Current Research in MRI

Original Article

Large Language Models on Magnetic Resonance Imaging Safety-Related Questions: Accuracy of ChatGPT-3.5, ChatGPT-4, Gemini, and Perplexity

Esat Kaba, Hande Melike Bülbül, Gülen Burakgazi, Merve Solak, Serdar Tabakoğlu, Ayşenur Topçu Varlık, Nur Hürsoy, Fatma Beyazal Çeliker

Department of Radiology, Recep Tayyip Erdogan University Faculty of Medicine, Rize, Turkey

Cite this article as: Kaba E, Bülbül HM, Burakgazi G, *et al*. Large language models on magnetic resonance imaging safety-related questions: accuracy of ChatGPT-3.5, ChatGPT-4, Gemini, and Perplexity. *Current Research in MRI* 2024;3(1):16-19.

Corresponding author: Esat Kaba, e-mail: esatkaba04@gmail.com

Received: March 13, 2024 **Revision requested:** April 3, 2024 **Last revision received:** April 4, 2024 **Accepted:** April 7, 2024 **Publication Date:** April 22, 2024

DOI:10.5152/CurrResMRI.2024.24095

Abstract

Objective: This study investigates the accuracy of large language models (LLMs) on magnetic resonance imaging (MRI) safety-related questions.
Methods: Three experienced radiologists independently prepared 20 multiple-choice questions based on the MRI safety guidelines published by the Turkish Magnetic Resonance Society. An initial prompt was entered into 4 different LLMs (ChatGPT-3.5, ChatGPT-4, Gemini, and Perplexity) and then a total of 60 questions were asked. The answers received were compared with the answers assigned by the radiologists according to guidelines. The performance of each model was obtained as accuracy.
Results: In 60 questions, the accuracy rates were 78.3% (47/60) for ChatGPT-3.5, 93.3% (56/60) for ChatGPT-4, 88.3% (53/60) for Gemini, and 86.7% (52/60) for Perplexity. In addition, ChatGPT-3.5 answered 19/20, 13/20, and 15/20, ChatGPT-4 answered 18/20, 18/20, and 20/20, Gemini answered 19/20, 18/20, and 16/20, and Perplexity answered 20/20, 15/20, and 17/20 correctly to question groups prepared by 3 radiologists, respectively.
Conclusion: Large language models, particularly the most stable and highest performing ChatGPT-4, may be useful to patients and health-care professionals in providing MRI safety-related information. They have the potential to assist in the future to protect health-care professionals and patients from MRI-related accidents.

Keywords: Large language model, MRI safety, questions, accuracy

INTRODUCTION

Large language models (LLMs) are a component of generative artificial intelligence (AI) that are trained with a lot of data and focus on the interaction between humans and computer language.[1] Such models are developed through unsupervised training where they learn the structures, patterns, and relationships of language by analyzing large amounts of textual data.[2] Training with a wide variety and size of data has significantly improved their ability to better understand human language and generate human-like text. After OpenAI launched ChatGPT in November 2022, it became very popular and reached millions of users quickly. Due to their high text analysis capabilities, their potential use in many fields of medicine is being explored.[3]

Potential uses of LLMs in radiology include topics such as radiology report generation, report structuring, report simplification, and radiological protocol determination.[4] In addition, studies testing the accuracy and reliability of LLMs' knowledge of radiology-related topics and evaluating their performance in creating patient educational materials have been published.[5] Patil et al[6] analyzed the answers to 318 questions in neuroradiology, general and physics, pediatric radiology, ultrasound, and nuclear medicine and compared the performance of ChatGPT and Bard. In this study, the accuracy rate of ChatGPT was remarkable at 87.11%. Large language models are also promising for facilitating patients' understanding of complicated radiologic terminology and providing summary reports to patients. One study investigated the accuracy of LLMs in summarizing full magnetic resonance imaging (MRI) reports of cancer patients and found satisfactory results.[7] However, as indicated in these studies, LLMs have some important limitations and much larger studies are needed to demonstrate the reliability of their potential use.[6,7]

Magnetic resonance imaging is a valuable technique in radiology due to its high soft tissue resolution and its absence of ionizing radiation, and its use is increasing worldwide.[8] It generates an electromagnetic force 30-60 thousand times stronger than the magnetic field strength of the earth.[9] This strong magnetic field can lead to fatal situations if safety rules are not followed.[10] Therefore, patients and health-care professionals must have adequate knowledge about MRI safety.

In this study, we tested the accuracy and reliability of LLMs on MRI safety to investigate their potential future usability by patients and health-care professionals.

MATERIAL AND METHODS

In this study, 3 radiologists with 19, 14, and 8 years of radiology experience, respectively, independently prepared 20 multiple-choice questions (questions group 1, 2, and 3) related to MRI safety. All questions were based on the MRI safety guidelines of the Turkish Magnetic Resonance Society (TMRD) (https://tmrd.org.tr/uploads/files/tmrd-mr-klavuzu.pdf). A total of 60 questions were created, and all questions were reviewed by another radiologist with 5 years of experience. Four sample questions and their options are shown in Table 1. Since no human or animal subjects were used, ethical approval or informed consent was not required.

The initial prompt was then entered into ChatGPT-3.5 and 4 (https://chat.openai.com/), Gemini, (https://gemini.google.com/app), and Perplexity (https://www.perplexity.ai/) chatbots. All questions were entered into 4 different language models with default parameters in March 2024. All questions and options are entered in English. Role-modeling technique was used in prompting. The initial prompt entered into the chatbots is as follows;

Initial Prompt

"As a highly experienced radiologist with 25 years of experience, answer these questions about magnetic resonance imaging (MRI) safety, there is only one correct answer."

The results were then analyzed and compared with the correct answers assigned by the radiologists according to guidelines. The flowchart of our study is shown in Figure 1.

RESULTS

A total of 60 questions prepared by 3 radiologists were asked to 4 different LLMs. The accuracy rates were 78.3% (47/60) for ChatGPT 3.5, 93.3% (56/60) for ChatGPT-4, 88.3% (53/60) for Gemini, and 86.7% (52/60) for Perplexity. In other words, ChatGPT-4 performed the best compared to other LLM models, answering 56 out of 60 questions correctly.

ChatGPT-3.5 answered 19/20, 13/20, 15/20, ChatGPT-4 18/20, 18/20, 20/20, Gemini 19/20, 18/20, 16/20, and Perplexity 20/20, 15/20, 17/20 correctly to 3 radiologists' questions, respectively (Table 2).

The comparative graph of LLMs' answers to radiologist question groups is given in Figure 2.

MAIN POINTS

- In this study, we analyzed the responses of large language models (LLMs) to magnetic resonance imaging (MRI) safety-related questions.
- ChatGPT-4 outperformed the other LLMs by answering 56 out of 60 multiple-choice questions correctly, with an accuracy of 93.3%.
- The accuracy rates of Gemini, Perplexity, and ChatGPT-3.5 are 88.3%, 86.7%, and 78.3%, respectively.
- Large language models can potentially assist patients and health-care professionals in MRI safety, as in many domains of radiology.

Table 1. Four Sample Questions

1. Which of the following can cause strong gravitational or ejection effects (missile or projectile effect) and related injuries?

a. Static magnetic field
b. Gradient magnetic field
c. Radiofrequency energy

2. In which MRI security zone is the MRI device located?

a. Zone 1
b. Zone 2
c. Zone 3
d. Zone 4

3. Which gas is discharged during the quench process in MRI?

a. Nitrogen
b. Hydrogen
c. Helium
d. Oxygen

4. Which of the implants and materials implanted in the human body is not classified as active in terms of MRI safety?

a. Aneurysm clips
b. Pacemaker
c. Implantable cardioverter defibrillator
d. Neurostimulation system

MRI, magnetic resonance imaging.

The correct and incorrect answers of 4 different LLMs to a total of 60 questions are visualized in Figure 3.

DISCUSSION

In this study, we investigated the accuracy of LLMs in MRI safety-related questions. Three different radiologists independently prepared a total of 60 questions based on the TMRD MRI safety guidelines. ChatGPT-3.5 and 4, Gemini, and Perplexity chatbots were used with default parameters. All questions were presented to LLMs with a role-modeling initial prompt technique. The answers were compared to the gold standard of correct answers. As a result, ChatGPT-4 showed the highest performance with 93.3% accuracy, answering 56 out of 60 questions correctly. ChatGPT-3.5 gave the correct answer to 47 out of 60 questions and showed the lowest performance with 78.3% accuracy. Gemini and Perplexity performed competitively with 88.3% and 86.7%, respectively.

In the field of natural language processing, LLMs, which are constantly making groundbreaking advances, are models for understanding, designing, reconstructing, and processing text.[11] Since November 2022, many companies have increasingly made continuously updated LLMs available to the general public. Although LLMs are trained for human-like speech, their perspective has broadened as the application has developed. Aiming to maximize productivity in health care and medical fields as in many industrial fields, these LLMs promise potential areas of use in radiology, which is directly affected by high technological developments. On the other hand, many radiology-based studies testing the accuracy and reliability of these LLMs, which have some concerns and limitations, have been published.[12,13]

Lee et al[14] investigated the accuracy of ChatGPT's answers to MRI-related questions.[14] In this study, the authors asked 50 simple MRI-related questions and categorized the answers as correct,

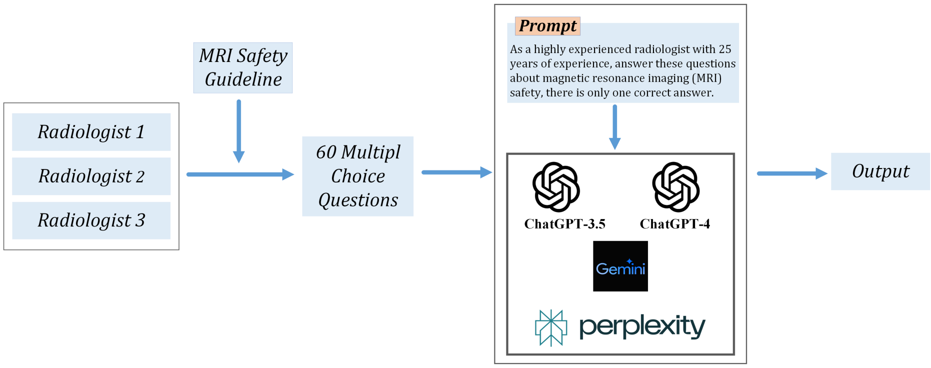
Radiologist 1
Radiologist 2
Radiologist 3
MRI Safety Guideline
60 Multipl Choice Questions
Prompt
As a highly experienced radiologist with 25 years of experience, answer these questions about magnetic resonance imaging (MRI) safety, there is only one correct answer.
ChatGPT-3.5
ChatGPT-4
Gemini
perplexity
Output

Figure 1. The flowchart of this study.

Table 2. Accuracy of Large Language Models

	Accuracy (Questions Group 1)	Accuracy (Questions Group 2)	Accuracy (Questions Group 3)	Accuracy (Total)
ChatGPT-3.5	95% (19/20)	65% (13/20)	75% (15/20)	78.3% (47/60)
ChatGPT-4	90% (18/20)	90% (18/20)	100% (20/20)	93.3% (56/60)
Gemini	95% (19/20)	90% (18/20)	80% (16/20)	88.3% (53/60)
Perplexity	100% (20/20)	75% (15/20)	85% (17/20)	86.7% (52/60)

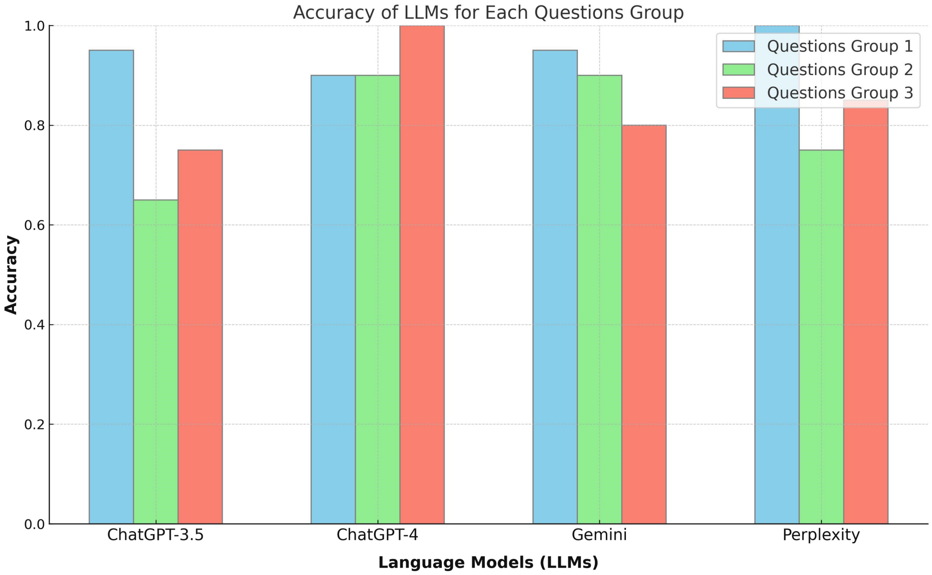

Figure 2. Comparative graph of large language models' answers to radiologist question groups.

partially correct, and incorrect. In addition, they asked 75 multiple-choice questions on basic electromagnetism, MR magnets, gradients, radiofrequency and coils, and site planning, and the answers were analyzed by independent researchers. As a result, ChatGPT answered the 50 questions in the first step 86% and 88% correctly according to the 2 observers, respectively. The 75 multiple-choice questions in the second step were answered correctly between 40% and 66.7% depending on the topic. In our study, ChatGPT-3.5 answered 95%, 65%, and 75% correctly to the question groups prepared by 3 radiologists based on MR safety guidelines, respectively. ChatGPT-4 answered 90%, 90%, and 100% of the questions correctly, respectively. From this point of view, ChatGPT-4 is more successful and shows a more balanced accuracy value than the ChatGPT-3.5 version.

Magnetic resonance imaging is one of the most important radiology examinations without ionizing radiation and provides a high level of anatomical detail and functional imaging.[15] It has strong static and gradient magnetic fields and radiofrequency energy. The strong static magnetic field causes a ferromagnetic effect and when used inappropriately, metal materials in the patient's body or the acquisition room can cause fatal injuries.[16] It also has the potential for many safety hazards, including tissue burns, nerve stimulation, acoustic noise, hearing loss, and contrast media complications.[17] There are also some rules to be considered for pregnant, children, and claustrophobic patients.[18] For all these reasons, radiologists, MRI technicians, and patients should have detailed information to prevent fatal errors. At this point, LLMs can benefit health-care professionals and patients regarding MRI safety in the future. If they can be securely integrated into hospital systems in

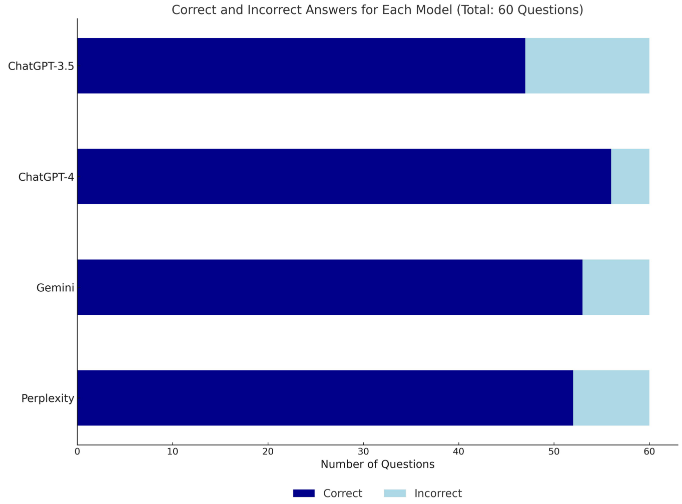

Figure 3. The correct and incorrect answers of 4 different large language models to a total of 60 questions.

the future, an application can be created that can detect the presence of any metal (e.g., pacemaker, aneurysm clip, orthopedic prosthesis) in the patient's body from the patient history and alert the technician. Thus, some potential errors that may occur under a heavy workload can be prevented. In addition, patients can access answers to their questions about MRI safety in a simplified way through LLMs. According to the results of our study, ChatGPT-4, which shows the highest performance with 93.3% accuracy, is promising for this purpose. However, LLMs, which have limitations such as timeliness problems, ethical concerns, confidentiality, bias, and hallucination generation, need much wider validation to be used in this field.[19,20]

There are several limitations of our study. First, the number of questions prepared was small. This is because the TMRD MRI safety guideline was used. This guideline is summarized. Large language models can be tested with a wider variety of questions than more comprehensive guidelines in the literature. Second, the questions were only prepared and entered into the LLMs in English. In the future, the performance of LLMs in different languages could be compared. Third, it should be noted that ChatGPT-4 is a paid LLM, and this may create limitations in its widespread use. Finally, the questions had a multiple-choice format with only 1 correct answer. In future larger studies, open-ended questions could be asked, and the answers provided by the LLMs could be analyzed by experienced radiologists.

In conclusion, although MRI is generally safe, it is a modality that can cause accidents with fatal consequences if several safety rules are not followed. Large language models can potentially provide information and decision support to radiologists, technicians, and patients on MRI safety, as in many areas of radiology. Furthermore, given the widespread use of LLMs in the community, they have the potential to increase the general population's level of knowledge about MRI safety.

Ethics Committee Approval: N/A.

Informed Consent: N/A.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – E.K., H.M.B., G.B., M.S., S.T., A.T.V., N.H., F.B.Ç.; Design – E.K., H.M.B., G.B., M.S.; Supervision – E.K., H.M.B., G.B., M.S., F.B.Ç.; Data Collection and/or Processing – E.K., H.M.B., G.B., M.S., S.T., A.T.V., N.H., F.B.Ç.; Materials – E.K., HMB, G.B., M.S., S.T., A.T.V., N.H., F.B.Ç.; Analysis and/or Interpretation – E.K., H.M.B., G.B., M.S., S.T., A.T.V., N.H., F.B.Ç.; Literature Search – E.K., H.M.B., G.B., M.S., S.T., A.T.V., N.H., F.B.Ç.; Writing Manuscript – E.K., H.M.B., G.B., M.S., S.T., A.T.V., N.H., F.B.Ç. Critical Review – E.K., H.M.B., G.B., M.S., S.T., A.T.V., N.H., F.B.Ç.

Declaration of Interests: The authors have no conflict of interest to declare.

Funding: The authors declared that this study has received no financial support.

REFERENCES

1. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *J Med Syst*. 2024;48(1):22. [CrossRef]
2. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. *Nat Rev Phys*. 2023;5(5):277-280. [CrossRef]
3. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9:e48291. [CrossRef]
4. Kim S, Lee CK, Kim SS. Large language models: A guide for radiologists. *Korean J Radiol*. 2024;25(2):126-133. [CrossRef]
5. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing patient communication with chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J Am Coll Radiol*. 2024;21(2):353-359. [CrossRef]
6. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Can Assoc Radiol J*. 2023:8465371231193716. [CrossRef]
7. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit Health*. 2023;9: 20552076231221620. [CrossRef]
8. Sammet S. Magnetic resonance safety. *Abdom Radiol (NY)*. 2016;41(3): 444-451. [CrossRef]
9. Panych LP, Madore B. The physics of MRI safety. *J Magn Reson Imaging*. 2018;47(1):28-43. [CrossRef]
10. Dempsey MF, Condon B, Hadley DM. MRI safety review. *Semin Ultrasound CT MR*. 2002;23(5):392-401. [CrossRef]
11. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J Med Syst*. 2024 Feb 17;48(1):22. doi: [CrossRef]
12. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology*. 2023;307(5):e231259. [CrossRef]
13. Kaba E. Zero-, Single-, and Few-Shot Learning in Large language models to identify incidental findings from radiology reports. In: *AJR Am J Roentgenol*. 2024:1. [CrossRef]
14. Lee KH, Lee RW. ChatGPT's accuracy on magnetic resonance imaging basics: characteristics and limitations depending on the question type. *Diagnostics (Basel)*. 2024;14(2):171. [CrossRef]
15. Starekova J, Hernando D, Pickhardt PJ, Reeder SB. Quantification of liver fat content with CT and MRI: state of the art. *Radiology*. 2021;301(2):250-262. [CrossRef]
16. Tsai LL, Grant AK, Mortele KJ, Kung JW, Smith MP. A practical guide to MR imaging safety: what radiologists need to know. *RadioGraphics*. 2015;35(6):1722-1737. [CrossRef].
17. Cross NM, Hoff MN, Kanal KM. Avoiding MRI-related accidents: A practical approach to implementing MR safety. *J Am Coll Radiol*. 2018;15(12):1738-1744. [CrossRef]
18. Expert Panel on MR Safety, Kanal E, Barkovich AJ, et al. ACR guidance document on MR safe practices: 2013. *J Magn Reson Imaging*. 2013;37(3):501-530. [CrossRef]
19. Sallam M, Chat GPT. ChatGPT Utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887. [CrossRef]
20. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology- a recent scoping review. *Diagn Pathol*. 2024;19(1):43. [CrossRef]