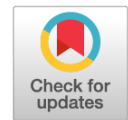


DOI: <https://doi.org/10.17816/DD678373>

EDN: QSANCA



# Application of Large Language Models in Radiological Diagnostics: A Scoping Review

Yuriy A. Vasilev<sup>1</sup>, Roman V. Reshetnikov<sup>1</sup>, Olga G. Nanova<sup>1</sup>, Anton V. Vladzmyrskyy<sup>1</sup>, Kirill M. Arzamasov<sup>1</sup>, Olga V. Omelyanskaya<sup>1</sup>, Maria R. Kodenko<sup>1</sup>, Rustam A. Erizhokov<sup>1</sup>, Anastasia P. Pamova<sup>1</sup>, Seal R. Seradzhi<sup>1</sup>, Ivan A. Blokhin<sup>1</sup>, Anna P. Gonchar<sup>1,2</sup>, Pavel B. Gelezhe<sup>1</sup>, Dina A. Akhmedzyanova<sup>1</sup>, Yuliya F. Shumskaya<sup>1</sup>

<sup>1</sup> Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;

<sup>2</sup> Moscow City Hospital named after S.S. Yudin, Moscow, Russia

## ABSTRACT

**BACKGROUND:** Modern large language models show potential for application in radiological diagnostics across a wide range of routine tasks.

**AIM:** The work aimed to conduct a scoping review of the application of large language models in radiological diagnostics by analyzing possible use-case scenarios and assessing the methodological quality of relevant studies.

**METHODS:** Two search strategies were employed: a primary search (PubMed and eLibrary) targeting full-text publications with well-developed methodology, and a supplementary search (PubMed) aimed at broader coverage of large language model use cases in radiological diagnostics during 2023–2025. Extracted data included bibliometric characteristics, study objectives, use-case scenarios of large language models, nosological profiles, key methodological parameters, and both quantitative and qualitative indicators of diagnostic performance—for both the models and the specialists involved, including their number and experience. The quality was assessed using the modified QUADAS-CAD questionnaire.

**RESULTS:** The primary search yielded 9 studies for analysis; the supplementary search yielded 216. A total of 9 major use-case scenarios for large language models in radiology were identified. The most common among them was the rephrasing of radiology reports in order to improve their accessibility for patient understanding. Models predominantly used were GPT-4 and BERT, along with GPT-3.5, Llama 2, Med42, GPT-4V, and Gemini Pro. The large language model GPT-4 demonstrated high diagnostic accuracy in identifying brain tumors (73.0%), myocarditis (83.0%), and in making decisions on invasive procedures for acute coronary syndrome (86.0%). In turn, it demonstrated low diagnostic accuracy for nervous system disorders of various etiologies (50.0%) and for musculoskeletal diseases (43.0%). The BERT model exhibited high diagnostic accuracy in detecting pulmonary nodules (99.0%) and signs of intracranial hemorrhage (sensitivity and specificity: 97.0% and 90.0%, respectively), as well as in report classification (accuracy: 84.3%).

Most articles (88.9%) carried a high risk of bias. The main reasons for this included small and imbalanced sample sizes, overlap between training and test datasets, and insufficiently precise preparation and description of reference standards.

**CONCLUSION:** The diagnostic performance of large language models varies significantly across articles. Their clinical implementation requires standardized, methodologically robust research, including larger and more balanced samples, optimization of the structure and volume of datasets, separation of training and testing samples, thorough preparation and description of reference standards, as well as the accumulation of empirical data for specific radiological tasks.

**Keywords:** artificial intelligence; large language models; radiological diagnostics; radiology report; systematic review.

## To cite this article:

Vasilev YuA, Reshetnikov RV, Nanova OG, Vladzmyrskyy AV, Arzamasov KM, Omelyanskaya OV, Kodenko MR, Erizhokov RA, Pamova AP, Seradzhi SR, Blokhin IA, Gonchar AP, Gelezhe PB, Akhmedzyanova DA, Shumskaya YuF. Application of Large Language Models in Radiological Diagnostics: A Scoping Review. *Digital Diagnostics*. 2025;6(2):268–285. DOI: 10.17816/DD678373 EDN: QSANCA

Submitted: 06.05.2025

Accepted: 12.06.2025

Published online: 17.06.2025

DOI: <https://doi.org/10.17816/DD678373>

EDN: QSANCA

# Применение больших языковых моделей в лучевой диагностике: обзор предметного поля

Ю.А. Васильев<sup>1</sup>, Р.В. Решетников<sup>1</sup>, О.Г. Нанова<sup>1</sup>, А.В. Владзимирский<sup>1</sup>, К.М. Арзамасов<sup>1</sup>,  
О.В. Омелянская<sup>1</sup>, М.Р. Коденко<sup>1</sup>, Р.А. Ерижоков<sup>1</sup>, А.П. Памова<sup>1</sup>, С.Р. Сераджи<sup>1</sup>, И.А. Блохин<sup>1</sup>,  
А.П. Гончар<sup>1,2</sup>, П.Б. Гележе<sup>1</sup>, Д.А. Ахмедзянова<sup>1</sup>, Ю.Ф. Шумская<sup>1</sup>

<sup>1</sup> Научно-практический клинический центр диагностики и телемедицинских технологий, Москва, Россия;

<sup>2</sup> Городская клиническая больница им. С.С. Юдина, Москва, Россия

## АННОТАЦИЯ

**Обоснование.** Современные большие языковые модели обладают потенциалом использования в лучевой диагностике для решения широкого спектра рутинных задач.

**Цель исследования.** Провести обзор предметного поля применения больших языковых моделей в лучевой диагностике с анализом возможных сценариев их использования и оценкой качества методологии соответствующих исследований.

**Методы.** Провели два варианта поиска — первичный (PubMed и eLibrary), ориентированный на выявление полнотекстовых публикаций с максимально проработанной методологией, и дополнительный (PubMed), направленный на широкий охват сценариев применения больших языковых моделей в лучевой диагностике за период 2023–2025 гг. Извлекли библиометрические данные, формулировку исследовательской задачи, сценарий применения больших языковых моделей, нозологический профиль, ключевые методологические параметры, а также количественные и качественные показатели диагностической эффективности как моделей, так и участвующих специалистов, включая их число и опыт. Качество исследований оценивали с использованием модифицированного опросника QUADAS-CAD.

**Результаты.** При первичном поиске для анализа отобрано 9 публикаций, при дополнительном — 216. Найдено 9 основных сценариев применения больших языковых моделей в лучевой диагностике. Наиболее распространёнными из них было переформулирование рентгенологических заключений с целью повышения их доступности восприятия пациентами. Преимущественно использовали модели GPT-4 и BERT, а также GPT-3.5, Llama 2, Med42, GPT-4V и Gemini Pro. Большая языковая модель GPT-4 продемонстрировала высокую точность при диагностике опухолей головного мозга (73,0%), миокардитов (83,0%), а также в случае принятия решений о проведении инвазивной процедуры при остром коронарном синдроме (86,0%). В свою очередь, она продемонстрировала низкую диагностическую точность в отношении патологий нервной системы различной этиологии (50,0%) и заболеваний опорно-двигательной системы (43,0%). Модель BERT показала высокую диагностическую точность в задачах детекции лёгочных узелков (99,0%) и признаков внутричерепного кровоизлияния (чувствительность и специфичность — 97,0 и 90,0% соответственно), а также при классификации заключений (точность 84,3%).

Большинство работ (88,9%) содержат вероятность систематической ошибки. Основные причины этого: маленький объём и несбалансированность выборок, пересечение обучающих и тестовых наборов данных, недостаточно аккуратная подготовка и описание референсных стандартов.

**Заключение.** Показатели диагностической точности больших языковых моделей сильно варьируют между разными исследованиями. Для их внедрения в клиническую практику необходимо проведение стандартизированных и методологически качественных исследований, включающих увеличение объёма и сбалансированности выборок, оптимизацию структуры и объёма наборов данных, формирование неперекрывающихся обучающих и тестовых выборок, тщательную подготовку и описание референсных стандартов, а также накопление эмпирических данных по отдельным задачам лучевой диагностики.

**Ключевые слова:** искусственный интеллект; большие языковые модели; лучевая диагностика; рентгенологический протокол; систематический обзор.

### Как цитировать:

Васильев Ю.А., Решетников Р.В., Нанова О.Г., Владзимирский А.В., Арзамасов К.М., Омелянская О.В., Коденко М.Р., Ерижоков Р.А., Памова А.П., Сераджи С.Р., Блохин И.А., Гончар А.П., Гележе П.Б., Ахмедзянова Д.А., Ю.Ф. Шумская Ю.Ф. Применение больших языковых моделей в лучевой диагностике: обзор предметного поля // Digital Diagnostics. 2025. Т. 6, № 2. С. 268–285. DOI: 10.17816/DD678373 EDN: QSANCA

DOI: <https://doi.org/10.17816/DD678373>

EDN: QSANCA

# 大语言模型在放射诊断中的应用：范围综述

Yuriy A. Vasilev<sup>1</sup>, Roman V. Reshetnikov<sup>1</sup>, Olga G. Nanova<sup>1</sup>, Anton V. Vladzimirskyy<sup>1</sup>, Kirill M. Arzamasov<sup>1</sup>, Olga V. Omelyanskaya<sup>1</sup>, Maria R. Kodenko<sup>1</sup>, Rustam A. Erizhokov<sup>1</sup>, Anastasia P. Pamova<sup>1</sup>, Seal R. Seradzhi<sup>1</sup>, Ivan A. Blokhin<sup>1</sup>, Anna P. Gonchar<sup>1,2</sup>, Pavel B. Gelezhe<sup>1</sup>, Dina A. Akhmedzyanova<sup>1</sup>, Yuliya F. Shumskaya<sup>1</sup>

<sup>1</sup> Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;

<sup>2</sup> Moscow City Hospital named after S.S. Yudin, Moscow, Russia

## 摘要

**论证。**现代大语言模型具备在放射诊断中应用于解决广泛常规任务的潜力。

**目的：**综述大语言模型在放射诊断中的应用范围，分析其使用场景，并评估相关研究的方法学质量。

**方法。**开展两轮文献检索：初步检索（PubMed和eLibrary）聚焦于具备详实方法学的全文研究，补充检索（PubMed）旨在广泛覆盖2023 - 2025年间大语言模型在放射诊断中应用的各种情境。提取了书目信息、研究任务的表述、大语言模型的应用场景、疾病谱、关键方法学参数，以及诊断效能的定量与定性指标，涵盖模型本身及参与专家，包括其人数与经验。采用改良版QUADAS-CAD问卷对研究质量进行评估。

**结果。**初步检索纳入9项研究，补充检索纳入216项。共识别出在放射诊断中应用大语言模型的9种主要场景。其中最常见的是为提升患者理解而对放射学报告进行改写。最常使用的模型包括GPT-4和BERT，以及GPT-3.5、Llama 2、Med42、GPT-4V和Gemini Pro。大语言模型GPT-4在脑肿瘤（73.0%）、心肌炎（83.0%）以及急性冠状动脉综合征中介入治疗决策（86.0%）方面表现出较高的诊断准确性。但在诊断不同病因的神经系统疾病（50.0%）和肌肉骨骼疾病（43.0%）方面准确性较低。BERT模型在肺结节检测（99.0%）和颅内出血征象识别（灵敏度97.0%、特异度90.0%）方面表现优异，在放射学报告分类中准确率为84.3%。大多数研究（88.9%）存在系统性偏倚的可能。其主要原因包括：样本量小且分布不均、训练集与测试集重叠、参考标准准备和描述不够严谨。

**结论。**大语言模型的诊断准确性在不同研究间差异显著。其进入临床实践前，亟需开展标准化且方法学严谨的研究，包括扩大并平衡样本量、优化数据集结构与规模、明确划分训练集与测试集、严谨制定和描述参考标准，并针对特定放射诊断任务积累实证数据。

**关键词：**人工智能；大语言模型；放射诊断；放射学报告；系统综述。

## 引用本文：

Vasilev YuA, Reshetnikov RV, Nanova OG, Vladzimirskyy AV, Arzamasov KM, Omelyanskaya OV, Kodenko MR, Erizhokov RA, Pamova AP, Seradzhi SR, Blokhin IA, Gonchar AP, Gelezhe PB, Akhmedzyanova DA, Shumskaya YuF. 大语言模型在放射诊断中的应用：范围综述. *Digital Diagnostics*. 2025;6(2):268–285. DOI: 10.17816/DD678373 EDN: QSANCA

收到: 06.05.2025

接受: 12.06.2025

发布日期: 17.06.2025

## BACKGROUND

Large language models (LLMs) are artificial intelligence (AI) models that use deep learning technologies to process natural language, generate and interpret texts. Although the first language models were developed several decades ago, significant progress was only achieved with the release of the architecture of Transformers architecture in 2018. This architecture uses large amounts of text data to train models and forms the basis for current LLMs [1–3].

Recently, there has been an increasing focus on using LLMs in radiological diagnostics. Current models demonstrate a high level of training and successfully pass assessment tests in radiological diagnostics, providing answers with accuracy comparable to, and sometimes exceeding, that of radiologists [1, 2]. LLMs can assist

radiologists by generating structured radiology reports, interpreting and using unstructured diagnostic scan protocols, and identifying clinically significant changes [3, 4].

However, the practical application of LLMs in radiological diagnostics and their accuracy and reproducibility rates for various clinical cases have not been thoroughly evaluated.

## AIM

The work aimed to conduct a scoping review of the application of LLMs in radiological diagnostics by evaluating possible use-case scenarios and assessing the methodological quality of relevant studies.

## METHODS

This paper was prepared in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines [5].

### Search Strategy

The primary search [6] was performed using two search engines: PubMed and eLibrary. The publication search spanned 2023–2025.

The following search query was used in PubMed: (((Large language models) OR (LLM)) OR (Natural language processing)) AND (Radiology).

The following filter options were selected to identify the most robust evidence: Free Full Text, Full Text, Classical Article, Clinical Study, Comparative Study, Controlled Clinical Trial, Multicenter Study.

The following keywords were used in the eLibrary search query: *большие языковые модели (large language models), рентгенология (radiology)*.

Additionally, a manual search was performed using the reference lists of the selected publications and two reviews [3, 4].

The search strategy involved two steps.

- First, we analyzed the titles and abstracts of all the publications that were found using the above search

queries. Articles that met the objectives of this paper were selected for further analysis. We excluded study protocols, plans, and publications about the use of LLMs in radiology examination questions, and publications that did not describe the diagnostic performance of models.

- Second, we analyzed the full texts and their availability from the pool of selected papers and then compiled a sample for the primary analysis.

A supplementary PubMed search was conducted without restrictions on text availability: (((Large language models) OR (LLM)) OR (Natural language processing)) AND (Radiology) NOT review[pt] NOT letter[pt] NOT editorial[pt] NOT comment[pt].

The publication search spanned 2023–2025. The last search was performed on May 26, 2025

and included only publications with English abstracts in peer-reviewed journals. Exclusion criteria:

- Articles that were not original papers
- Articles not related to generative AI
- Articles that did not address the practical application of LLMs in radiological diagnostics, including publications on model performance in answering medical assessment questions.

Eight experts selected the papers. Two experts further assessed the final list of included publications. The experts were researchers with more than 10 years of experience in medical informatics.

### Information Extraction and Paper Quality Assessment

The following characteristics were extracted from the full texts of the selected articles:

- Bibliometric data such as the first author, title of publication, year of publication, DOI, journal and its impact factor, and country of study origin;
- The aim of the study, the type of target condition, the study area, and the main characteristics of the study (e.g., sample size, study design, external validation, and type of LLMs used);
- Diagnostic performance metrics of models (sensitivity, specificity, accuracy);
- Comparative diagnostic performance of AI models versus medical professionals;
- Number of radiologists and their qualifications;
- Information about conflicts of interest.

The quality of the selected publications was assessed using a Quality Assessment of Diagnostic Accuracy Studies Computer-Aided Detection (QUADAS-CAD) tool, which was developed for studies using AI [7].

The following data was extracted to generate supplementary search results:

- Bibliometric data
- Aims and methodology of papers
- Radiology modality
- Target condition

- Type of LLM
- Characteristics of study populations
- Study design
- Number of radiologists and their experience
- Type of reference standard
- Type of data for LLM
- General conclusion of the study.

Quality for supplementary search results was not assessed.

## RESULTS

### Primary Search Results

Fig. 1 shows the results of the primary systematic search.

### Key characteristics of articles

Tables 1–3 summarize key characteristics of the selected publications and related studies.

Three studies [8–10] used radiology reports, and one of these studies used both text reports and images [8]. One study focused on neuroradiology and the diagnosis of brain tumors [9], another evaluated musculoskeletal diseases [8], and the third evaluated myocarditis [10].

Two studies detected findings in radiology reports [11, 12]. Of these, one study focused on diagnosing lung cancer by detecting pulmonary nodules [11], and the other evaluated signs of intracranial hemorrhage [12].

One study evaluated the performance of LLMs in answering clinical questions using text data and images [13]. Additionally, the analysis included various target conditions, but the publication did not provide data on their distribution.

One study used text data from medical records to diagnose nervous system diseases and identify clinically significant changes based on radiological findings [14]. One

study categorized head computed tomography (CT) reports by their importance [15].

One study evaluated the feasibility of using clinical data and CT protocol findings to make decisions on invasive procedures in patients with acute coronary syndrome [16].

The review included publications describing five multicenter studies [9, 10, 12–14] and four single-center studies [8, 11, 15, 16]. No studies with external validation were identified. Of the nine studies above, only one was prospective [16], and the other eight were retrospective.

Three studies used large sample sizes:

- 101,703 CT reports [11]
- 34,188 brain CT reports [12]
- 3738 head CT reports [15].

Five studies used relatively small samples (86–396 cases) [8–10, 14, 16]. One study did not specify the sample size [13]. Four publications described demographic characteristics [9–11, 16], but most of them were not balanced, except for one study [11]. Seven studies characterized the distribution of conditions within the samples [8–12, 14, 15], but they were all imbalanced in their representation of different diseases.

Two studies used MRI reports [9, 10], four used CT reports [11, 12, 15, 16], and one used report obtained by any available modality [8]. Two publications did not specify the type of imaging [13, 14].

ChatGPT-based models were the most common AI architectures, particularly GPT-4<sup>®</sup> (OpenAI, USA), which was used in five papers [8–10, 13, 14, 16]. A BERT<sup>®</sup> model (Google, USA) was used in three studies [11, 12, 15].

Three papers compared different LLMs [8, 12, 13]. For example, Horiuchi et al. [8] compared the performance of the GPT-4<sup>®</sup> model (OpenAI, USA) that processes text, with GPT-4V, a multimodal model designed to process text

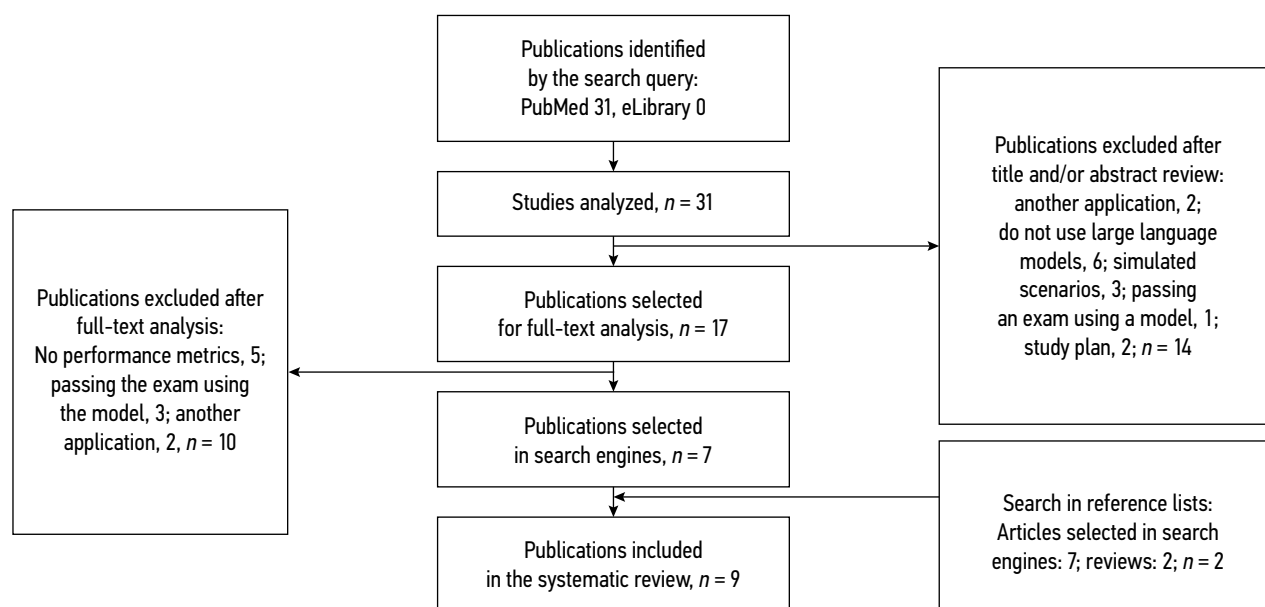


Fig. 1. Flowchart of the primary systematic search for publications.

Table 1. List of reviewed publications and their key characteristics

Authors	Year	Title	Journal	Impact factor	Study objective	Country	Modality	Target condition
Mitsuyama et al. [9]	2024	Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors	European Radiology	4.7	Evaluate the diagnostic performance of GPT-4® (OpenAI, USA) in brain tumor analysis compared with neuro-radiologists and general radiologists	Japan	MRI	Brain tumors
Horiuchi et al. [8]	2025	ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology	European Radiology	4.7	Evaluate the diagnostic accuracy of GPT-4® and GPT-4V® (OpenAI, USA) and musculoskeletal radiologists	Japan	Ultrasound, MRI, CT	Musculoskeletal diseases
Grolleau et al. [11]	2024	Incidental pulmonary nodules: Natural language processing analysis of radiology reports	Respiratory Medicine and Research	2.2	Evaluate detection parameters for pulmonary nodules during the year: rates, changes, and clinical and radiological characteristics	France	Chest CT	Lung tumors
Han et al. [13]	2024	Comparative analysis of multimodal large language model performance on clinical vignette questions	JAMA	64	Evaluate the performance of large language models in answering clinical questions	Germany	Uncertain	Uncertain
Horiuchi et al. [14]	2024	Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases	Neuroradiology	2.4	Evaluate the diagnostic performance of GPT-4® (OpenAI, USA) in neuroradiology	Japan	Uncertain	Nervous system disorders of various etiologies
Wabaya et al. [15]	2024	Comparison of natural language processing algorithms in assessing the importance of head computed tomography reports written in Japanese	Japanese Journal of Radiology	2.9	Develop a five-point scale for radiology report classification and compare the relevant performance of natural language processing algorithms using head CT reports in Japanese	Japan	CT	Head disorders
Kaya et al. [10]	2024	Generative Pre-trained Transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: A multicenter study	Journal of Cardiovascular Magnetic Resonance	9.1	Evaluate GPT-4® (OpenAI, USA) performance in medical decision-making based on cardiac MRI findings in suspected myocarditis	Germany	MRI	Myocarditis
Khoruzhaya et al. [12]	2024	Compare an ensemble of machine learning algorithms and BERT in their ability to interpret text descriptions of brain CT scans for the presence of intracranial hemorrhage	Current medical technologies	0.7 RINC 1.243	Train and test machine learning models and compare their performance with that of the BERT® language model (Google, USA), which was pre-trained on medical data for simple binary classification	Russia	CT	Intracranial hemorrhage
Cagnina et al. [16]	2025	Assessing the need for coronary angiography in high-risk non-ST-elevation acute coronary syndrome patients using artificial intelligence and computed tomography	The International Journal of Cardiovascular Imaging	1.9	Evaluate the performance of GPT-4® (OpenAI, USA) in determining the need for invasive coronary angiography for patients with high-risk acute coronary syndrome and no ST-segment elevation, using both standard clinical data and coronary CT findings.	Switzerland	CT	Acute coronary syndrome

Note. CT, computed tomography; MRI, magnetic resonance imaging; RSC, Russian Science Citation Index; USA, United States of America.

**Table 2.** Key characteristics of the studies presented in the included publications

Authors	Models	Period	Inclusion Criteria	Exclusion Criteria:	Study design	No. of radiologists and their experience	Reference test	External validation
Mitsuyama et al. [9]	GPT-4® (OpenAI, USA)	2017–2021	—	Recurrence events (96 radiology reports)	<ul style="list-style-type: none"> <li>• multicenter (2 sites)</li> <li>• retrospective</li> </ul>	<ul style="list-style-type: none"> <li>• 3 neuroradiologists</li> <li>• 4 general radiologists</li> </ul>	Pathological diagnosis of a neurosurgically resected tumor	No
Horiuchi et al. [8]	GPT-4® (OpenAI, USA); GPT-4V® (OpenAI, USA)	January 2014 to September 2023	—	No text reports (22 publication)	<ul style="list-style-type: none"> <li>• single-center</li> <li>• retrospective</li> </ul>	<ul style="list-style-type: none"> <li>• 2 radiologists with 4 and 7 years of experience</li> </ul>	Reported diagnoses	No
Grolleau et al. [11]	BERT® (Google, USA)	2020	—	—	<ul style="list-style-type: none"> <li>• single-center</li> <li>• retrospective</li> </ul>	—	Expert interpretation of reports	No
Han et al. [13]	GPT-4® (OpenAI, USA); GPT-3.5® (OpenAI, USA); Llama 2® (Meta, USA); Med42® (Hippocratic AI, USA); GPT-4V® (OpenAI, USA); Gemini Pro® (Google DeepMind, UK)	January 2017 to August 2023	Clinical cases from the Journal of the American Medical Association (JAMA); clinical images from the New England Journal of Medicine	—	<ul style="list-style-type: none"> <li>• multicenter</li> <li>• retrospective</li> </ul>	—	Uncertain	No
Horiuchi et al. [14]	GPT-4® (OpenAI, USA)	October 2021 to September 2023	—	—	<ul style="list-style-type: none"> <li>• multicenter</li> <li>• retrospective</li> </ul>	—	Reported diagnoses	No
Wataya et al. [15]	BERT® (Google, USA)	2020	Main body of the reports	No brain or skull description (10 reports)	<ul style="list-style-type: none"> <li>• single-center</li> <li>• retrospective</li> </ul>	<ul style="list-style-type: none"> <li>• 6 neuroradiologists with 4, 4, 4, and 6 years of experience (reference test)</li> </ul>	Manual report annotation by radiologists	No
Kaya et al. [10]	GPT-4® (OpenAI, USA)	—	—	Significant artifacts or poor image quality that interfere with diagnosis	<ul style="list-style-type: none"> <li>• multicenter (8 sites)</li> <li>• retrospective</li> </ul>	<ul style="list-style-type: none"> <li>• 3 radiologists with 1, 2 and 4 years of experience</li> <li>• 2 radiologists with 8 and 10 years of experience (reference test)</li> </ul>	Report interpretation by radiologists	No
Khoruzhaya et al. [12]	<ul style="list-style-type: none"> <li>• decision tree;</li> <li>• random forest;</li> <li>• logistic regression;</li> <li>• nearest neighbor algorithm;</li> <li>• support vector machines;</li> <li>• Catboost;</li> <li>• XGboost; MedRuBERTiny2</li> </ul>	—	—	—	<ul style="list-style-type: none"> <li>• multicenter (56 sites)</li> <li>• retrospective</li> </ul>	—	Uncertain	No
Cagnina et al. [16]	ChatGPT-4® (OpenAI, CLIA)	2020 and 2022	All patients who require invasive coronary angiography	—	<ul style="list-style-type: none"> <li>• single-center</li> <li>• prospective</li> </ul>	—	Results of invasive coronary angiography	No

Table 3. Sample characteristics and reported conditions

Authors	Patients / reports, n	Age groups, years	Sex ratio	Race	Representation of condition	Conflict of interests
Mitsuyama et al. [9]	150 radiology reports	Site A: 53.0 ± 17.0; Site B: 69.0 ± 15.0	56 men	No	Site A: meningioma (34); pituitary adenoma (17); neurilemmoma (12); angioma (5); craniopharyngioma (4); hemangioblastoma (4); high-grade glioma (10); low-grade glioma (3); epidermoid cyst (2); sarcoma (2); arachnoid cyst (1); chordoma (1); lymphoma (1); metastatic tumors (1); Rathke cleft cyst (1); central neurocytoma (1) Site B: meningioma (16); pituitary adenoma (6); neurilemmoma (4); angioma (0); craniopharyngioma (0); hemangioblastoma (1); high-grade glioma (5); low-grade glioma (1); epidermoid cyst (0); sarcoma (0); arachnoid cyst (0); chordoma (0); lymphoma (10); metastatic tumors (6); Rathke cleft cyst (2); central neurocytoma (0)	No
Horiuchi et al. [8]	106 cases of musculoskeletal diseases in Skeletal Radiology (Test Yourself)	No	No	No	According to the 2020 World Health Organization Classification of Tumors of Soft Tissue and Bone, all cases were divided into two groups: neoplastic ( $n = 45$ ) and non-neoplastic ( $n = 61$ ) Neoplastic cases: bone tumors ( $n = 24$ ) and soft tissue tumors ( $n = 22$ ) Non-neoplastic cases (classified based on their origin): muscle / soft tissue / nerve ( $n = 12$ ), arthritis / arthropathy ( $n = 10$ ), infection ( $n = 8$ ), congenital/developmental anomalies and dysplasia ( $n = 6$ ), trauma ( $n = 6$ ), metabolic disease ( $n = 5$ ), anatomical trait ( $n = 4$ ), and others ( $n = 10$ )	No
Grolleau et al. [11]	101,703 CT scan descriptions	64.7 ± 19.6	55.2% men	No	Reports mentioning nodules: 971 (48.8%); reports not mentioning nodules: 1020 (51.2%)	No
Han et al. [13]	Uncertain	No	No	No	—	Yes
Horiuchi et al. [14]	100 scan descriptions	No	No	No	All cases were classified by anatomical site: brain ( $n = 77$ ), spine ( $n = 11$ ), and head and neck ( $n = 12$ ) Cases involving brain injury were classified as either tumors of the central nervous system ( $n = 19$ ) or tumors of other sites ( $n = 58$ )	No
Wataya et al. [15]	3738 head CT reports	No	No	No	The percentages of each type of report in the entire dataset were as follows: no findings, 15.0%; minor findings, 26.7%; routine follow-up, 44.2%; careful follow-up, 7.7%; examination or therapy, 6.4%	Yes
Kaya et al. [10]	396 patients	Myocarditis: 38.6 ± 17.7 No myocarditis: 44.4 ± 17.6	Myocarditis: 41 women and 130 men No myocarditis: 97 women and 128 men	—	Non-ischemic cardiomyopathies (60.1%); chemotherapy-induced toxicity (0.3%); dilated cardiomyopathy (6.1%); hypertrophic cardiomyopathy (3.0%); left ventricular non-compaction (0.3%); myocarditis (41.2%); pericarditis (5.8%); sarcoidosis (0.8%); Takotsubo cardiomyopathy (1.8%); ischemic cardiomyopathy (5.8%); acquired heart disease (1.5%); indeterminate findings (1.5%); and no findings (36.1%)	Yes
Khoruzhaya et al. [12]	34,188 brain CT reports	No	No	No	Of 1194 reports in the test sample, 927 showed no signs of intracranial hemorrhage and 267 showed signs of intracranial hemorrhage	No
Cagnina et al. [16]	86 patients	62.0 ± 13.0	27% women	No	—	No

and imaging data. Han et al. [13] compared the performance of four LLMs: GPT-4<sup>®</sup> and GPT-3.5<sup>®</sup> (OpenAI, USA), Llama 2<sup>®</sup> (Meta, USA), Med42<sup>®</sup> (Hippocratic AI, USA) (the last two were open source), and two multimodal models that process both text and images: GPT-4V<sup>®</sup> (OpenAI, USA) and Gemini Pro (Google DeepMind, UK). Khoruzhaya et al. [12] compared the diagnostic performance of seven machine learning-based models, their ensemble combinations, and the BERT<sup>®</sup> model (Google, USA).

Four studies [8–10, 13] compared the diagnostic performance of LLMs and radiologists.

Three papers declared conflicts of interest [10, 13, 15], and six papers declared no conflicts of interest. The included studies did not evaluate either the time required for clinical implementation of LLMs or their potential cost-effectiveness.

### Supplementary Search Results

Fig. 2 shows the results of the supplementary systematic search. Supplement 1 summarizes key characteristics of the selected publications and related studies. Most of these studies (78 publications, or 36.1%) were conducted in the USA, followed by Germany and Japan (23 publications each, or 10.6%), and the People's Republic of China (20 publications, or 9.3%). The analysis included a total of 216 publications from 25 countries.

The majority of the papers (185, or 85.6%) used text information, which was the predominant type of data in the included studies. Sixteen studies (7.4%) used medical images, 14 (6.5%) used a combination of text and imaging data, and 1 (0.5%) evaluated audio recordings.

CT was the most frequently used imaging modality (80 publications, or 37.0%), followed by MRI (55 publications, or 25.5%), and radiography (39 publications, or 36.1%). Some studies also used findings from ultrasounds, mammograms, positron emission tomography–computed tomography (PET/CT), and scintigraphy.

There were nine main practical applications of LLMs in radiological diagnostics. The most common application involved rewording radiology reports to make them more

accessible and understandable for patients (Table 4). Additionally, 21 studies (9.7%) compared the performance of different LLMs in various diagnostic applications. These papers were included in the review because they help determine optimal models for radiological diagnostics applications. Note that their performance may depend on the specific task, input data, and how well the model is adapted to the clinical context. These studies suggest that larger, more recent models tend to be more accurate.

A significant percentage of studies on the clinical use of LLMs have poor methodology and result presentation. Although several sets of recommendations are available to standardize AI in healthcare research and its reporting [17, 18], only 2.3% of the papers found through a supplementary search complied with these recommendations.

### Diagnostic Accuracy of Models and Radiologists

The reviewed papers demonstrated the wide range of target conditions and tasks that LLMs can perform in radiological diagnostics. The diagnostic performance for each application is discussed below.

Three studies [8–10] evaluated diagnoses based on radiology reports. The obtained values of diagnostic performance ranged from low to high. All three papers evaluated the diagnostic performance of GPT-4<sup>®</sup> (OpenAI, USA). Mitsuyama et al. [9] demonstrated that this model performed relatively well in neuroradiology (for diagnosing brain tumors), achieving accuracy rates of 73.0% for final diagnosis and 94.0% for differential diagnosis involving three possible options. Radiologists demonstrated comparably high accuracy rates: 65.0%–79.0% and 73.0%–89.0%, respectively. Horiuchi et al. [8] evaluated the use of GPT-4<sup>®</sup> (OpenAI, USA) for diagnosing musculoskeletal diseases and obtained lower accuracy rates: 43.0% for final diagnoses and 58.0% for differential diagnoses involving three possible options. Radiologists also demonstrated low accuracy rates: 41.0%–53.0% and 58.0%–67.0%, respectively. The differences in accuracy rates between the model and the radiologists

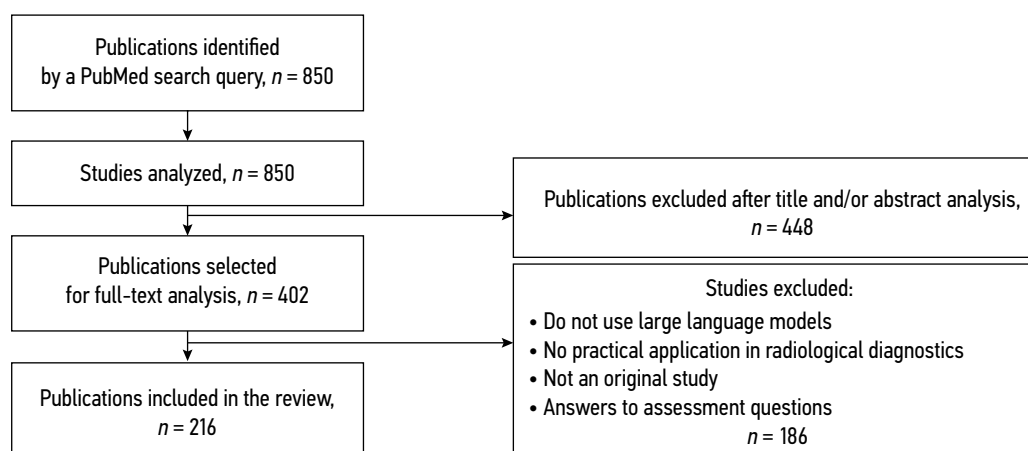


Fig. 2. Flowchart of the supplementary systematic search for publications.

**Table 4.** Applications of large language models in radiological diagnostics

Scenario	No. of papers, n (%)	Positive effect	Limitations
Simplifying radiology reports	76 (35.2)	Enhanced patient access to information and reduced communication barriers	<ul style="list-style-type: none"> <li>Loss of accuracy and detail when simplifying complex medical information</li> </ul>
Identifications of abnormalities	31 (14.3)	Early disease detection and higher diagnostic accuracy	<ul style="list-style-type: none"> <li>False positive and false negative results</li> <li>Need for clinical verification</li> </ul>
A differential diagnosis	27 (12.5)	Decision support in complex cases and reduced time to diagnosis	<ul style="list-style-type: none"> <li>Risk of suggesting incorrect diagnoses</li> <li>Dependence on training data quality</li> </ul>
Extraction of structured data	22 (10.2)	Accelerated analysis of large datasets and enhanced analytics	<ul style="list-style-type: none"> <li>Errors in automatic annotation</li> <li>Need for expert validation of data</li> </ul>
Generation of radiology reports	15 (6.9)	Reduced report turnaround time, standardization, and increased productivity	<ul style="list-style-type: none"> <li>Errors in complex cases and potential loss of an individualized approach</li> </ul>
Classification of studies by data assessment systems (e.g., PI-RADS or BI-RADS)	11 (5.1)	Improved accuracy of diagnostic categorization and standardization of clinical decisions	<ul style="list-style-type: none"> <li>Risk of errors in non-standard cases</li> <li>Dependence on input data quality</li> </ul>
Automation of data processing and interpretation in radiology	6 (2.8)	Automation of complex processing of medical texts and images, improved work efficiency	<ul style="list-style-type: none"> <li>Need for strict monitoring of the models' operation, risk of misinterpretation because of errors in clinical data</li> </ul>
Prognosis of disease outcomes	4 (1.8)	Improved treatment planning	<ul style="list-style-type: none"> <li>Challenges with model generalizability across diverse populations; the need for big data</li> </ul>
Detection of errors in radiology reports	3 (1.4)	Improved medical documentation and reduced likelihood of errors	<ul style="list-style-type: none"> <li>Not all errors can be detected automatically</li> <li>Risk of missing rare errors</li> </ul>

*Note.* PI-RADS, Prostate Imaging Reporting and Data System; BI-RADS, Breast Imaging Reporting and Data System.

were not statistically significant ( $p = 0.78$  and  $p = 0.99$  for a trainee radiologist, and  $p = 0.22$  and  $p = 0.26$  for a certified radiologist; according to the chi-square test). In this study, the multimodal GPT-4V<sup>®</sup> model (OpenAI, USA) proved ineffective, achieving accuracy rates of 8.0% and 14.0%, respectively. However, using GPT-4<sup>®</sup> (OpenAI, USA) as an auxiliary tool increased radiologists' accuracy by 3.0%–6.0%, but the significance of this increase was not reported. Similar use of GPT-4V<sup>®</sup> (OpenAI, USA) did not improve accuracy in any case.

Kaya et al. [10] demonstrated high performance of GPT-4<sup>®</sup> (OpenAI, USA) in diagnosing myocarditis. The model achieved sensitivity, specificity, and accuracy rates of 90.0%, 78.0%, and 83.0%, respectively. These rates were comparable to those of radiologists with one year of experience (90.0%, 84.0%, and 86.0%, respectively;  $p = 0.14$ , chi-squared test). However, they were statistically inferior to those of more experienced medical professionals (86.0%–85.0%, 91.0%–96.0%, and 89.0%–91.0%, respectively;  $p < 0.01$ ).

Two studies evaluated the automatic detection of abnormal findings in reports using the BERT<sup>®</sup> model (Google, USA) [11, 12]. In both cases, LLMs showed high diagnostic performance. For example, Grolleau et al. [11] reported that the model achieved a sensitivity, specificity, and accuracy of 98.0%, 99.0%, and 99.0%, respectively, in pulmonary nodule detection. Similarly, Khoruzhaya et al. [12] found that the BERT<sup>®</sup> model (Google, USA) had a sensitivity and specificity of 97.0% and 90.0%, respectively, for identifying signs of intracranial hemorrhage. These values were significantly higher than those obtained from

other machine learning techniques that were trained using the same data. For example, the median sensitivity and specificity were 94.0% and 78.0% for all machine learning models using different text vectorization techniques, and 94.0% and 84.0% for ensemble machine learning models. The differences were significant ( $p < 0.05$ , McNemar's test).

Han et al. [13] evaluated the performance of LLMs in answering clinical questions using text and images. However, the article did not provide more detailed information about the content of the questions used to test the models. The comparative analysis used four LLMs: GPT-4<sup>®</sup> and GPT-3.5<sup>®</sup> (OpenAI, USA); Llama 2<sup>®</sup> (Meta, USA), and Med-42<sup>®</sup> (Hippocratic AI, USA), and two multimodal models that process both text and images: GPT-4V<sup>®</sup> (OpenAI, USA) and Gemini Pro<sup>®</sup> (Google DeepMind, UK). The results showed that GPT-4V<sup>®</sup> (OpenAI, USA) achieved accuracy rates of 73.3% and 88.7% in answering 140 and 348 questions from *the Journal of the American Medical Association (JAMA)* and *The New England Journal of Medicine (NEJM)*, respectively. These results were significantly better than those of other models (median accuracy rates of 53.6% and 51.4%, respectively;  $p < 0.05$ , t-test) and radiologists (51.4% for *NEJM* questions). These high accuracy rates demonstrated by GPT-4V<sup>®</sup> (OpenAI, USA) contrast sharply with the results of Horiuchi et al. [8].

Horiuchi et al. [14] reported low accuracy rates of GPT-4<sup>®</sup> (OpenAI, USA) in diagnosing nervous system disorders of various etiologies based on medical records and clinically relevant findings in radiology reports. The accuracy rates were 50.0% and 63.0% for establishing final and differential diagnoses, respectively. These results are significantly

lower than those reported by Mitsuyama et al. [9]. Horiuchi et al. [14] did not find any significant differences in model accuracy rates based on the anatomical location of the abnormal process (brain, spinal cord, or head and neck region;  $p = 0.89$ , Fisher exact test). However, the authors reported significant differences in accuracy rates depending on the type of neoplasm. Accuracy rates for establishing final and differential diagnoses were 16.0% and 26.0%, respectively, for central nervous system tumors, and 62.0% and 74.0%, respectively, for tumors of other sites ( $p < 0.01$ ).

Wataya et al. [15] compared the ability of various AI models to classify the clinical importance of brain CT reports using a special five-level scale:

- category 0: no findings
- category 1: minor findings
- category 2: routine follow-up
- category 3: careful follow-up
- category 4: examination or therapy.

Four models were compared:

- logistic regression
- bidirectional long-short-term memory (BiLSTM)
- a general BERT® model (Google, USA) trained using non-medical Japanese texts from Wikipedia
- a domain-specific BERT® model (Google, USA) trained using Japanese medical records from the University of Tokyo Hospital database.

The paper presented multiple paired comparisons without adjusting the significance level for the multiple tests. Both variants of the BERT® language model (Google, USA) demonstrated significant superiority over logistic regression (78.7%) and the BiLSTM model (76.5%). The general BERT® model achieved an accuracy of 81.6%, whereas the domain-specific BERT® model achieved an accuracy of 84.3% ( $p = 0.001$ – $0.020$  for different pairwise comparisons; Mann–Whitney U-test). Although the difference in accuracy rates between two BERT® models (Google, USA) did not reach significance ( $p = 0.06$ ), the authors reported that the domain-specific model performed best.

Cagnina et al. [16] evaluated the diagnostic performance of GPT-4® (OpenAI, USA) in determining the need for invasive procedures in patients with acute coronary syndrome, based on clinical data and CT reports. The model demonstrated high sensitivity and accuracy rates of 95.0% and 86.0%, respectively, with moderate specificity of 64.0%.

### Assessment of Study Methodology Quality

Table 5 presents the results of the methodological quality assessment for the reviewed studies using a modified QUADAS-CAD tool.

Most of the reviewed papers (88.9%) were susceptible to risk of bias and overestimation of diagnostic performance of LLMs (Fig. 3). The reasons were as follows.

- First, the size and composition of the samples should be noted. A substantial percentage of the studies (66.7%) used samples with a few cases. Additionally, the samples

are often imbalanced in terms of both demographic characteristics (88.9%) and the representation and severity of target conditions (77.8%). One study does not indicate which clinical conditions were used [13], which suggests the risk of bias in domain D1 (Patient Selection). Data selection criteria were unclear in 55.6% of the studies because of poor reporting or complete absence of information.

- Second, some studies (33.3%) that used open-source data reported that the training and test samples may overlap [8, 13, 15]. This obviously overestimates the diagnostic performance of the relevant LLM. This is a signal question in domain D2 (Index Test).
- Third, most studies did not justify the size of the included datasets. Samples were typically selected at random or based on the data availability.
- Fourth, the procedure for creating the reference standard was not fully clear in several studies: 22.2% did not report how the standard was obtained [8, 13], 44.4% did not report whether a single standard was used [8, 12–14], and 33.3% did not disclose the qualifications of the medical professionals involved in preparing and verifying the reference standard [8, 12, 13].

Six out of nine (66.7%) of the reviewed studies have a high risk of a type I error, which occurs when statistically significant differences are found between groups that do not actually exist due to an absence of significance level adjustments for multiple comparisons [19, 20]. None of the studies that performed pairwise multiple comparisons [8, 10–15] made such an adjustment.

## DISCUSSION

Our two-stage search strategy (primary and supplementary) provided both a detailed review of studies with robust methodologies and transparent results and a comprehensive overview of diverse LLM use case-scenarios in radiological diagnostics.

### Application of Large Language Models in Radiological Diagnostics

Although only a few publications were selected for the scoping review during the primary search, LLMs can be used for various applications in radiological diagnostics. The most common diagnostic strategy is to analyze radiology reports, medical records, and clinically significant findings. Additionally, the LLM was used for detecting clinically significant findings in radiology reports, classifying reports, making decisions about the need for invasive procedures based on clinical data and reports, and assessing the accuracy of answers to clinical questions. The range of diagnosed conditions is also considerable. The reviewed studies included diagnoses of nervous system disorders (including tumors and intracranial hemorrhages), musculoskeletal diseases, lung cancer, and myocarditis.

**Table 5.** Methodological quality assessment using a modified QUADAS-CAD tool

Questions	Mitsuyama et al. [9]	Horiuchi et al. [8]	Grolleau et al. [11]	Han et al. [13]	Horiuchi et al. [14]	Wabaya et al. [15]	Kaya et al. [10]	Khoruzhaya et al. [12]	Cagnina et al. [16]
<i>Patient selection (D1)</i>									
<b>Were the training and test datasets balanced in terms of the severity (including the absence) of the target condition?</b>	No	No	Yes	Uncertain	No	No	No	Yes	No
<b>Were the training and test datasets balanced in terms of demographic factors?</b>	No	No	Yes	Uncertain	No	No	No	No	No
Did the study prevent any inappropriate exclusions?	Uncertain	Uncertain	Uncertain	Uncertain	Yes	Uncertain	Yes	Yes	Yes
<i>Index test (D2)</i>									
<b>If a neural network was used, were the training and testing datasets mutually exclusive?</b>	Yes	Uncertain	Yes	Uncertain	Yes	Uncertain	Yes	Yes	Yes
If a neural network was used, was the sample size for each dataset justified?	No	No	No	No	No	No	No	Yes	No
If a diagnostic threshold was used, was it pre-specified?	Uncertain	Uncertain	Uncertain	Uncertain	Uncertain	Uncertain	Uncertain	Yes	Yes
If a decision threshold was used (for AI), was it pre-specified?	Uncertain	Uncertain	Uncertain	Uncertain	Uncertain	Yes	Yes	Yes	Yes
<i>Reference standard (D3)</i>									
Is the reference standard able to correctly classify the target condition?	Yes	Uncertain	Yes	Uncertain	Yes	Yes	Yes	Yes	Yes
Was the reference standard established by experts with the necessary qualifications?	Yes	Uncertain	Yes	Uncertain	Yes	Yes	Yes	Uncertain	Yes
<i>Generation of results</i>									
Was the process of generating the results transparent?	Yes	Yes	Yes	Uncertain	Yes	Yes	Yes	Yes	Yes
Was the reference standard the same for all patient data?	Yes	No	Yes	Uncertain	Uncertain	Yes	Yes	Uncertain	Yes

**Note.** Signal questions are highlighted in bold. AI, artificial intelligence; QUADAS-CAD, Quality Assessment of Diagnostic Accuracy Studies-Computer-Aided Detection, a special, modified questionnaire for assessing the risk of bias and the applicability of studies in artificial intelligence technologies.

	Patient selection	Index test	Reference standard	Generation of results	Total score
Mitsuyama et al. [9]	−	?	+	+	?
Horiuchi et al. [8]	−	−	−	−	−
Grolleau et al. [11]	+	?	+	+	+
Han et al. [13]	−	−	−	−	−
Horiuchi et al. [14]	−	?	+	?	−
Wataya et al. [15]	−	−	+	+	−
Kaya et al. [10]	−	?	+	+	?
Khoruzhaya et al. [12]	?	+	?	?	?
Cagnina et al. [16]	−	+	+	+	?

+ Low                      − High  
? Some risks

**Fig. 3.** The risk of bias was assessed using a modified QUADAS-CAD questionnaire; QUADAS-CAD, Quality Assessment of Diagnostic Accuracy Studies-Computer-Aided Detection, a special, modified questionnaire for assessing the risk of bias and the applicability of studies in artificial intelligence technologies.

This work uses predominantly two models: GPT-4® (OpenAI, USA) and BERT® (Google, USA). Additionally, the analysis used GPT-3.5® (OpenAI, USA), Llama 2® (Meta, USA), and Med42® (Hippocratic AI, USA), and two multimodal models that process both text and images: GPT-4V® (OpenAI, USA) and Gemini Pro® (Google DeepMind, UK).

A supplementary search expanded the range of LLM applications and estimated their frequency using a large dataset. One of the most common applications in this case is simplifying radiology reports to reduce communication barriers between medical professionals and patients (for example, by creating dedicated chatbots). Applications involving the identification of abnormalities and differential diagnosis were reported less frequently and ranked second. The least common use-case scenarios for LLMs were predicting disease outcomes and identifying errors in radiology reports. Each reviewed scenario has its own limitations, including risks such as false positive and negative results, reduced accuracy, and the potential loss of an individualized approach. These limitations can be overcome by applying a robust LLM methodology to each use scenario and improving the methodological quality of the studies.

In most cases, LLMs primarily use text-based data, whereas the use of images or multimodal inputs is less

common. Audio data is rarely used. In terms of modality distribution, CT leads, followed by MRI and radiography. Ultrasound, mammography, PET/CT, and scintigraphy are less common. This distribution is typical of studies in radiological diagnostics.

### Diagnostic Accuracy of Large Language Models in Radiology

The diagnostic accuracy of LLMs can vary significantly across studies. Even the same model may perform differently when addressing different tasks. For example, the GPT-4® model (OpenAI, USA) demonstrated high diagnostic accuracy in identifying brain tumors and myocarditis based on radiology reports. The GPT-4® model (OpenAI, USA) showed high sensitivity and accuracy in making decisions about the need for invasive procedures in patients with acute coronary syndrome, based on clinical data and radiology reports, but its specificity was relatively low. However, the model demonstrated low diagnostic performance in identifying nervous system disorders of various etiologies (based on patient medical records and imaging data) and musculoskeletal diseases (based on imaging reports). GPT-4V® (OpenAI, USA) showed particularly low accuracy when diagnosing musculoskeletal diseases and was significantly inferior to GPT-4® (OpenAI, USA). In contrast, GPT-4V® (OpenAI, USA) was found to have the highest performance among the six models tested, including GPT-4® (OpenAI, USA), in terms of the quality of answers to clinical questions.

All studies using the BERT® model (Google, USA) demonstrated high diagnostic accuracy in detecting pulmonary nodules and signs of intracranial hemorrhage and classifying brain CT reports according to their clinical significance.

Two studies compared the diagnostic accuracy of BERT® (Google, USA) with that of other machine learning models in detecting signs of intracranial hemorrhage and classifying brain CT reports. BERT® demonstrated higher accuracy in both applications.

Four studies compared the diagnostic accuracy of the LLMs and radiologists. Two of these studies involved radiologists with various qualifications. In diagnosing myocarditis, the model performed comparably to novice radiologists but was outperformed by highly experienced medical professionals. In diagnosing musculoskeletal disorders, GPT-4® (OpenAI, USA) performed similarly to trainee and board-certified radiologists. In two other studies that did not specify the experience level of the radiologists, GPT-4® (OpenAI, USA) demonstrated comparable accuracy in diagnosing brain tumors, and GPT-4V® (OpenAI, USA) outperformed radiologists when answering clinical questions.

One study found that the diagnostic accuracy of GPT-4® (OpenAI, USA) varied greatly depending on the type of nervous system disorder being diagnosed. For example, the accuracy of diagnosing central nervous system tumors using radiology reports was significantly less accurate than diagnosing tumors in other sites.

## Evaluation of Methodological Quality of Studies on Using Large Language Models in Radiological Diagnostics

An assessment of the risk of bias in the selected papers based on the primary search revealed that most of them (88.89%) are at risk of bias and overestimation of diagnostic performance for various reasons.

Most studies used small samples. Additionally, the samples were imbalanced in terms of demographic characteristics and condition representation. In some cases, the distribution of conditions in samples is unclear.

Overlapping of training and test samples can lead to an overestimation of the models' diagnostic performance. When using open-source data, it is important to create samples that exclude cross-sections and avoid including cases present in both the training and test samples.

Many of the reviewed studies only reported accuracy rates, which are the least informative metric of diagnostic performance, and did not report other key metrics, such as sensitivity and specificity. Without these parameters, it is difficult to estimate the percentage of correctly classified true positive and true negative results, which is crucial for clinical practice.

Most papers presented pairwise comparisons of diagnostic performance metrics for different models, models and radiologists, and different target conditions. However, none of these papers adjusted the significance level for multiple comparisons. This may substantially overestimate the significance level of the differences found. At least, the results should be reported and interpreted with statistical adjustments for multiple comparisons. Additionally, the number of comparisons varied across papers, which made it difficult to generalize and compare their results.

Compliance with the relevant recommendations and checklists [21–25] for evaluating the diagnostic accuracy of AI tools can help address these issues and greatly improve publication quality.

A review of the 216 studies identified in the supplementary search revealed that most of them were of low methodological quality in terms of compliance with current standards for studies of LLMs and result presentation.

### Limitations of Systematic Review

Studies were searched using two search and analysis engines in two languages (English and Russian), as well as the reference lists of the chosen articles. This approach did not cover all relevant publications. Instead, it provided

a representative sample reflecting the general trend. The studies vary widely in application, but their low quality at this stage complicates the formulation of generalized conclusions about the diagnostic accuracy of LLMs. Further accumulation [26, 27] and systematic analysis of data in each application are necessary.

## CONCLUSION

LLMs show promising results in radiological diagnostics and perform well in various applications, ranging from simplifying report texts to supporting clinical decisions. However, the accuracy of current diagnostic data varies depending on the application, and this data is often obtained with the high risk of bias. Therefore, widespread clinical implementation of LLMs is currently premature. Further improvement of methodological quality and standardization of methods for assessing diagnostic performance are required to obtain a reliable evidence base.

## ADDITIONAL INFORMATION



**Supplement 1:** List of the included studies from the additional search and their basic characteristics.  
doi: 10.17816/DD678373-4340320

**Author contributions:** Yu.A. Vasilev, A.V. Vladzimirskyy, O.V. Omelyanskaya: development of the research concept; R.V. Reshetnikov; O.G. Nanova, K.M. Arzamasov, M.R. Kodenko, R.A. Erizhokov, A.P. Pamova, S.R. Seradzhi, I.A. Blokhin, A.P. Gonchar, P.B. Gelezhe, D.A. Akhmedzyanova, Yu.F. Shumskaya: literature review, data analysis, writing the text of the manuscript; R.V. Reshetnikov; O.G. Nanova: text editing. All the authors approved the version of the manuscript to be published and agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

**Ethics approval:** Not applicable.

**Funding sources:** This article was prepared by a team of authors within the framework of a scientific and practical project in the field of medicine (No. EGISU: 125051305989-8) "A promising automated workplace of a radiologist based on generative artificial intelligence".

**Disclosure of interests:** The authors have no relationships, activities, or interests for the last three years related to for-profit or not-for-profit third parties whose interests may be affected by the content of the article.

**Statement of originality:** No previously published material (text, images, or data) was used in this work.

**Data availability statement:** All data generated during this study are available in the article and its supplementary material.

**Generative AI:** No generative artificial intelligence technologies were used to prepare this article.

**Provenance and peer-review:** This paper was submitted unsolicited and reviewed following the standard procedure. The peer review process involved two members of the editorial board and the in-house science editor.

## REFERENCES | СПИСОК ЛИТЕРАТУРЫ

1. Cherif H, Moussa C, Missaoui AM, et al. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. *JMIR Medical Education*. 2024;10:e52818. doi: 10.2196/52818 EDN: OFMTDE
2. Kim W, Kim BC, Yeom HG. Performance of large language models on the Korean Dental licensing examination: a comparative study. *International Dental Journal*. 2025;75(1):176–184. doi: 10.1016/j.identj.2024.09.002 EDN: JDFMDL

3. Busch F, Hoffmann L, dos Santos DP, et al. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*. 2024;35(5):2589–2602. doi: 10.1007/s00330-024-11107-6 EDN: PNFKNR
4. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. *Diagnostic and Interventional Imaging*. 2023;104(6):269–274. doi: 10.1016/j.diii.2023.02.003 EDN: FGMMTY
5. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*. 2018;169(7):467–473. doi: 10.7326/M18-0850
6. Vasilev YuA, Vladzimirskyy AV, Omelyanskaya OV, et al. *Methodological recommendations for preparing a systematic review*. Moscow: Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies; 2023. (In Russ.) EDN: XKXHDA
7. Kodenko MR, Vasilev YA, Vladzimirskyy AV, et al. Diagnostic accuracy of ai for opportunistic screening of abdominal aortic aneurysm in CT: a systematic review and narrative synthesis. *Diagnostics*. 2022;12(12):3197. doi: 10.3390/diagnostics12123197 EDN: ERWYPX
8. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *European Radiology*. 2024;35(1):506–516. doi: 10.1007/s00330-024-10902-5 EDN: JAHWFM
9. Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *European Radiology*. 2024;35(4):1938–1947. doi: 10.1007/s00330-024-11032-8 EDN: UHMLBQ
10. Kaya K, Gietzen C, Hahnfeldt R, et al. Generative Pre-trained Transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: A multicenter study. *Journal of Cardiovascular Magnetic Resonance*. 2024;26(2):101068. doi: 10.1016/j.jocmr.2024.101068 EDN: TSRLJX
11. Grolleau E, Couraud S, Jupin Delevaux E, et al. Incidental pulmonary nodules: Natural language processing analysis of radiology reports. *Respiratory Medicine and Research*. 2024;86:101136. doi: 10.1016/j.resmer.2024.101136 EDN: DHDPIX
12. Khoruzhaya AN, Kozlov DV, Arzamasov KM, Kremneva EI. Comparison of an ensemble of machine learning models and the BERT language model for analysis of text descriptions of brain CT reports to determine the presence of intracranial hemorrhage. *Sovremennye tehnologii v medicine*. 2024;16(1):27–36. doi: 10.17691/stm2024.16.1.03 EDN: AXXVVD
13. Han T, Adams LC, Bresslem KK, et al. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA*. 2024;331(15):1320–1321. doi: 10.1001/jama.2023.27861 EDN: KPFLZG
14. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. 2023;66(1):73–79. doi: 10.1007/s00234-023-03252-4 EDN: SRFGAA
15. Wataya T, Miura A, Sakisuka T, et al. Comparison of natural language processing algorithms in assessing the importance of head computed tomography reports written in Japanese. *Japanese Journal of Radiology*. 2024;42(7):697–708. doi: 10.1007/s11604-024-01549-9 EDN: VAKPBV
16. Cagnina A, Salihu A, Meier D, et al. Assessing the need for coronary angiography in high-risk non-ST-elevation acute coronary syndrome patients using artificial intelligence and computed tomography. *The International Journal of Cardiovascular Imaging*. 2024;41(1):55–61. doi: 10.1007/s10554-024-03283-9 EDN: JMBFSX
17. Gallifant J, Afshar M, Ameen S, et al. The TRIPPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*. 2025;31(1):60–69. doi: 10.1038/s41591-024-03425-5 EDN: KAPIXF
18. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: a technical report. *NEJM AI*. 2025;2(6). doi: 10.1056/Alp2401106
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1995;57(1):289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
20. Hollestein LM, Lo SN, Leonardi-Bee J, et al. MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies. *British Journal of Dermatology*. 2021;185(6):1081–1083. doi: 10.1111/bjd.20600 EDN: QQWVVP
21. Collins GS, Moons KGM, Dhiman P, et al. TRIPPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi: 10.1136/bmj-2023-078378 EDN: WSTQKK
22. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. doi: 10.1136/bmjopen-2016-012799
23. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi: 10.1136/bmj.h5527
24. Vasiliev YuA, Vlazimirskyy AV, Omelyanskaya OV, et al. Methodology for testing and monitoring artificial intelligence-based software for medical diagnostics. *Digital Diagnostics*. 2023;4(3):252–267. doi: 10.17816/DD321971 EDN: UEDORU
25. Vasilev YuA, Bobrovskaya TM, Arzamasov KM, et al. Medical datasets for machine learning: fundamental principles of standartization and systematization. *Manager Zdravookhranenia*. 2023; (4):28–41. doi: 10.21045/1811-0185-2023-4-28-41 EDN: EPGAMD
26. Vinogradova IA, Nizovtsova LA, Omelyanskaya OV. Innovative strategic session in the scientific activity of the Center for Diagnostics and Telemedicine. *Digital Diagnostics*. 2022;3(4):414–420. doi: 10.17816/DD111833 EDN: DLRLVI
27. Kalinina ML, Svitachev AP, Biswas D, Vishnu P. Comparison of awareness and attitudes toward artificial intelligence among Russian- and English-speaking students at Orenburg State Medical University. *Digital Diagnostics*. 2023;4(1S):62–65. doi: 10.17816/DD430346 EDN: DIKOYA

## AUTHORS' INFO

\* **Olga G. Nanova**, Cand. Sci. (Biology);

address: 24 Petrovka st, bldg 1, Moscow, Russia, 127051;

ORCID: 0000-0001-8886-3684;

eLibrary SPIN: 6135-4872;

e-mail: nanova@mail.ru

**Yuriy A. Vasilev**, MD, Cand. Sci. (Medicine);

ORCID: 0000-0002-5283-5961;

eLibrary SPIN: 4458-5608;

e-mail: npcmmr@zdrav.mos.ru

**Roman V. Reshetnikov**, Cand. Sci. (Physics and Mathematics);

ORCID: 0000-0002-9661-0254;

eLibrary SPIN: 8592-0558;

e-mail: ReshetnikovRV1@zdrav.mos.ru

## ОБ АВТОРАХ

\* **Нанова Ольга Геннадьевна**, канд. биол. наук;

адрес: Россия, 127051, г. Москва, ул. Петровка, д. 24, стр. 1;

ORCID: 0000-0001-8886-3684;

eLibrary SPIN: 6135-4872;

e-mail: nanova@mail.ru

**Васильев Юрий Александрович**, канд. мед. наук;

ORCID: 0000-0002-5283-5961;

eLibrary SPIN: 4458-5608;

e-mail: npcmmr@zdrav.mos.ru

**Решетников Роман Владимирович**, канд. физ.-мат. наук;

ORCID: 0000-0002-9661-0254;

eLibrary SPIN: 8592-0558;

e-mail: ReshetnikovRV1@zdrav.mos.ru

**Anton V. Vladzimirskyy**, MD, Dr. Sci. (Medicine);  
ORCID: 0000-0002-2990-7736;  
eLibrary SPIN: 3602-7120;  
e-mail: VladzimirskijAV@zdrav.mos.ru

**Kirill M. Arzamasov**, MD, Dr. Sci. (Medicine);  
ORCID: 0000-0001-7786-0349;  
eLibrary SPIN: 3160-8062;  
e-mail: ArzamasovKM@zdrav.mos.ru

**Olga V. Omelyanskaya**;  
ORCID: 0000-0002-0245-4431;  
eLibrary SPIN: 8948-6152;  
e-mail: o.omelyanskaya@npcmr.ru

**Maria R. Kodenko**, Cand. Sci. (Engineering);  
ORCID: 0000-0002-0166-3768;  
eLibrary SPIN: 5789-0319;  
e-mail: KodenkoMR@zdrav.mos.ru

**Rustam A. Erizhokov**, MD;  
ORCID: 0009-0007-3636-2889;  
eLibrary SPIN: 2274-6428;  
e-mail: ErizhokovRA@zdrav.mos.ru

**Anastasia P. Pamova**, MD, Cand. Sci. (Medicine);  
ORCID: 0000-0002-0041-3281;  
eLibrary SPIN: 5146-4355;  
e-mail: PamovaAP@zdrav.mos.ru

**Seal R. Seradzhi**;  
ORCID: 0009-0000-3990-6668;  
e-mail: SeradzhiSR@zdrav.mos.ru

**Ivan A. Blokhin**, MD, Cand. Sci. (Medicine);  
ORCID: 0000-0002-2681-9378;  
eLibrary SPIN: 3306-1387;  
e-mail: BlokhinIA@zdrav.mos.ru

**Anna P. Gonchar**, MD, Cand. Sci. (Medicine);  
ORCID: 0000-0001-5161-6540;  
eLibrary SPIN: 3513-9531;  
e-mail: GoncharAP@zdrav.mos.ru

**Pavel B. Gelezhe**, MD, Cand. Sci. (Medicine);  
ORCID: 0000-0003-1072-2202;  
eLibrary SPIN: 4841-3234;  
e-mail: GelezhePB@zdrav.mos.ru

**Dina A. Akhmedzyanova**, MD;  
ORCID: 0000-0001-7705-9754;  
eLibrary SPIN: 6983-5991;  
e-mail: AkhmedzyanovaDA@zdrav.mos.ru

**Yuliya F. Shumskaya**, MD;  
ORCID: 0000-0002-8521-4045;  
eLibrary SPIN: 3164-5518;  
e-mail: shumskayayf@zdrav.mos.ru

**Владзими́рский Анто́н Вячесла́вович**, д-р мед. наук;  
ORCID: 0000-0002-2990-7736;  
eLibrary SPIN: 3602-7120;  
e-mail: VladzimirskijAV@zdrav.mos.ru

**Арзамасов Кирилл Михайлович**, д-р мед. наук;  
ORCID: 0000-0001-7786-0349;  
eLibrary SPIN: 3160-8062;  
e-mail: ArzamasovKM@zdrav.mos.ru

**Омелянская Ольга Васильевна**;  
ORCID: 0000-0002-0245-4431;  
eLibrary SPIN: 8948-6152;  
e-mail: o.omelyanskaya@npcmr.ru

**Коденко Мария Романовна**, канд. техн. наук;  
ORCID: 0000-0002-0166-3768;  
eLibrary SPIN: 5789-0319;  
e-mail: KodenkoMR@zdrav.mos.ru

**Ерижоков Рустам Арсеньевич**;  
ORCID: 0009-0007-3636-2889;  
eLibrary SPIN: 2274-6428;  
e-mail: ErizhokovRA@zdrav.mos.ru

**Памова Анастасия Петровна**, канд. мед. наук;  
ORCID: 0000-0002-0041-3281;  
eLibrary SPIN: 5146-4355;  
e-mail: PamovaAP@zdrav.mos.ru

**Сераджи Сеал Рахмануддин**;  
ORCID: 0009-0000-3990-6668;  
e-mail: SeradzhiSR@zdrav.mos.ru

**Блохин Иван Андреевич**, канд. мед. наук;  
ORCID: 0000-0002-2681-9378;  
eLibrary SPIN: 3306-1387;  
e-mail: BlokhinIA@zdrav.mos.ru

**Гончар Анна Павловна**, канд. мед. наук;  
ORCID: 0000-0001-5161-6540;  
eLibrary SPIN: 3513-9531;  
e-mail: GoncharAP@zdrav.mos.ru

**Гележе Павел Борисович**, канд. мед. наук;  
ORCID: 0000-0003-1072-2202;  
eLibrary SPIN: 4841-3234;  
e-mail: GelezhePB@zdrav.mos.ru

**Ахмедзянова Дина Альфредовна**;  
ORCID: 0000-0001-7705-9754;  
eLibrary SPIN: 6983-5991;  
e-mail: AkhmedzyanovaDA@zdrav.mos.ru

**Шумская Юлия Федоровна**;  
ORCID: 0000-0002-8521-4045;  
eLibrary SPIN: 3164-5518;  
e-mail: shumskayayf@zdrav.mos.ru

\* Corresponding author / Автор, ответственный за переписку