European Journal of Radiology 193 (2025) 112416

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Research article

Chat GPT-4 shows high agreement in MRI protocol selection compared to board-certified neuroradiologists

Zeynep Bendella [a,b,*,1], Barbara Daria Wichtmann [a,b,1], Ralf Clauberg [a], Vera C. Keil [c], Nils C. Lehnen [a,b], Robert Haase [a,b], Laura C. Sáez [c,d], Isabella C. Wiest [e,f], Jakob Nikolas Kather [f], Christoph Endler [g], Alexander Radbruch [a,b], Daniel Paech [a,h,2], Katerina Deike [a,i,2]

[a] Clinic of Neuroradiology, University Hospital Bonn, Bonn, Germany
[b] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[c] Department of Radiology & Nuclear Medicine, Amsterdam UMC, Vrije Universiteit, Cancer Center Amsterdam, Amsterdam, the Netherlands
[d] Hospital Universitario Son Llátzer (HUSLL), Palma, Mallorca, Spain
[e] Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[f] Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany
[g] Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Bonn, Germany
[h] Department of Radiology, Brigham and Womeńs Hospital, Harvard Medical School, Boston, MA, USA
[i] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

ARTICLE INFO

Keywords:
ChatGPT-4
Large language model (LLM)
MRI protocol
Radiology request form

ABSTRACT

Objectives: The aim of this study was to determine whether ChatGPT-4 can correctly suggest MRI protocols and additional MRI sequences based on real-world Radiology Request Forms (RRFs) as well as to investigate the ability of ChatGPT-4 to suggest time saving protocols.

Material & methods: Retrospectively, 1,001 RRFs of our Department of Neuroradiology (in-house dataset), 200 RRFs of an independent Department of General Radiology (independent dataset) and 300 RRFs from an external, foreign Department of Neuroradiology (external dataset) were included. Patients' age, sex, and clinical information were extracted from the RRFs and used to prompt ChatGPT- 4 to choose an adequate MRI protocol from predefined institutional lists. Four independent raters then assessed its performance. Additionally, ChatGPT-4 was tasked with creating case-specific protocols aimed at saving time.

Results: Two and 7 of 1,001 protocol suggestions of ChatGPT-4 were rated "unacceptable" in the in-house dataset for reader 1 and 2, respectively. No protocol suggestions were rated "unacceptable" in both the independent and external dataset. When assessing the inter-reader agreement, Coheńs weighted κ ranged from 0.88 to 0.98 (each $p < 0.001$).

ChatGPT-4′s freely composed protocols were approved in 766/1,001 (76.5 %) and 140/300 (46.67 %) cases of the in-house and external dataset with mean time savings (standard deviation) of 3:51 (minutes:seconds) (±2:40) minutes and 2:59 (±3:42) minutes per adopted in-house and external MRI protocol.

Conclusion: ChatGPT-4 demonstrated a very high agreement with board-certified (neuro-)radiologists in selecting MRI protocols and was able to suggest approved time saving protocols from the set of available sequences.

Abbreviations: AI, Artificial Intelligence; ChatGPT-4, Chat Generative Pre-trained Transformer 4; LLM, large language model; MRI, magnetic resonance imaging; NLP, Natural Language Processing; RRF, radiology request form.

* Corresponding author at: Clinic of Neuroradiology, University Hospital Bonn, Rheinische-Friedrich-Wilhelms-Universität Bonn, Venusberg-Campus 1, 53127 Bonn, Germany.

E-mail address: zeynep.bendella@ukbonn.de (Z. Bendella).

[1] These authors contributed equally to this work.

[2] These authors jointly supervised this work.

https://doi.org/10.1016/j.ejrad.2025.112416
Received 4 June 2025; Received in revised form 7 August 2025; Accepted 10 September 2025
Available online 13 September 2025

1. Introduction

Selecting the appropriate magnetic resonance imaging (MRI) protocol to elucidate a patient's clinical diagnosis poses a significant challenge for radiologists, particularly for less experienced radiologists [1,2]. It is essential to choose relevant sequences to aid in diagnosing and initiating the correct therapy. This process can be time-consuming and is also subject to inter-reader variability, as different radiologists may have varying preferences for specific sequences [3]. The Radiology Request Form (RRF) plays a critical role in the radiological workflow by facilitating communication between referring doctors and radiologists, thus enabling the selection of suitable radiological examinations and scanning protocols [4].

However, conducting unnecessary or incomplete examinations can lead to inefficient workflows, potential misdiagnoses, and treatment delays. As different MRI sequences highlight various aspects of the examination area, selecting the right combination is crucial to balance sensitivity to specific changes while minimizing sensitivity to others [5,6].

Brown and Marotta evaluated three Natural Language Processing (NLP) models −random forest, support vector machine, and k-nearest neighbor – for automating key tasks in the brain MRI workflow: protocol selection, deciding on the necessity of an intravenous contrast agent, and examination prioritization. Their models achieved accuracies of 82.9 %, 83.0 %, and 88.2 % for each respective task. Accuracy was defined as the proportion of correctly classified cases in their test dataset [7].

Furthermore, Kalra et al. developed NLP-based machine learning models for protocol assignment across various anatomical areas and imaging modalities, achieving a precision range of 76 % to 82 % in a dataset of over 18,000 CT and MRI scans [8]. This study supports the feasibility of classical NLP-driven automation in radiology protocol selection, highlighting the potential of Artificial Intelligence (AI)-based clinical decision support tools.

However, the potential of AI-based Large Language Models (LLMs), which are trained on extensive text data to recognize and simulate complex language patterns, has not been comprehensively surveyed in a clinical context [9,10].

Released in spring 2023, Chat Generative Pre-trained Transformer 4 (ChatGPT-4) is an example of such an LLM, based on the widely adopted transformer architecture. It has undergone extensive training across multiple languages enabling it to generate responses that closely mimic human-like interactions with text input [11,12]. The Generative Pre-trained Transformer (GPT) architecture of ChatGPT-4 utilizes a neural network with self-attention mechanisms to process natural language and produce contextually relevant responses to the input text [13–17]. Consequently, ChatGPT has been investigated for its value in supporting especially the following areas in radiology: process optimization, report generation, education of medical staff and differential diagnosis [18–27]. In the area of diagnostic imaging selection, which can be seen as an intersection between process optimization and differential diagnosis, Barash et al. evaluated whether ChatGPT-4 could recommend appropriate imaging studies for eight typical diagnoses in the emergency department, such as appendicitis, diverticulitis, and pulmonary embolism [28]. They provided ChatGPT-4 the clinical information of the patients and found that ChatGPT-4 succeeded in 38 out of 40 cases. Furthermore, Gertz et al. reported an 84 % agreement between ChatGPT-4s choice of imaging modality, contrast agent application and acquisition of contrast enhancement phases compared to the reference standard when providing real-world RRFs [22].

To further analyze the value of ChatGPT in process optimization, clinical decision- making in MRI, and the training of medical staff, this study aimed to evaluate the performance of ChatGPT-4 in selecting correct MRI protocols and sequences based on clinical information, age, and sex of patients, as extracted from real-world RRFs. Additionally, this study investigated whether ChatGPT can independently compose appropriate MRI protocols and thereby save time.

2. Materials and Methods

2.1. Ethical approval

The retrospective study received approval from the local Ethics Committee for Clinical Trials on Humans and Epidemiological Research with Personal Data, IRB number: 312/23-EP. No patient-identifying information was supplied to ChatGPT-4.

2.1.1. Datasets

Data acquisition was conducted between August 2023 and July 2024. During this period, a total of n = 1,001, 200 and 300 consecutive original in- and outpatient RRFs were retrospectively collected through a review of the radiologic information system. Specifically, 1,001 RRFs were obtained from the in-house department of neuroradiology (in-house dataset), 200 from the independent department of general radiology of our university medical center (independent dataset) and 300 from an entirely external department of neuroradiology (external dataset). The in-house dataset covered the full range of institutional neuroradiological MRI protocols, as detailed in Supplementary Table 1. All available RRFs during the defined study period were included without further selection. The independent and external datasets were used to assess generalizability across different clinical settings and therefore did not reflect the complete protocol spectrum.

Table 1 gives an overview of the three different datasets included in this study.

Clinical indications were manually extracted from the RRFs by experienced (neuro-) radiologists (readers) for each dataset. No standardization or pre-processing was applied beyond the removal of patient-identifying information, allowing ChatGPT to work directly with real-world referral language.

2.1.2. ChatGPT-4 and Prompt engineering

ChatGPT-4 was accessed through the web interface. During the study period, only a free and a paid version existed, and ChatGPT-4 was available exclusively in the paid version.

Prompt engineering was was performed on a small in-house dataset (n = 30) originally in German and only translated into English for the purposes of this paper using ChatGPT-4. No protocol- or sequence-specific tuning was performed. Except for these 30 in-house cases used to optimize the prompt/instructions provided to ChatGPT, no further modifications were made to the prompt. The prompt was tested repeatedly to ensure consistent and reproducible responses.

The same prompt was equally applied to all cases, with contextualization limited to the respective institutional MRI protocol lists and the case-specific input, i.e., age, gender, and clinical indication from the RRFs. The clinical parameters used as input were not altered and included only the patient's age, gender, and the unmodified free-text clinical indication as stated in the original RRF. The exact prompt used for protocol composition, including the original German version and its English translation, is provided in Supplementary Table 2 and in the following paragraph.

For each case, a new browser session was initiated to reduce potential contextual memory effects. All ChatGPT-4 interactions were performed using a paid subscription account to ensure consistent access to GPT-4. This account was held by one of the study authors, a board-certified (neuro-)radiologist with 9 years of clinical experience.

2.1.3. Selection of pre-defined MRI protocols

Based on the patient's clinical information, age, and sex, ChatGPT-4 was instructed to identify the most suitable and time-efficient MRI protocol from pre-specified lists of MRI protocols of the three different institutions as well as additional sequences from the given sequences in the MRI protocols, each with its acquisition time listed. The lists of the

Table 1
Overview of applied datasets.

Dataset	Number of cases	Department	Comment on the department	Initial Reading	Consensus Reading	Reason for Inclusion
Prompt engineering	30	Neuroradiology	In-house cases, but no overlap with the in-house dataset	Not applicable	Not applicable	Prompt engineering
In-house	1,001	Neuroradiology		Reader 1 and 2 independently from each other	Reader 1 and 2 together	Large-scale assessment
Independent	200	General Radiology	Same hospital as in-house dataset, but independent department	Reader 1 and 2 independently from each other	Reader 1 and 2 together	Confirmation in a broader general radiological context
External	300	Neuroradiology & ENT	Cases from the Netherlands	Reader 3 and 4 independently from each other	Reader 3 and 4 together	True external validation

MRI protocols of the different institutions are presented in Supplementary Table 1,3, 4; the exact input from the 1,001 in-house RRFs is presented in Supplementary Table 5. The existing MRI protocols were based on the recommendations of the ESR and ACR-ASNR-SPR guidelines and were further optimized internally [29,30]. Apart from acquisition time of each sequence, ChatGPT was not provided with any further information on the individual sequences or the protocols.

The specific prompt for ChatGPT-4 translated into English was: "Your task is as follows: Based on the clinical information as well as age and sex of the patient from the radiology request form, you are to determine the MRI protocol for the requested examination from the list of MRI protocols which follows later in the prompt. It is important to make time-efficient decisions and, if necessary, to supplement the main protocol with individual sequences to address the research question adequately. For this purpose, the acquisition time of each sequence is written in brackets after each sequence, in the format "mm:ss". Please also provide the total acquisition time of the chosen MRI protocol. There are the following main MRI protocols…"

After listing the MRI protocols (compare Supplementary Table 1, 3, 4) and the case- specific input consisting of age, gender and clinical information of the patient (compare Supplementary Table 5 with exemplary input from the in-house dataset) the following instruction was given: „Please also consider the potential complications of the suspected diagnosis which guides your choice of MRI protocol and sequences. Which main protocol and which additional sequences would you choose? The examination must remain time-efficient, therefore do not list any sequence twice.

2.1.4. Radiological reference rating

ChatGPT-4́s protocol suggestions were compared with expert decisions made by three board-certified (neuro-)radiologists (reader 1 (9 years of experience), reader 2 (18 years of experience), reader 3 (10 years of experience, external site)) and one board-certified radiologist (reader 4 (7 years of experience)), who were blinded to ChatGPT-4′s suggestions.

Rating categories were defined as follows:

- Identical: Full agreement between ChatGPT-4 and the radiologist on both the

  selected main MRI protocol and any additional sequences.

- Acceptable: Differences existed, but the protocol was still deemed sufficient to

  address the clinical question.

- Unacceptable: The protocol failed to adequately address the clinical question or indication.

All ChatGPT-4 protocol suggestions were evaluated against expert consensus, which served as the reference standard throughout the study for classification as "identical," "acceptable," or "unacceptable." All reported agreement metrics are derived from these consensus ratingas. The in-house and independent datasets were reviewed by Reader 1 and Reader 2, while the external dataset was assessed by Reader 3 and Reader 4 as seen in table 1 and 3.

2.1.5. Analysis of potential time savings

Furthermore, in an additional investigation, independent from the task previously presented, ChatGPT was instructed to freely compose individual protocols for each of the 1,001 in-house cases and all 300 external cases from the MRI sequences within the respective MRI protocol list (compare Supplementary Tables 1, 2 and 4), with the requirement that it should maintain full diagnostic validity but be as time-efficient as possible. Time differences reflect variations in the combination and duration of MRI sequences chosen by ChatGPT-4 compared to those selected by reader 1 across the in-house dataset.

2.2. Statistical analysis

Descriptive statistical analyses were conducted using Microsoft Excel (Version 2007 Microsoft Corp., Redmond, USA). Inter-reader variability analysis was performed with SPSS (Version 27) using the weighted Coheńs kappa test to provide weighted κ- values. A Wilcoxon signed-rank test and paired sample *t* test were performed to compare the time savings. A p-value of < 0.05 was considered significant.

All data are presented as mean ± standard deviation, unless otherwise specified.

3. Results

3.1. Selection of pre-defined MRI protocols

3.1.1. In-house dataset

In the analysis of the 1,001 in-house cases, the three most frequently selected MRI protocols (according to reader 1) were: tumor protocol (25.17 %), multiple sclerosis (MS) protocol with contrast agent (23.40 %), and ischemia protocol (13.99 %) (compare Supplementary Table 1). In 112 of 1,001 cases (11.19 %) additional sequences were required to supplement the main protocol (compare Table 2). The overall number of "identical", "acceptable" and "unacceptable" cases in the in-house dataset was 989, 10 and 2 for reader 1 and 943, 51 and 7 for reader 2, respectively. Table 3 summarizes the results of the readings of all datasets.

Unacceptable decisions of ChatGPT versus reader 1 and reader 2 were due to 1 and 4 differences in main protocol decisions, 1 and 1 differences in additional sequences, and 0 and 2 differences in both main protocol and additional sequences. All unacceptable cases, along with exemplary acceptable and identical cases, are shown in Supplementary Table 6.

For all acceptable cases, the counts of differences in the choices of

Table 2
Choice of specific MRI sequences in addition to the main MRI protocol (n = 112/1,001 cases).

Additional sequences/protocols	Absolute count	Relative count (% in n = 112)
coronal DWI	21	18.75
3D Inflow Angiography	15	13.39
supra-aortic contrast-enhanced MRA	29	25.89
sagittal T2 TSE	34	30.36
PCA	6	5.36
axial SWip	11	9.82
coronal T2 mDIXON	10	8.93
coronal T1 mDIXON CE	6	5.36
Orbita	3	2.68
coronal T2 1024 2 mm	7	6.25
axial T2 1024 2 mm	2	1.79
coronal temporal angulated FLAIR	3	2.68
axial FLAIR CE	2	1.79
keyhole	5	4.46
axial 3D T2 DRIVE	3	2.68
HWS CE	2	1.79
HWS Ligamenta	1	0.89
whole spine CE	2	1.79
coronal T1 non-contrast	1	0.89
coronal T1 CE	1	0.89

Abbreviations: Contrast-enhanced (CE), fluid attenuated inversion recovery (FLAIR), diffusion weighted imaging (DWI), susceptibility weighted imaging (SWIp), magnetic resonance angiography (MRA), driven equilibrium (DRIVE), phase contrast angiography (PCA), cervical spine (HWS).

Table 3
Overview of reading results.

Dataset	Reader	Classification		
		identical	acceptable	unacceptable
In-house (n = 1,001)	1/ChatGPT	989	10	2
	2/ChatGPT	943	51	7
	1/2	951	45	5
Independent (n = 200)	1/ChatGPT	198	2	0
	2/ChatGPT	197	3	0
	1/2	198	2	0
External (n = 300)	3/ChatGPT	277	23	0
	4/ChatGPT	270	30	0
	3/4	273	27	0

main protocol and/or additional sequences are summarized in Supplementary Table 7.

3.1.2. Independent and external dataset

In both the independent and external dataset, no protocol suggestions of ChatGPT were classified as "unacceptable". The number of differing protocol suggestions between ChatGPT and the radiologists were 2 of 200 cases (reader 1, independent dataset), 3 of 200 cases (reader 2, independent dataset) and 23 and 30 of 300 cases (reader 3 and 4, external dataset).

3.1.3. Assessment of inter-reader variability

The weighted kappa-values of all readings ranged between 0.88 and 0.98, indicating a very good agreement. All associated p-values were < 0.001, demonstrating that the observed agreement was statistically significant and not due to chance. Table 4 presents the assessment of inter-reader variability of all readings.

3.1.4. Analysis of potential time-savings

When ChatGPT-4 was tasked to freely compose time-efficient MRI protocols from the available sequences, these suggestions were approved in n = 766/1,001 (76.52 %) of the in-house cases and n = 140/300 (46.67 %) of the external cases, leading to clear time savings with a significance of p < 0.001 (mean time saving (standard deviation) = 3:51 min:seconds (±2:40) minutes and 2:59 (±3:42) minutes per adopted MRI protocol of the in-house and external dataset, respectively, compare Table 5).

Table 4
Results from inter-reader variability assessment.

Dataset	Reader	Weighted Cohens Kappa κ	p-value
In-house (n = 1,001)	Reader 1 vs. 2	0.941	<0.001
	Reader 1 vs. ChatGPT	0.933	<0.001
	Reader 2 vs. ChatGPT	0.881	<0.001
Independent (n = 200)	Reader 1 vs. 2	0.974	<0.001
	Reader 1 vs. ChatGPT	0.979	<0.001
	Reader 2 vs. ChatGPT	0.969	<0.001
External (n = 300)	Reader 3 vs. 4	0.951	<0.001
	Reader 3 vs. ChatGPT	0.933	<0.001
	Reader 4 vs. ChatGPT	0.937	<0.001

4. Discussion

This study reports a high agreement in MRI sequence selection between ChatGPT-4 and experienced board-certified (neuro-)radiologists, when original information from RRFs was utilized as input. This high agreement was achieved in three different datasets, covering a large scale in-house neuroradiology dataset (N = 1,001), a general radiology dataset from an independent department (N = 200), and an external dataset from a foreign institution (N = 300). The use of diverse datasets reflects the model's robustness across different institutional settings. Furthermore, ChatGPT-4 achieved relevant measurement time savings when it was allowed to freely compose protocols from the available set of sequences, demonstrating its potential to improve clinical efficiency.

In more than three-quarters of the in-house cases and nearly half of the external cases, the MRI protocol suggestions by ChatGPT-4 were diagnostically appropriate and saved 17 % (in-house) and 16 % (external) of the acquisition time compared to the predefined MRI protocols in the adopted cases. The lower number of approved, time-saving protocols in the external compared to the in-house dataset might be due to the overall shorter MRI protocols in the external institution compared to our MRI protocols (mean protocol acquisition time (± SD), external dataset: 18:29 (±07:04) minutes, in-house dataset: 22:03 (±04:24) minutes).

Table 5
Time savings when ChatGPT-4 freely composed time-efficient MRI protocols from the given sequences.

Dataset	No. of adopted MRI protocols according to ChatGPTs suggestion	Mean absolute time saving per adopted protocol (hh:mm:ss)	Standard deviation (hh:mm:ss)	Mean relative time saving per adopted protocol (%)	*p*-value
In-house	n = 766/1,001 (76.52 %)	00:03:51	± 00:02:40	17 %	<0.001
External	n = 140/300 (46.67 %)	00:02:59	± 00:03:42	16 %	<0.001

Time savings were calculated only for protocols adopted as diagnostically acceptable and shorter than the predefined standard protocols. Both datasets, including the In-house and independent dataset (n = 200), were evaluated independently.

Notably, these time savings were particularly pronounced when ChatGPT-4 was allowed to freely compose MRI protocols based solely on the clinical question and available sequences without being constrained by predefined institutional protocols. This flexible approach enabled the model to tailor imaging strategies more precisely to the diagnostic need, often omitting unnecessary sequences while maintaining diagnostic quality. While predefined protocols aim to standardize imaging and ensure comprehensive diagnostic assessment, ChatGPT-4 demonstrated the ability to dynamically adjust protocol complexity in a time-efficient manner, offering a promising tool to optimize workflow and resource use in clinical practice.

Numerous studies to date have reported on a potential improvement of daily clinical workflows through AI, particularly with the use of ChatGPT-4 [12,31,32]. However, while many of these papers suggest a prospective benefit, our work demonstrates an immediately implementable use case: The input consisted of anonymized and otherwise unmodified examination requests, meaning theoretically, any radiological resident could immediately utilize this support from ChatGPT in the browser without any implementation effort. Previous studies have shown that ChatGPT-4 can support decisions such as selecting examination regions, contrast administration, or imaging modalities, achieving a correct decision rate of 84 % across all cases [22,32,33]. While ChatGPT can provide helpful initial protocol suggestions – requiring revision in only a small percentage of cases – it must always be used under expert supervision. Particularly in educational settings, there is a risk that less experienced radiologists may over-rely on AI output, potentially adopting incorrect decisions. Prior studies have shown that junior clinicians are more susceptible to AI bias than experienced ones [34]. Thus, AI should complement, not replace, structured radiology training and critical thinking.

Our study, however, focused on the determination of MRI protocols and sequences, a potentially more complex task as this requires ChatGPT-4 to not only identify potential differential diagnoses but also to ascertain the most effective MRI approach for evaluating these conditions [21]. Therefore, our research expands upon the existing literature, as Chat-GPT performed a particularly advanced medical task out-of-the-box with very high precision in a large cohort of 1,001 cases, which was confirmed in an independent and external validation. Remarkably, ChatGPT-4 was able to justify its choice for each selected sequence on a case-by-case basis, even though the prompt did not include any explanations for the sequences. Additionally, an analysis of whether the choice of additional sequences could be traced back to specific trigger words found no keywords. Instead, it appeared that certain descriptions of symptoms, much like with human experts, must have guided the selection of supplementary sequences.

These findings underscore that ChatGPT-4 should be regarded strictly as a supplementary tool that supports, but does not replace, the clinical judgment of the responsible radiologist. Human oversight remains essential to ensure diagnostic safety, particularly in cases where incorrect or suboptimal suggestions could lead to adverse consequences such as unnecessary administration of contrast agents. In our study, a small number of protocol suggestions were classified as “unacceptable,” reflecting scenarios where ChatGPT-4’s output failed to fully address the clinical question. While such cases were rare, they illustrate that not all exceptional situations can be anticipated or fully covered through prompt design alone. One possible improvement may include more specific instructions in the prompt, such as advising against contrast use in pediatric patients. Therefore, oversight by a qualified physician remains an indispensable aspect of using AI tools like ChatGPT-4 in clinical settings.

The integration of AI in medical education, particularly in the training of radiology residents, is a subject of debate. However, the consensus is increasingly recognizing AI’s potential in various educational aspects. AI can prioritize urgent cases for prompt review by supervising physicians, tailor learning materials to the specific needs of residents, and enhance the quality of radiology reports drafted by them [23,35–37]. This study underscores how residents could effectively utilize ChatGPT-4 to assist in determining the appropriate MRI protocol based on available clinical data. In all 1,501 analyzed cases, ChatGPT-4 provided justifications for its choices of MRI protocols and additional sequences. These detailed, case-specific explanations present a unique educational opportunity for trainees, allowing them to follow the model’s clinical reasoning and compare it with expert decisions under supervision. As such, the application of ChatGPT-4 may be particularly beneficial for less experienced radiologists not only as a training aid but also as a support tool in clinical routine. Beyond education, our findings suggest that ChatGPT-4 has the potential to enhance workflow efficiency, improve resource utilization, and ultimately contribute to cost-effective radiological practice. One limitation of using a continuously evolving LMM such as ChatGPT is the potential variability in output across different model versions. As ChatGPT is regularly updated, identical prompts may not consistently produce the same responses over time. While newer versions may offer improved performance, this dynamic nature poses challenges for reproducibility and clinical validation. Similarly, even slight changes in the wording or structure of a prompt can lead to different outcomes. This prompt sensitivity highlights the importance of maintaining consistent prompt formulations for clinical use and ensuring regular revalidation when new prompts or model updates are introduced. To address this issue, locally hosted LLMs are currently under investigation [38]. These models can offer greater control, stability, and compliance with data protection regulations, particularly in clinical environments. However, they may lack the continuous optimization and performance gains of cloud-based models, highlighting a trade-off between consistency and innovation [39].In light of these considerations, we suggest that future clinical implementations of LLMs in clinical workflows should include a structured prompt review and revalidation process following each major model update. Based on our experience, such re-evaluation could involve testing the updated model on a small representative dataset (e.g., 30 referral cases) to ensure that prompt performance remains aligned with clinical expectations and intended use.

To overcome current limitations of ChatGPT in clinical use, structured validation and feedback are essential. As outlined by Pianykh et al. (2020), continuous learning principles, such as expert supervision, version tracking, and outcome-based revalidation, can help maintain safety and improve performance over time, even in non-adaptive AI systems [34].

As with any retrospective analysis, our study has several limitations. A more comprehensive evaluation of ChatGPT-4′s performance, including integration of additional clinical data such as structured reports or laboratory findings, could further enhance the model’s clinical relevance. Additionally, the current version of ChatGPT-4 has not undergone formal clinical validation, and its use in healthcare remains limited to supportive, non-decisive functions. Any application of such AI tools must adhere to regulatory standards concerning data protection and patient safety, which are not yet fully defined or standardized in the context of large language models. Some deviations between ChatGPT-4 and expert decisions were observed, particularly in cases classified as “acceptable” or “unacceptable.” These discrepancies may result from variable clinical language, ambiguous referral phrasing, or limitations in the model’s contextual understanding. Future efforts could focus on standardizing referral formats or refining prompts to include explicit clinical priorities or constraints. Furthermore, our study focused primarily on neuroradiological scenarios. To assess the broader applicability of ChatGPT-4 in radiology, future research should investigate its performance in musculoskeletal, thoracic, abdominal, and other subspecialties using similarly structured validation frameworks. Future studies will explore other open source LLMs hosted publicly as well as locally.

In conclusion, while ChatGPT-4 emerges as a promising tool in supporting neuroradiological practices, its integration should be undertaken with care, highlighting the synergistic relationship between AI

and human expertise. It must be emphasized that the final decision-making authority regarding the protocol to be applied and the full responsibility remain with the radiologist.

However, the insights gained from this study advocate for the feasibility of integrating AI tools like ChatGPT-4 as a means of support within clinical workflows, potentially leading to more efficient and patient-tailored radiological assessments. Furthermore, it could serve as a valuable adjunct for educational and training purposes, complementing the standard care provided by experienced neuroradiologists. Importantly, this application of ChatGPT-4 is readily deployable in the clinical environment without requiring additional implementation efforts or raising new data protection concerns.

CRediT authorship contribution statement

Zeynep Bendella: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Barbara Daria Wichtmann:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. **Ralf Clauberg:** Writing – review & editing, Validation, Formal analysis, Data curation. **Vera C. Keil:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation. **Nils C. Lehnen:** Writing – review & editing, Validation, Formal analysis, Data curation. **Robert Haase:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Laura C. Sáez:** Writing – review & editing, Validation, Methodology, Formal analysis. **Isabella C. Wiest:** Writing – review & editing, Validation, Formal analysis, Data curation. **Jakob Nikolas Kather:** Writing – review & editing, Validation, Supervision, Formal analysis, Data curation. **Christoph Endler:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis. **Alexander Radbruch:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Daniel Paech:** Writing – review & editing, Writing – original draft, Project administration, Conceptualization. **Katerina Deike:** Writing – original draft, Validation, Supervision, Formal analysis, Conceptualization.

Funding

The authors state that this work has not received any funding.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: KD, DP, RH and AR are co-founder and shareholder of the relios.vision GmbH. BDW has received speaker honoraria from Philips Healthcare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2025.112416.

References

[1] H.J. Huhtanen, M.J. Nyman, A. Karlsson, J. Hirvonen, Machine learning and deep learning models for automated protocoling of emergency brain MRI using text from clinical referrals, Radiol. Artif. Intell. 7 (2025) e230620.
[2] N.K. Pinter, The right imaging protocol for the right patient, Continuum (Minneap Minn) 29 (2023) 16–26.
[3] A.S. Shetty, T.J. Fraum, D.R. Ludwig, et al., Body MRI:imaging protocols, techniques, and lessons learned, Radiographics 42 (2022) 2054–2074.
[4] M. Bernardy, C.G. Ullrich, J.V. Rawson, et al., Strategies for managing imaging utilization, J. Am. Coll. Radiol. 6 (2009) 844–850.
[5] E.R. McVeigh, R.M. Henkelman, M.J. Bronskill, Optimization of survey protocols for MRI, Magn. Reson. Med. 13 (1990) 177–191.
[6] P.J. Bairstow, J. Persaud, R. Mendelson, L. Nguyen, Reducing inappropriate diagnostic practice through education and decision support, Int. J. Qual. Health Care 22 (2010) 194–200.
[7] A.D. Brown, T.R. Marotta, A Natural Language Processing-based Model to Automate MRI Brain Protocol selection and Prioritization, Acad. Radiol. 24 (2017) 160–166.
[8] A. Kalra, A. Chakraborty, B. Fine, J. Reicher, Machine learning for automation of radiology protocols for quality and efficiency improvement, J. Am. Coll. Radiol. 17 (2020) 1149–1158.
[9] Y. Tadavarthi, V. Makeeva, W. Wagstaff, et al., Overview of Noninterpretive Artificial Intelligence Models for Safety, Quality, Workflow, and Education applications in Radiology Practice, Radiol. Artif. Intell. 4 (2022) e210114.
[10] G. Litjens, T. Kooi, B.E. Bejnordi, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[11] GPT-4. OpenAI. Available at: https://openai.com/gpt-4. Accessed April 6, 2023.
[12] M. Mannil, J. von Spiczak, R. Manka, H. Alkadhi, Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography, Invest. Radiol. 53 (2018) 338–343.
[13] L. Floridi, M. Chiriatti, GPT-3: its nature, scope, limits, and consequences, Minds Mach 30 (2020) 681–694.
[14] S. Biswas, ChatGPT and the future of medical writing, Radiology 307 (2023) e223312.
[15] F.C. Kitamura, ChatGPT is shaping the future of medical writing but still requires human judgment, Radiology 307 (2023) e230171.
[16] J. Clusmann, F.R. Kolbinger, H.S. Muti, The future landscape of large language models in medicine, Commun Med (lond) 3 (2023) 141.
[17] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S. W. Ting, Large language models in medicine, Nat. Med. 29 (2023) 1930–1940.
[18] N.C. Lehnen, F. Dorn, I.C. Wiest, et al., Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis, Radiology 311 (2024) e232741.
[19] R. Bhayana, B. Nanda, T. Dehkharghanian, et al., Large language models for automated synoptic reports and resectability categorization in pancreatic cancer, Radiology 311 (2024) e232714.
[20] R.J. Gertz, T. Dratsch, A.C. Bunck, et al., Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accurancy, Radiology 311 (2024) e232714.
[21] H.L. Haver, E.B. Ambinder, M. Bahl, E.T. Oluyemi, J. Jeud, P.H. Yi, Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT, Radiology 307 (2023) e230424.
[22] R.J. Gertz, A.C. Bunck, S. Lennartz, et al., GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study, Radiology 307 (2023) e230877.
[23] K.S. Amin, M.A. Davis, R. Doshi, A.H. Haims, P. Khosla, H.P. Forman, Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports, Radiology 309 (2023) e232561.
[24] R. Doshi, K.S. Amin, P. Khosla, S.S. Bajaj, S. Chheang, H.P. Forman, Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis, Radiology 310 (2024) e231593.
[25] J. Kottlors, G. Bratke, P. Rauen, et al., Feasibility of differential diagnosis based on imaging patterns using a large language model, Radiology 308 (2023) e231167.
[26] D. Li, K. Gupta, M. Bhaduri, P. Sathiadoss, S. Bhatnagar, J. Chong, Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases, Radiology 310 (2024) e232411.
[27] T.H. Kung, M. Cheatham, A. Medenilla, et al., Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models, PLoS Digital Health, PLoS Digital Health 2 (2023) e0000198.
[28] Y. Barash, E. Klang, E. Konen, V. Sorin, ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection, J. Am. Coll. Radiol. 20 (2023) 998–1003.
[29] (ESR). ESR iGuide: Clinical decision support for imaging referrals Vienna: ESR; (2025) Available via https://www.myesr.org/eu-international-affairs/policy-topics/esr-iguide/. Accessed 9 Apr 2025.
[30] American society of Neuroradiology._Practice Guidelines_ASNR (2025) Available via https://www.asnr.org/practice-guidelines/. Accessed 9 Apr 2025.
[31] Y. Kim, K.J. Lee, L. Sunwoo, et al., Deep learning in diagnosis of maxillary sinusitis using conventional radiography, Invest. Radiol. 54 (2019) 7–15.
[32] M. Perkuhn, P. Stavrinou, F. Thiele, et al., Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine, Invest. Radiol. 53 (2018) 647–654.
[33] L.C. Adams, D. Truhn, F. Busch, et al., Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study, Radiology 307 (2023) e230725.
[34] O.S. Pianykh, G. Langs, M. Dewey, et al., Continuous learning AI in radiology: Implementation principles and early applications, Radiology 297 (2020) 6–14.
[35] P.J. Slanetz, D. Daye, P.H. Chen, L.R. Salkowski, Artificial intelligence and machine learning in radiology education is ready for prime time, J. Am. Coll. Radiol. 17 (2020) 1705–1707.
[36] S.H. Tajmir, T.K. Alkasab, Toward augmented radiologists: changes in radiology education in the era of machine learning and artificial intelligence, Acad. Radiol. 25 (2018) 747–750.

[37] S. Gaube, H. Suresh, M. Raue, et al., Do as AI say: susceptibility in deployment of clinical decision-aids, npj Digital Med. 4 (2021) 31.
[38] Y.J. Park, A. Pillai, J. Deng, et al., Assessing the research landscape and clinical utility of large language models: a scoping review, BMC Med. Inform. Decis. Mak. 24 (2024) 72.
[39] J. Jonnagaddala, Z.S.Y. Wong, Privacy preserving strategies for electronic health records in the era of large language models, npj Digital Med. 8 (2025) 34.