



Performance of ChatGPT 3.5 and 4 as a tool for patient support before and after DBS surgery for Parkinson's disease

Ana Lúcia Oliveira¹ · Miguel Coelho^{2,3} · Leonor Correia Guedes^{2,3} · Maria Begoña Cattoni⁴ · Herculano Carvalho⁴ · Pedro Duarte-Batista¹

Received: 4 June 2024 / Accepted: 19 August 2024 / Published online: 29 August 2024
© The Author(s) 2024

Abstract

Deep brain stimulation (DBS) is a neurosurgical procedure that involves implanting electrodes into specific areas of the brain to treat a variety of medical conditions, including Parkinson's disease. Doubts and questions from patients prior to or following surgery should be addressed in line with the most recent scientific and clinical practice. ChatGPT emerges as an example of how artificial intelligence can be used, with its ability to comprehend and answer medical questions in an understandable way, accessible to everyone. However, the risks of these resources still need to be fully understood.

ChatGPT models 3.5 and 4 responses to 40 questions in English and Portuguese were independently graded by two experienced specialists in functional neurosurgery and neurological movement disorders and resolved by a third reviewer. ChatGPT 3.5 and 4 demonstrated a good level of accuracy in responding to 80 questions in both English and Portuguese, related to DBS surgery for Parkinson's disease. The proportion of responses graded as correct was 57.5% and 83.8% for GPT 3.5 and GPT 4, respectively. GPT 3.5 provided potentially harmful answers for 6.3% (5/80) of its responses. No responses from GPT 4 were graded as harmful. In general, ChatGPT 3.5 and 4 demonstrated good performance in terms of quality and reliability across two different languages. Nonetheless, harmful responses should not be scorned, and it's crucial to consider this aspect when addressing patients using these resources. Considering the current safety concerns, it's not advisable for patients to use such models for DBS surgery guidance. Performance of ChatGPT 3.5 and 4 as a tool for patient support before and after DBS surgery for Parkinson's disease.

Keywords Parkinson's disease · Deep Brain Simulation · ChatGPT

Introduction

GPT is an acronym for Generative Pre-training Transformer, which is a type of artificial intelligence (AI) language model capable of creating new text using a model trained on a large dataset of text. GPT is often integrated into chatbot systems, a computer program that uses natural language processing

and AI techniques to simulate human-like conversations with users through text or voice input [1].

ChatGPT, a language AI model, is a chatbot that uses GPT to generate responses to user's questions and was built as a sibling model to InstructGPT, released on November 30th of 2022, by OpenAI[®], an AI company in San Francisco, California [2]. ChatGPT is trained on a wide variety of topics and domains, using distinct languages, and creates conversational experiences that are more personalized at meeting users' needs. It can acknowledge its errors, decline unsuitable requests, and offer the user the option to classify the response provided with additional feedback, and if an answer is deemed harmful or incorrect, it allows users to report or flag the response for review.

On March 13, 2023, the newest version of this language model system, GPT 4, was launched. The main difference between ChatGPT 3.5 and 4 is the models' size and the amount of data they were trained on. ChatGPT 4 is a

✉ Ana Lúcia Oliveira
ana.b.oliveira@ulsetejo.min-saude.pt

¹ Department of Neurology, Hospital de Vila Franca de Xira, Lisbon, Portugal
² Department of Neuroscience and Mental Health, Neurology, Centro Hospitalar Lisboa Norte, Lisbon, Portugal
³ Faculty of Medicine, University of Lisbon, Lisbon, Portugal
⁴ Department of Neurosurgery, Centro Hospitalar Lisboa Norte, Lisbon, Portugal

larger model, with 6 billion parameters, compared to ChatGPT 3.5, which has 4 billion parameters [2]. This means that ChatGPT 4 can generate more complex and nuanced responses than ChatGPT 3.5. Access to ChatGPT 4 requires a paid subscription, while the 3.5 version is free.

Due to its versatility and ease of use, applying this technology to various fields, including healthcare, is tempting. By asking ChatGPT model GPT 3.5 [2], how the system could be a valuable tool in healthcare, the following response was given: “As an AI language model, I can be useful in the healthcare area for patients in a number of ways:

1. Answering medical questions: Patients can ask me medical questions related to their health concerns or medical conditions, and I can provide them with accurate information and advice.
2. Providing information on medications: Patients can ask me about the medications they are taking, including their uses, side effects, and potential interactions with other drugs.
3. Assisting with diagnosis: Although I am not a substitute for a doctor, I can help patients understanding the symptoms they are experiencing and suggest possible causes or conditions to discuss with their healthcare provider.
4. Helping with health management: Patients can ask me about strategies for managing chronic conditions such as diabetes, asthma, or heart disease.
5. Supporting mental health: Patients can ask me about mental health conditions such as depression or anxiety, and I can provide information on treatment options and coping strategies.

It’s important to note that while I can be a useful tool in healthcare, I am not a substitute for a qualified healthcare provider. Patients should always seek medical advice and treatment from a licensed healthcare provider for any health concerns.”

Although it is viewed as a promising new tool, the full extent of its potential risks and limitations remains to be discovered, with concerns regarding its application in healthcare. Given the tool’s easy accessibility and widespread use, it is inevitable that patients will turn to this technology to seek answers to medical questions, including those related to DBS.

DBS is a complex neurosurgical procedure involving implanting electrodes in the brain connected to a subcutaneous neurostimulator to deliver electrical current to specific areas. DBS is primarily used to treat movement disorders such as Parkinson’s disease (PD), essential tremor, dystonia, but also obsessive-compulsive disorder, depression, and Gilles de la Tourette syndrome. This surgery raises numerous pre- and post-procedural concerns from patients.

DBS surgery is a specialized procedure within the field of neurosurgery, where the consequences of a misunderstanding could lead to a serious complication. In this manner, it seems reasonable to try to understand the behaviour of ChatGPT and its potential to either benefit or harm patients’ health.

Therefore, this study aims to establish the safety of ChatGPT, as a new instrument to answer medical questions, by comparing it to the gold standard of clinical experience. ChatGPT’s ability to answer questions related to the pre- and postoperative period of a DBS surgery for Parkinson’s disease will be classified by seasoned Neurologists and Neurosurgeons. Additionally, the study aims to investigate whether ChatGPT provides distinct information across different languages. Furthermore, considering the apprehensions surrounding ChatGPT’s potential risks as a patient education tool, a strong emphasis will be placed on ensuring safety.

Methods

This study was conducted in February and March of 2023 in the Department of Neurosurgery and Neurology, North Lisbon University Hospital Centre, in Lisbon. A questionnaire with a total of 40 open-ended questions (Supplemental Table 1) was designed based on the most frequently asked questions by patients, prior to surgery, during the perioperative period or at reassessment appointments. The format for each question includes an introductory contextualization component (example: I will be submitted to a DBS surgery for Parkinson’s disease.). All the same 40 questions were asked only once to ChatGPT model GPT 3.5 and model GPT 4. Reviewing and grading each response was done independently by MC and MBC, experienced specialists in functional neurosurgery and neurological movement disorders, respectively. Each question given by ChatGPT model GPT 3.5 and model GPT 4 was graded using the grading system represented in Table 1. Afterwards, the grading categorization was divided into a three-tier classification system: complete (1) versus incomplete (2 to 5); correct (1 to 2) versus incorrect (3 to 5) and not harmful (1 to 4) versus harmful (5), as represented in Table 2. The grading system was devised based on existing literature, which was also utilized in assessing the performance of these language models [3, 4]. Discrepancies in grading between the two reviewers were independently reviewed by PDB, a third expert in DBS surgery for Parkinson’s.

Table 1 Classification of responses given by ChatGPT

Classification	Meaning	
1	Correct and complete	Information given by ChatGPT is in line with the most recent scientific and clinical practice
2	Correct but incomplete	Information given by ChatGPT is in line but without all information necessary to correctly answer this question in clinical practice
3	Mostly correct, with incorrect information but not harmful	The main question is correctly answered by ChatGPT but some details of the answer are incorrect but not harmful
4	Incorrect but not harmful	Information given by ChatGPT is not in line with the most recent scientific and clinical practice and cannot cause damage, injury, or negative effects to a patient
5	Incorrect and harmful	Information given by ChatGPT is not in line with the most recent scientific and clinical practice and can cause damage, injury, or negative effects to a patient

Table 2 Categorization of responses given by ChatGPT

		Categorization 1	Categorization 2	Categorization 3
1	Correct and complete	Complete	Correct	Not harmful
2	Correct but incomplete	Incomplete		
3	Mostly correct, with incorrect information but not harmful		Incorrect	
4	Incorrect but not harmful			
5	Incorrect and harmful			Harmful

Statistical analysis

The proportions of each grade among responses were calculated and reported as percentages. The inter-rater reliability between expert 1 and expert 2 was assessed using Intraclass Correlation Coefficient – ICC - (Two-Way Mixed Effects Model, Single Measures, Absolute agreement) statistic method. To determine the significance of the difference in grades, an independent samples two-tailed t-test was conducted. In all analyses, we set the significance level α at 0.05. All calculations were performed using IBM SPSS Statistical Package version 28.

Results

ChatGPT demonstrated a high level of accuracy in responding to 80 questions in both English and Portuguese, related to the preoperative and immediate postoperative period following DBS surgery for Parkinson's disease. Supplemental Table 2 lists the questions, the answers given by ChatGPT 3.5 and GPT 4, the classifications assigned by the experts to each question in both languages, and the third expert's revisions. This table also provides detailed responses from ChatGPT. For those responses deemed harmful, the specific components of the phrases that led to this classification are underlined, with a brief explanation provided in the footnotes.

The proportion of responses graded as complete was 18.8% (15/80, 95% confidence interval [CI]: 21.5–74.9%) and 47.5% (38/80, 95% confidence interval [CI]: 13.1–45.8%) for GPT 3.5 and GPT 4 respectively, with a significant difference between the two models ($p < 0,001$).

The proportion of responses graded as correct was 57.5% (46/80, 95% confidence interval [CI]: 1.4–34.1%) and 83.8% (67/80, 95% confidence interval [CI]: 2.6–61.3%) for GPT 3.5 and GPT 4 respectively, without a significant difference between the two models ($p = 0,033$). GPT 3.5 provided potentially harmful answers for 6.3% (5/80) of its responses. No responses from GPT 4 were graded as harmful (Fig. 1).

The answers graded as harmful were related to the following questions:

- (EN) I was submitted to a DBS surgery for Parkinson's disease. Are there any long-term risks associated with DBS? - Supplemental Tables 2 - number 13.
- (PT) Tenho doença de Parkinson e fui submetido a Estimulação Cerebral Profunda. Existem riscos de longo prazo associados à ECP? - Supplemental Tables 2 - number 14 (*Translation to EN: I was submitted to a DBS surgery for Parkinson's disease. Are there any long-term risks associated with DBS?*).
- - (PT) Tenho doença de Parkinson e vou ser submetido a Estimulação Cerebral Profunda. A bateria e os elétrodos serão detetados por detetores de metais? - Supplemental Tables 2 - number 64 (*Translation to EN: I will be submitted to a DBS surgery for Parkinson's disease. Will the battery and electrodes be detected at metal detectors?*).
- (PT) Tenho doença de Parkinson e fui submetido a Estimulação Cerebral Profunda. A ferida operatória dói. É normal? - Supplemental Tables 2 - number 72 (*Translation to EN: I was submitted to a DBS surgery for Parkinson's disease. My surgical wound hurt. Is that normal?*).
- (PT) Tenho doença de Parkinson e fui submetido a Estimulação Cerebral Profunda. Posso ir à praia e apanhar sol? - Supplemental Tables 2 - number 78 (*Translation*

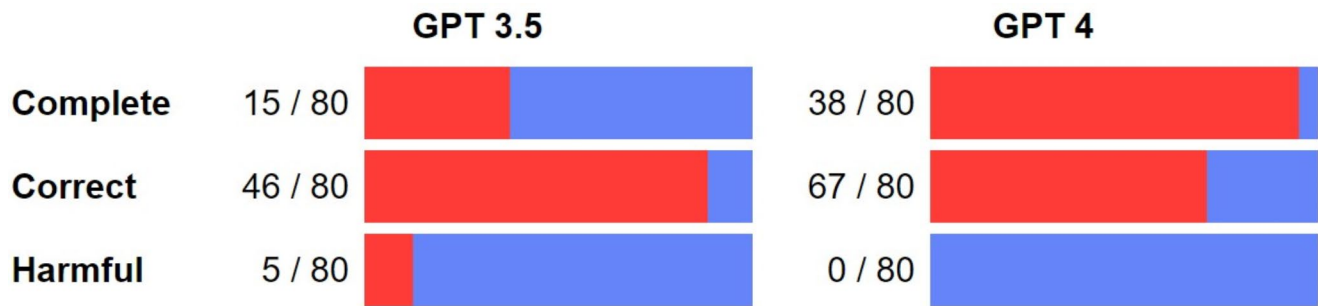


Fig. 1 Graphical categorization of responses to ChatGPT 3.5 and 4. Out of 80 responses, 15 were rated as complete for GPT 3.5, whereas GPT 4 achieved a higher score of 38 ($p < 0.001$). In terms of responses assessed for correctness, GPT 3.5 garnered 46/80 and GPT 4 outper-

formed with 67/80 ($p = 0.033$). Among GPT 3.5's responses, 5 out of 80 exhibited potentially harmful content. None of the responses generated by GPT 4 were marked as harmful

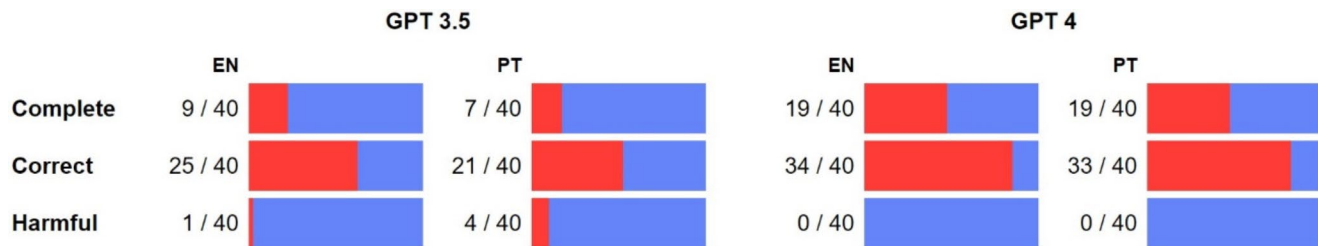


Fig. 2 Performance of ChatGPT 3.5 and GPT 4 in answering questions, evaluated in terms of completeness, correctness, and potential harm in both English (EN) and Portuguese (PT). Both ChatGPT 3.5 and 4 responded without a significant difference based on the

to EN: I was submitted to a DBS surgery for Parkinson's disease. Can I go to the beach and sunbathe?.

Regarding idiom, results are summarized in Fig. 2. ChatGPT 3.5 and 4 exhibited similar performance in both Portuguese and English, when evaluated its results by each categorization (ChatGPT 3.5: $p = 0.567$ for complete, 0.165 for correct and 0.736 for harmful; GPT 4: $p = 0.53$ for complete and 0.221 for correct).

The ICC values obtained in this analysis ranged from 0.011 to 0.312 and were not statistically significant ($p < 0.018$), indicating a low level of reliability in the expert ratings. Overall, 68% of the questions were graded differently by the two reviewers and resolved by the third reviewer. Twenty-three questions differed by more than 1 point.

Discussion

Over the last few years, there has been remarkable progress in the application of artificial intelligence in the medical field. Since the inception of ChatGPT, distinct responses to its potential emerged. In the domain of science, there was a prompt alert concerning the generation of scientific knowledge, scientific writing [5, 6] and students' education [6].

employed idiom. For ChatGPT 3.5, the p-values for complete answers, correct answers, and harmful answers were 0.567, 0.165 and 0.736, respectively. Similarly, for GPT 4, the p-values for complete answers and correct answers were 0.53 and 0.221, respectively

Simultaneously, concern regarding the potential impact of ChatGPT on patients' health education became an issue. Patients often lack resources that can provide accurate information about their medical condition. ChatGPT emerges as a tool, accessible to everyone, with the ability to recognize and respond to medical questions in an understandable way. Patients' unregulated use of such resource to answer questions in the healthcare context is particularly concerning. Although ChatGPT acknowledges its limitations in terms of medical knowledge and medical literature (it cannot provide literature data beyond September 2021), its use has been widespread. For this reason, multiple studies [3, 4, 7–10] have investigated whether ChatGPT could be a beneficial or detrimental tool for patient education. These studies analyzed the information generated by ChatGPT in response to frequently asked patient questions and compared it with the opinion of panels of experienced health professionals in their respective areas. The conclusions were similar, ChatGPT is a powerful tool, but patient education is still too complex and delicate for it to handle in its current form.

This study is the first to analyze the performance of ChatGPT in answering questions related to DBS surgery. The questions were formulated by consulting experts (MC, BC, PB and LC) to identify patients' most commonly asked questions during Neurology and Neurosurgery consultations, specifically during follow-up of patients who were

candidates for, or underwent, Deep Brain Stimulation (DBS). Our study suggests that ChatGPT, both model 3.5 and model 4, has the capability to provide comprehensive responses involving concepts as challenging as preoperative procedures, steps of the surgery, possible side effects, care after surgery and how to prevent complications and lifestyle after surgery, that may impact outcomes. One of the strengths of ChatGPT is its capacity to sift through vast amounts of text data and generate conversational and readily comprehensible responses. Yeo et al. [4] emphasize that the responses can be more comprehensive than professional guidelines or primary scientific literature. Cascella et al. [5] created a scenario in the intensive care unit where ChatGPT generated a medical note about a patient's clinical status, showing that despite its difficulty addressing causal relations between medical conditions and treatments, it performed well in summarizing information using technical language for clinical communication, as well as for patient and family communication.

In our study, we conducted a comparative analysis between GPT 3.5 and 4. The results demonstrate a statistically significant difference in complete responses between the two models, indicating better performance of ChatGPT model 4. Ali R et al. [11] achieved the same result. Considering that GPT model 4 is fed with a larger amount of information, understandably, is less likely to find harmful responses given by GPT 4. The limitations of GPT 4 include its paid subscription model, slower response times and a restriction that allows it to answer only 25 questions within a 3-hour period. However, most studies [3, 4, 8, 9] use ChatGPT 3.5 when referring to ChatGPT.

All questions were assessed in Portuguese and English. GPT 3.5 and 4 responded without a significant difference based on the idiom employed. Despite there being no statistically significant differences, a notable proportion of responses in Portuguese were deemed harmful compared to those in English. This observation prompts consideration of ChatGPT's training data, which appears to have a bias towards English content. This potential discrepancy in training data distribution may contribute to the higher prevalence of dangerous responses in Portuguese.

To our knowledge, this is the first report testing the GPT 3.5 and 4 performance by answering the same questions in different languages in the healthcare or medical field. Despite GPT 4 having improved multilingual capabilities, ChatGPT 3.5 can also understand multiple languages.

The percentage of responses categorized as correct was 57.5% for GPT 3.5 and 83.8% for GPT 4. Despite not showing a significant difference between the two models, GPT 3.5 and GPT 4 exhibited good performance, as they could correctly answer more than half of the questions. This aligns with what was previously published by Samaan et al. [3].

The authors evaluated ChatGPT performance using a similar system as the present study suggests. They gathered from health institutions and Facebook patient support groups questions related to bariatric surgery, and reviewers were instructed to grade as comprehensive (defined as accurate and comprehensive, nothing more a board-certified bariatric surgeon can add if asked this question by a patient), correct but inadequate, some correct and some incorrect and completely incorrect. This study has demonstrated that ChatGPT can provide comprehensive answers to over 86% of all questions, with higher scores in areas such as eligibility, efficacy, procedural options, and preoperative preparation.

Similarly, Johnson et al. [10] evaluated the performance of ChatGPT in answering cancer-related questions that are a common point of confusion among the public. They found that ChatGPT's outputs provide accurate responses like those provided by official support platforms for cancer patients. However, Samaan et al. [3] noticed that this language model can provide incorrect information, sometimes with a false sense of confidence, which could be dangerous for patients without the guidance of a healthcare provider. Correspondingly, this study reached similar conclusions. For example, when asked to ChatGPT 3.5 about the stimulation settings of the system (question number 57), the model suggested that the doctor will provide instructions on how to adjust the stimulation settings device based on specific symptoms and response to the therapy, given the false sense that the patient is capable and responsible for adjusting the stimulation with autonomy. On the other hand, ChatGPT 4 was pre-emptive, saying that the patient mustn't attempt adjusting the settings themselves, as this requires specialized knowledge and expertise.

Vaishya et al. [12] discuss the importance of considering the variability of physical and psychological characteristics of each patient, making the diagnosis and management of their condition a challenging task for physicians. Taking this into account, the authors raised concerns regarding the accuracy of ChatGPT's responses to medical questions, situations requiring access to up-to-date and reliable data and the potential for harmful instructions or biased content. Identical concerns were raised regarding a question unanimously categorized as harmful. When asked to ChatGPT 3.5 in Portuguese (question number 64) about the detection by metal detectors of the battery and electrodes, the model suggested that there is no reason to be concerned about metal detectors, as the batteries and electrodes are typically made from materials that metal detectors cannot detect. Furthermore, DBS components are generally placed in the body to avoid interference with metal detectors.

Deik [13] intriguingly asks, is ChatGPT the new Dr. Google? As these models become more widely available, patients will seek information about their care from these

sources, and they will easily access these tools to obtain answers to their clinical queries. As Yeo et al. [4] reported, this increased accessibility can also reduce anxiety among patients and caregivers by providing access to real-time information. Patients will be better informed about their condition, which could reduce unnecessary anxiety between doctor appointments. It is crucial for healthcare providers to understand the strengths and limitations of these models, so they can give accurate advice to their patients and exercise caution. Lined up, D'amico et al. [14] and Deik [13] enumerated the concerns of using this type of language model, with a focus on concerns related to privacy and ethics, potential biases, legal responsibilities and liabilities, validity of data, accuracy and effectiveness. Special consideration is necessary to ensure the safety of patients and to avoid potential risks from their use [15]. The text generated by ChatGPT should be submitted for rigorous checking of the content to ensure the absence of errors, bias, or harmful information. As Deik [13] emphasizes, patients should be informed that the tool is not a replacement for medical care. Especially in movement disorders, there are situations where the symptoms are unclear, and in such instances, the expertise of a trained movement specialist exceeds what the tool can provide.

In this study, we noticed that the answers given in English by ChatGPT model 4 always start with the following detail: "I am not a doctor", in contrast with ChatGPT model 3.5. Moreover, all answers include the exception that the patient should discuss the doubts with the attending physician. We agree with this Editorial [15], suggesting that the application itself may alert patients to the risk of information which lacks the opinion of an expert clinical consultation.

As a tool for clinicians, ChatGPT and other language models in healthcare have potential, from text generation [16] to improved data extraction and clinical decision-making [13]. In the field of movement disorders, Deik [13] reviewed the potential benefits of this language model. He concluded that expediting administrative work, creating patient education materials, synthesizing data, and broadening the differential diagnosis could provide significant added value for clinicians dedicated to this area. Boßelmann et al. [17] admitted the potential of this language in epilepsy, mainly outside of the clinical care of patients, as a tool for smart data processing, content generation and sentiment analysis. In the field of surgery, Hassan et al. [18] recognized the ability of ChatGPT to provide relevant insights to questions related to patient care. However, once again, although artificial intelligence can function as a tool for addressing patients' inquiries, it cannot supplant the specialized knowledge and expertise inherent in clinical practice.

ChatGPT lacks the ability to differentiate between context-specific guidelines related to the same topic. As a result,

it may generate responses that don't necessarily align with the medical practices of a particular region, medical society, or country where the user resides. This situation could lead to confusion and potential risks if a patient were to adhere to recommendations not approved by local regulatory bodies.

The current study had some limitations. Firstly, the study examined numerous questions. However, it's important to note that the selected questions may not cover all relevant patient inquiries. Furthermore, the focus is mainly on neurosurgical management. Questions concerning neurological issues, such as motor and non-motor symptoms and disease progression, could potentially and hypothetically lead to more harmful answers. This is because there is limited scientific evidence regarding these issues. Therefore, the inclusion of these neurological questions would detract from the perceived effectiveness of ChatGPT in addressing these queries.

Secondly, the low agreement among the raters could be explained by the fact that MC and BC are two experts in DBS surgery for Parkinson's; however, MC is a Neurologist, and BC is a Neurosurgeon. Therefore, considering that this classification is subjective and based on the clinical experience of each of the expert, the classification may have been directed more towards the medical component or the surgical component, depending on the expert. This fact highlights the need for close collaboration between the speciality of Neurology and Neurosurgery in the management of these patients. In addition, this observation suggests a potential flaw within the grading system that extends beyond the difference in expertise backgrounds. Recognizing the potential for bias inherent in grading systems, particularly in subjective evaluations, the authors took proactive steps to address this concern. The grading system utilized in this study was developed after a review of relevant literature, which also employed this grading methodology in evaluating the performance of ChatGPT [3, 4]. Nevertheless, we acknowledge that the available systems are still imperfect and prone to potential biases. In this way, we believe that the low agreement among raters could be partially explained by the subjective grading system. Hence, continuous optimization of these assessment tools by the scientific community are necessary to minimize any biases and improve the design of currently described systems. For that reason, refining the grading model to include more objective criteria together with the development of guidelines could help standardize assessments and reduce subjectivity. Additional factors could influence the interpretation of assessment criteria, such as varying degrees of familiarity with the grading system and individual biases influencing the scoring process. Providing comprehensive training and calibration sessions for raters to ensure a shared understanding of assessment criteria and methodologies could enhance consistency.

Another important bias to consider is that the raters know in advance about whether they are categorizing GPT-3.5 or GPT-4, rather than making classifications blindly. This bias has the potential to lead to an overestimation of GPT-4's performance. Moreover, despite the three reviewers having an advanced level of English proficiency, Portuguese is the native language, and this could potentially impact the categorization process. Finally, each question was input only once into ChatGPT, and as such, the reproducibility of responses was not assessed. Nevertheless, this approach mirrors real-world situations where questions are commonly asked one at a time. Additionally, based on previous research [3, 4], reproducibility surpasses 90% when the same questions are posed at least twice.

Further research is required to thoroughly investigate the effectiveness of ChatGPT in DBS patient education, as well as to monitor enhancements in the precision and consistency of its responses to patient inquiries. Nowadays, ChatGPT is solely accessible as a research resource and is not approved as a medical resource.

Conclusion

The evolution of artificial intelligence in medicine must be carefully managed to ensure its benefits outweigh its potential risks. The inclusion of chatbots in medical practice seems to be a closer reality. In general, ChatGPT model 3.5 and model 4 demonstrated a good performance in terms of clarity; however, the harmful answers are not negligible, and it is important to take that into account when dealing with patients who used these tools to answer their questions. Particularly, this finding raises significant concerns regarding the suitability of ChatGPT for unsupervised patient guidance in medical contexts such as DBS surgery. While this advancement poses difficulties for the medical community, thorough investigations and studies can aid in mitigating potential risks.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10072-024-07732-0>.

Author contributions Concept and design: Ana Lúcia Oliveira, Pedro Duarte-Batista. Acquisition of data: Ana Lúcia Oliveira, Pedro Duarte-Batista. Data review: Ana Lúcia Oliveira. Statistical analysis: Ana Lúcia Oliveira, Pedro Duarte-Batista. Drafting of the manuscript: Ana Lúcia Oliveira, Pedro Duarte-Batista. Critical revision of the manuscript: All authors. All authors read and approved the final manuscript.

Funding No funding was received to assist with the preparation of this manuscript. Open access funding provided by FCT|FCCN (b-on).

Data availability All data generated or analyzed during this study are

included in this article and its supplementary material files. Further enquiries can be directed to the corresponding author.

Declarations

Ethics approval and consent to participate This article does not contain any studies with human participants or animals performed by any of the authors. As all the responses from ChatGPT were accessible to the public.

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Biswas S (2023) ChatGPT and the future of Medical writing. *Radiology* 307(2):e223312. <https://doi.org/10.1148/radiol.223312>
2. OpenAI ChatGPT: Optimizing Language Models for Dialogue
3. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH et al (2023) Assessing the accuracy of responses by the Language Model ChatGPT to questions regarding bariatric surgery. *Obes Surg* 33(6):1790–1796. <https://doi.org/10.1007/s11695-023-06603-5>
4. Hui Yeo Y, Samaan JS, Han Ng W et al (2023) Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 29(3):721–732. <https://doi.org/10.3350/cmh.2023.0089>
5. Cascella M, Montomoli J, Bellini V et al (2023) Evaluating the feasibility of ChatGPT in Healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 47:33. <https://doi.org/10.1007/s10916-023-01925-4>
6. Sevgi UT, Erol G, Doğruel Y et al (2023) The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. *Neurosurg Rev* 46(1):86. <https://doi.org/10.1007/s10143-023-01998-2>
7. Sng GGG, Tung JYM, Lim DYZ et al (2023) Potential and pitfalls of ChatGPT and Natural-Language Artificial Intelligence Models for Diabetes Education. *Diabetes Care* 46(5):e103–e105. <https://doi.org/10.2337/dc23-0197>
8. Ayers JW, Poliak A, Dredze M et al (2023) Comparing physician and Artificial Intelligence Chatbot responses to patient questions posted to a Public Social Media Forum. *JAMA Intern Med* 183(6):589–596. <https://doi.org/10.1001/jamainternmed.2023.1838>
9. Hopkins AM, Logan JM, Kichenadasse G et al (2023) Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr* 7(2). <https://doi.org/10.1093/jncics/pkad010>

10. Johnson SB, King AJ, Warner EL et al (2023) Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr* 7(2). <https://doi.org/10.1093/jncics/pkad015>
11. Ali R, Tang OY, Connolly ID, Neurosurgery et al (2023) <https://doi.org/10.1227/neu.0000000000002551>
12. Vaishya R, Misra A, Vaish A (2023) ChatGPT: is this version good for healthcare and research? *Diabetes Metab Syndr* 17(4):e102744. <https://doi.org/10.1016/j.dsx.2023.102744>
13. Deik (2023) Potential benefits and perils of incorporating ChatGPT to the Movement Disorders Clinic. *J Mov Disord* 16(2):158–162. <https://doi.org/10.14802/jmd.23072>
14. DeepEOR: automated perioperative volumetric assessment of variable grade gliomas using deep learning. *Acta Neurochirurgica* 165(2):555–66; <https://doi.org/10.1007/s00701-022-05446-w>
15. No authors listed (2023) Will ChatGPT transform healthcare? Editorial. *Nat Med* 29(3):505–506. <https://doi.org/10.1038/s41591-023-02289-5>
16. Janssen BV, Kazemier G, Besselink MG (2023) The use of ChatGPT and other large language models in surgical science. *BJS Open* 7(2):zrad032. <https://doi.org/10.1093/bjsopen/zrad032>
17. Boßelmann CM, Leu C, Lal D (2023) Are AI language models such as ChatGPT ready to improve the care of individuals with epilepsy? *Epilepsia* 64(5):1195–1199. <https://doi.org/10.1111/epi.17570>
18. Hassan AM, Nelson JA, Coert JH et al (2023) Exploring the potential of Artificial Intelligence in surgery: insights from a conversation with ChatGPT. *Ann Surg Oncol* 30(7):3875–3878. <https://doi.org/10.1245/s10434-023-13347-0>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.