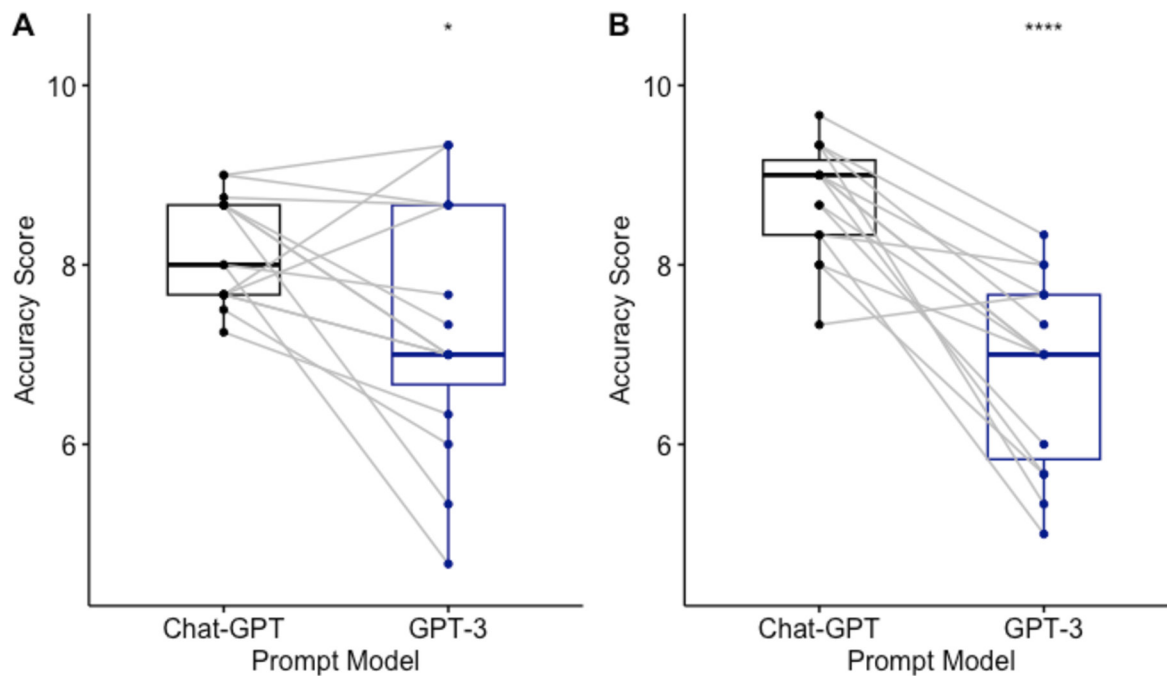


Figure 1 Accuracy scores for ChatGPT and GPT-3 generate (A) letters and (B) management plans, independently scored by three senior orthopaedic clinicians. Grey lines show paired prompts. Compared using paired t-test; *, $p < 0.05$, ****, $p < 0.001$. GPT, Generated Pre-trained Transformer.

cases, management plans were comprehensive and complete.

DISCUSSION

Our results demonstrate that currently available LLMs are capable of summarising text from clinician notes to readable clinical letters. Generally, these letters required limited prompting and had a readability score appropriate for those educated to a level of 14 years and above. Our letters were less readable than previous studies.¹⁴ This is likely because we did not include a specific readability instruction in our prompts. In addition, the prompts contained medical jargon including acronyms (eg, 'Web A'), which were frequently transferred to the letters produced by ChatGPT and GPT3. Such technical language will be less readable to a general reader. However, if prompted, ChatGPT and GPT-3 have both demonstrated the ability to simplify written text,¹⁴ and over time this can be standardised. Thus, while letters should still be checked by clinicians before sending, LLMs could be a useful tool in summarising clinical notes straight to readable patient clinical letters, with limited additional time and resources needed.

The accuracy of the letters was inconsistent, with notable omissions and inappropriate insertions by both models. This has been a recurring issue noted in multiple previous studies.^{8 11 12} In our study, ChatGPT and GPT-3 were able to infer information from the notes for some cases, but for others, this nuance was overlooked. ChatGPT was more accurate overall for letters and management plans, which may be because it uses the more sophisticated GPT-4, and it has been further refined with human reinforcement.

Both models added information to the letters. This could be useful, such as including lifestyle advice on smoking cessation, weight loss, pain management and rehabilitation following an injury without prompting. However, the information added was not always correct or appropriate. The addition of information, for example, when casts will be changed or removed, despite the prompts including defined management plans, could prove unsafe if these language models were freely applied in real clinical settings to give medical advice without appropriate oversight from clinicians. Given that language models generate text by adding words or sequences of words with the highest probability of following the prior text,²³ the highest probability text overall may not in every case be the correct one. This is especially true when considering a holistic approach to patient care.

The inaccuracies noted in this study may raise additional concerns regarding the liability of the developer of the LLM, and the clinician who uses the model to produce a clinical letter.²⁴ If incorrect medical advice is sent to the patient, this could potentially lead to harm. Attribution of this harm may be disputed if the clinician did not include the incorrect advice in their prompt, but signed off on the final letter containing the incorrect information. With the emergence of AI and automated processes, it is essential that, as well as improving the accuracy of these language models, these possibilities are considered and appropriately addressed by legal regulations and guidelines to ensure proper liability is set.

Ultimately, ChatGPT is not an approved medical device under the Health Insurance Portability and Accountability Act,^{24 25} and thus could not be used for

the purpose demonstrated. For a LLM to succeed in a clinical setting, it would need to be trained on robust medical data, with appropriately qualified clinicians providing feedback to ensure the accuracy of its outputs, and appropriately handle sensitive patient information.

Robust, medical-specific data could improve the accuracy of the outputs provided. However, there remains a risk of biases from the training data translating to the generated letters. Racial and sex biases were reported for GPT-3,²⁶ although with the added human feedback loop, some of these were addressed for ChatGPT. However, not all bias is as easy to detect, and there is an additional risk of an adversary intentionally introducing biases to favour or cause harm to certain patient groups.²⁷

Senior clinician review and feedback of any LLM, as well as patient involvement in the development process is also essential to ensure implementation of any LLM is accurate, appropriate and accepted. Additionally, clinicians should be provided with training to understand how to write prompts of a satisfactory quality, and to critically appraise any output generated. This may also assist with addressing any fears around ‘deskilling’ of clinicians; the LLM should aim to streamline the clinical letter writing process, but the clinician should still know what is appropriate to include in the letter, and ensure the final letter is of the expected standard.

Unlike ChatGPT or GPT-3, an LLM appropriate for clinical use would also need to have a secure database that adheres to current data protection standards, and maintain the privacy of any patient who’s information is input to generate a clinical letter. ChatGPT had a breach of privacy, where ~1.2% of user’s history or personal information could be accessed by other individuals.²⁸ If patient-identifiable data were used, this would be a significant breach of confidentiality,²⁹ and could raise serious ethical and legal issues and have widespread negative effects for the patient and the healthcare provider using the technology. Patients would also likely need to be appropriately informed of any use of an LLM, under the General Data Protection Regulation.²⁹ And, as part of its section on rights on individual, be able to withdraw their data from any model, and have the right to an explanation for any automatic processes involving their data.^{30 31} It is possible that their informed consent may also be necessary, given health is a protected characteristic under the regulation.^{31 32} However, if any LLM is demonstrated to be within the public’s interest, this may not be necessary.^{31 32}

This study compares 15 different orthopaedic scenarios in the elective and fracture clinical settings. While not entirely exhaustive, it provides a representative sample of the types of commonly encountered scenarios which an LLM might be used to summarise clinical notes into patient letters. In doing so, it

sufficiently demonstrates the utility of such an LLM tool in this clinical setting. This study focused on the potential of LLMs, and so the readability and accuracy of the generated letters was not compared with letters written by actual clinicians for patients. Given that there is scope for LLMs to aid in text summarisation, if an appropriate LLM approved for clinical use was generated, it would be important to compare this with the current standard to ensure and understand any benefit that was attained.

The current strength of ChatGPT and language models lies within their creation of readable text, not accurate text, and it is likely that the success of any LLM in a healthcare setting would be limited to this. If the drawbacks of current LLMs and legal and ethical issues could be addressed, it is clear that a healthcare specific language model trained on accurate and secure data would provide an excellent tool for increasing the efficiency of clinicians through usable summarisation of large volumes of data into a single clinical letter.

For fracture clinical documentation, such a tool would likely prove beneficial, for this use. The possibility of converting clinical notes or dictations straight to highly readable letters by automation of the more repetitive aspects, would be an attractive, time saving prospect. However, all clinical information should be specified by the clinician in the initial prompt, and the final letter checked for accuracy before sending any letter.

Twitter Jessica Caterson @jess_caterson and Arwel Tomos Poacher @arwelpoacher

Contributors JC: Study conceptualisation and design, data collection and analysis, first draft, edit and review. ATP: Study conceptualisation and design, data collection and analysis, first draft, edit, review, Guarantor. OA: First draft, edit and review. NC-M: review, data collection and data collection and analysis, first draft, edit and review. MH: First draft, edit and review. AJ: First draft, edit and review.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is

properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Arwel Tomos Poacher <http://orcid.org/0000-0002-4200-4929>

REFERENCES

- British Orthopaedic Association. Consultant advisory book. 2023. Available: <https://www.boa.ac.uk/standards-guidance/consultant-advisory-book.html> [Accessed 13 Apr 2023].
- Hook SE, Banister GC, Topliss C, et al. Letters and notes in orthopaedic surgery. *Ann R Coll Surg Engl* 2006;88:292–6.
- Longworth A, Davies D, Amirfeyz R, et al. Notes and Letters in Orthopaedic Surgery Revisited: Can Surgeons Change? *Bulletin* 2010;92:86–8.
- British Orthopaedic Association. England and Wales T&O Waiting Times data for, March . 2022 Available: <https://www.boa.ac.uk/resources/england-and-wales-t-o-waiting-times-data-for-march-2022.html#:~:text=There> [Accessed 13 Apr 2023].
- IBM. What is Natural Language Processing? | IBM, Available: <https://www.ibm.com/uk-en/topics/natural-language-processing> [Accessed 13 Apr 2023].
- OpenAI. Introducing ChatGPT, Available: <https://openai.com/blog/chatgpt> [Accessed 13 Apr 2023].
- Gupta R, Park JB, Bisht C, et al. Expanding Cosmetic Plastic Surgery Research With ChatGPT. *Aesthetic Surgery Journal* 2023;43:930–7.
- Manohar N, Prasad SS. Use of ChatGPT in Academic Publishing: A Rare Case of Seronegative Systemic Lupus Erythematosus in A Patient With HIV Infection. *Cureus*
- Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof* 2023;20:1.
- Seney V, Desroches ML, Schuler MS. Using ChatGPT to Teach Enhanced Clinical Judgment in Nursing Education. *Nurse Educ* 2023;48:124.
- Mogali SR. Initial impressions of ChatGPT for anatomy education. *Anat Sci Educ* February 7, 2023.
- Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *The Lancet Infectious Diseases* 2023;23:405–6.
- Rao A, Kim J, Kamineni M, et al. Evaluating chatgpt as an adjunct for radiologic decision-making. *Radiology and Imaging* [Preprint].
- Ali SR, Dobbs TD, Hutchings HA, et al. Using ChatGPT to write patient clinic letters. *The Lancet Digital Health* 2023;5:e179–81.
- Patel SB, Lam K. ChatGPT: the future of discharge summaries? *The Lancet Digital Health* 2023;5:e107–8.
- Readability score | Readability test | reading level Calculator | readable. Available: <https://readable.com/> [Accessed 13 Apr 2023].
- Wang L-W, Miller MJ, Schmitt MR, et al. Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social and Administrative Pharmacy* 2013;9:503–16.
- Burke V, Greenberg D, Commission on Adult Basic Education (U.S). Adult basic education: an interdisciplinary journal for adult literacy educators. *Adult Basic Educ Lit J Commission on Adult Basic Education* 1990.
- Kincaid J, Fishburne R, Rogers R, et al. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula). *Inst Simul Train*
- McLaughlin G. SMOG grading—A new readability formula in the journal of reading. 1969.
- What's new in October 2019? Introducing Reach – Readable, Available: <https://readable.com/blog/whats-new-in-october-2019/> [Accessed 13 Apr 2023].
- R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.
- Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback.
- Haupt CE, Marks M. AI-Generated Medical Advice-GPT and Beyond. *JAMA* 2023;329:1349–50.
- Harvey H, Pogose M. How to get ChatGPT regulatory approved as a medical device, Available: <https://www.hardianhealth.com/blog/how-to-get-regulatory-approval-for-medical-large-language-models> [Accessed 13 Apr 2023].
- Chiu K-L, Collins A, Alexander R. Detecting Hate Speech with GPT-3.
- Zou A, Wang Z, Kolter JZ, et al. Universal and Transferable Adversarial Attacks on Aligned Language Models.
- OpenAI. March 20 Chatgpt outage: here's what happened. Available: <https://openai.com/blog/march-20-chatgpt-outage#technical-details> [Accessed 13 Apr 2023].
- Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR), Available: <https://gdpr-info.eu/art-4-gdpr/> [Accessed 13 Apr 2023].
- What is automated individual decision-making and profiling; 2018.
- Art. 22 GDPR - Automated individual decision-making, including profiling - GDPR.eu, Available: <https://gdpr.eu/article-22-automated-individual-decision-making/> [Accessed 27 Nov 2022].
- What is valid consent?; In detail