Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detecting Errors in Coronary Computed Tomography Angiography Reports: Comparison Among Three Large Language Models and Human Reader

Jing Chen
Shanghai Jiao Tong University
Yihao Wang
Shanghai Jiao Tong University
Linlin Sun
Shanghai Jiao Tong University
Li Zhu
Shanghai Jiao Tong University
Huiyuan Zhu
Shanghai Jiao Tong University
Xingxing Cen
Shanghai Jiao Tong University
Hong Yu
yuhongphd@163.com
Shanghai Jiao Tong University

Research Article

Keywords: Large language models, Artificial intelligence, Coronary artery, Computed tomography

Posted Date: March 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8851444/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Research on the application of large language models (LLMs) in coronary computed tomography angiography (CCTA) is still in its early stages, with a lack of comprehensive comparative evaluations across multiple LLMs. This study aims to assess the performance of LLMs in identifying errors in CCTA reports, comparing their effectiveness with human reader, and evaluating their time and cost efficiency.

Methods

In this retrospective study, 600 radiology reports were collected from January to December 2024 at a single institution. A total of 373 errors, categorized into five common types (omission, insertion, spelling, confusion, and other), were intentionally introduced into 300 reports. Other 300 error-free reports were served as the reference standard. Three LLMs (DeepSeek-R1, DeepSeek-32B, and GPT-4o) and one senior radiologist, were tasked with identifying errors. The detection performance and reading time were evaluated using Pearson's chi-square tests and paired-sample t tests.

Results

DeepSeek-R1 achieved the highest detection rate among LLMs at 80.7% (301/373, 95% CI: 76.4, 84.4), surpassing GPT-4o (64.1%, 239/373, 95% CI: 59.1, 68.8) and DeepSeek-32B (71.3%, 266/373, 95% CI: 66.5, 75.7). specifically, DeepSeek-R1 outperformed or was non-inferior to the radiologist in omission (74.0% vs. 88.3%, $P = 0.069$), insertion (100% vs. 96%, $P = 0.24$), spelling (92.9% vs. 78.6%, $P = 0.048$), and confusion errors (100% vs. 100%, $P > 0.99$), despite a lower total error detection rate compared to radiologist (80.7% vs. 88.7%, $P = 0.006$). All LLMs processed markedly faster (6.8–72.4s vs 450.6s; all $P < 0.001$) and incurred significantly lower costs ($0.004 - $0.066 vs. $0.692 per report; all $P < 0.001$) than the radiologist.

Conclusion

DeepSeek-R1 can serve as an effective supplementary tool for proofreading in CCTA, offering benefits of efficiency and cost reduction.

Introduction

Radiology serves as a link between intricate medical imaging and clinical practice. Coronary computed tomography angiography (CCTA) reports involve details such as the degree and location of stenosis, structural abnormalities - whose accuracy directly influences revascularization decisions. Even minor discrepancies can alter patient outcomes.(1, 2) Yet these reports are still generated through a labor-intensive "resident draft - board certified review" cycle, further strained by rising workloads, high-pressure environments, and unreliable speech recognition.(1-4) Common mistakes in preliminary radiology reports consist of numerical, interpretation errors, omissions, insertions, and inconspicuous descriptor misregistration between the findings and impression sections.(1, 4-6) Proofreading radiology reports holds a crucial role in the radiologic workflow and is often a time-intensive task with a large workload.

The rise of large language models (LLMs) has transformed natural language processing and, through their ability to manage and interpret vast text data, offers significant opportunities to advance radiology.(7-11) In chest imaging, LLMs have already been applied to automate report generation(12-15), detect textual errors(16, 17), convert free text into structured data without disrupting radiologists' workflow(18), extract TNM classifications(19, 20), and track interval changes between consecutive imaging studies.(21, 22) GPT-4's proofreading performance rivals that of human readers in English settings and does not rely on human clinical experience(17). GPT-4 Omni (GPT-4o) represents the state-of-the-art for multilingual tasks.(23) However, previous research indicated that GPT performs worse in non-English environments than in English settings for certain tasks.(24) The evidence in non-English contexts is limited, and its application in CCTA remains unexplored.

DeepSeek, a recent and powerful generative AI model, is designed with advanced reasoning capabilities and powered by a Mixture of Experts framework. This model offers significant advantages over traditional LLMs by improving accuracy, interpretability, and transparency, making it especially valuable in clinical settings like oncology.(25-27) The performance of these models in CCTA is unknown.

We therefore conducted the study aiming to investigate the performance of DeepSeek (including DeepSeek-R1 and DeepSeek-32B) and GPT-4o in proofreading CCTA reports, as well as their reading time and cost, in comparison with human reader.

Materials and Methods

This single-center, retrospective study was approved by the ethics committee (IS24112), and informed consent was waived due to the retrospective design. No patient-identifiable information was provided to DeepSeek-R1, DeepSeek-32B or GPT-4o.

2.1 Study Approach and Data Collection

Between January 2024 and December 2024, 600 original CCTA radiology reports were selected from Shanghai Chest hospital, a tertiary cardiothoracic specialist hospital. All the reports were reviewed and corrected by two senior radiologists (L. Sun and H. Yu, with 18 and 27 years of experience, respectively) to ensure the accuracy. The data set was randomly divided into two groups using a free online randomization tool (https://www.randomizer.org): the correct group and the error group, each consisting of 300 reports. In the error group, five most common errors (shown in Table 1), summarized in previous

investigations(1, 4-6), were intentionally introduced by an attending radiologist (H. Zhu, with 9 years of experience), with a maximum of 2 errors per report. Finally, 373 errors were intentionally introduced. The study flowchart is illustrated in Fig 1.

2.2 Error and Time Calculations

Human Reader - Based on previous studies showing GPT-4's performance is comparable to humans, regardless of clinical experience(17), only one senior radiologist was selected. The senior radiologist, (J. Chen, with 12 years of experience), assessed each report for potential errors, while another attending (H. Zhu with 9 years of experience), used a stopwatch to measure the time taken for evaluation.

GPT-4o. - The LLM Application Programming Interfaces (API) provided by Microsoft Azure (https://admin-m2zzmgln-eastus2.openai.azure.com) was utilized.

DeepSeek-R1. - The LLM API interface provided by Tencent Cloud (https://api.lkeap.cloud.tencent.com/v1) was invoked.

DeepSeek-32B. - The LLM was deployed on the hospital's local server.

The LLM inference application was developed using a Python Script (3.12.2) to invoke an API for evaluation. In addition, the Python script records the time taken to invoke the API for LLM inference. Similar to human reader, the LLM is required to assess each radiology report for potential errors under a zero-shot prompting(10). The specific prompt is as follows: "Role: You are a senior radiologist. Task: I will now provide you with a coronary CTA radiology report, divided into 'Findings' and 'Impression' sections. Please review the report for any errors and assess the consistency between the 'Findings' and 'Impression', listing all discrepancies and potential issues."

2.3 Cost Calculation

To convert Chinese Yuan (RMB) to US Dollars, the current exchange rate of 1:0.1382 (as of March 8, 2025) is used.

Human Readers. - We referred to the statistical data from the Chinese job search website JobUI (https://www.jobui.com) and found that the annual salary for a senior radiologist in Shanghai is approximately USD 33,168.

DeepSeek-32B. - The hospital's local deployment is hosted on a server with 8 A30 graphics cards. We treat the server maintenance fee as an annual cost, which is approximately 10% of the purchase price, totaling around USD 4,837.

GPT-4o. - The API pricing for Microsoft Azure is set at USD 5 per million tokens for input and USD 20 per million tokens for output text.

DeepSeek-R1. - The pricing for Tencent Cloud's API is set at USD 0.5528 per million tokens for input and USD 2.2112 per million tokens for output, with the output price also covering the charge for the thought chain.

Statistical Analysis

All analyses were conducted using software (Python version 3.12.2). We evaluated the number of errors correctly detected and the time taken to process each radiology report as outcome measures. Paired - sample *t* tests were used to compare the time and cost of evaluating radiology reports by LLMs versus radiologist. Pearson's chi-square test was employed to compare the number of errors detected by LLMs and radiologist. The 95% Confidence Interval (CI) was calculated using the Wilson method.(28)

Bonferroni correction was applied to for multiple comparisons, and a two-sided $P < 0.05$ was considered statistically significant. Interrater reliability was assessed using the Cohen's kappa coefficient, as follows: 0.01-0.20, none to slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; and 0.8-1.00, almost perfect agreement. A higher kappa value indicates stronger agreement. Cohen *d* was used as the effect size for the differences in processing time, with *d* values of 0.2, 0.5, and 0.8 constituting small, medium, and large effects, respectively.

Results

Errors Detection Performance

In detecting omission, DeepSeek-R1 (57/77; 74%, 95% CI: 63.3, 82.5) outperformed DeepSeek-32B (53/77; 68.8%, 95% CI: 57.8,78.1; $P < 0.001$ and GPT-4o (44/77; 57.1%, 95% CI: 46, 67.6; $P < 0.001$) and showed comparable performance to that of the radiologist (68/77; 88.3%, 95% CI: 79.3, 93.7; $P = 0.069$), with no statistically significant difference. However, DeepSeek-32B and GPT-4o were significantly trailed behind radiologist ($P = 0.009$, $P < 0.001$, respectively). (Fig. 2A; Table 2). In detecting insertion, DeepSeek-R1 (75/75; 100%, 95% CI: 95.1, 100) outperformed DeepSeek-32B (74/75; 98.7%, 95% CI: 92.8, 99.8; $P < 0.001$) and GPT-4o (73/75; 97.3%, 95% CI: 90.8, 99.3; $P < 0.001$) and showed comparable performance to that of the radiologist (72/75; 96%, 95% CI: 88.9, 98.7; $P = 0.24$), with no statistically significant difference. DeepSeek 32B and GPT-4o showed similar performance to that of the radiologist, with no statistically significant difference ($P = 0.933$ and $P > 0.99$). (Fig. 2B; Table 2). In detecting spelling, DeepSeek-R1 (65/70; 92.9%, 95% CI: 84.3, 96.9) outperformed both DeepSeek-32B (42/70; 60%, 95% CI: 48.3, 70.7) and GPT-4o (34/70; 48.6%, 95% CI: 37.2, 60), with $P < 0.001$ for both comparisons. Additionally, DeepSeek-R1 demonstrated significantly better performance than the radiologist (55/70; 78.6%, 95% CI: 67.6, 86.6; $P = 0.048$). Both DeepSeek-32B and GPT-4o trailed behind the radiologist ($P = 0.051$ and $P < 0.001$, respectively). (Fig. 2C; Table 2). Regarding confusion, both DeepSeek-R1 and radiologist detected all confusion (83/83; 100%, 95% CI: 95.6, 100; $P > 0.99$). DeepSeek-32B (78/83; 94%, 95% CI: 86.7, 97.4) showed comparable performance to that of the radiologist ($P > 0.069$), while GPT-4o (72/83; 86.7%, 95% CI: 77.8, 92.4) significantly trailed behind the radiologist ($P = 0.003$). DeepSeek-R1 outperformed both DeepSeek-32B and GPT-4o ($P < 0.001$ for both). (Fig. 2D; Table 2). Finally, regarding other errors, DeepSeek-R1 (21/68;

30.9%, 95% CI: 21.2, 42.6) outperformed DeepSeek-32B (19/68; 27.9%, 95% CI: 18.7, 39.6; $P < 0.001$) and GPT-4o (16/68; 23.5%, 95% CI: 15, 34.9; $P < 0.001$) and showed significantly lower performance compared to the radiologist (53/68; 77.9%, 95% CI: 66.7, 86.2; $P < 0.001$). Both DeepSeek-32B and GPT-4o significantly trailed behind the radiologist ($P < 0.001$ for both). (Fig. 2E; Table 2). Overall, DeepSeek-R1 demonstrated the best performance among the LLMs with a total detection rate was 80.7% (301/373, 95% CI: 76.4, 84.4), outperforming both GPT-4o (64.1%, 239/373, 95% CI: 59.1, 68.8; $P < 0.001$) and DeepSeek-32B (71.3%, 266/373, 95% CI: 66.5, 75.7; $P < 0.001$). However, its performance was still lower than that of human radiologist, who achieved a detection rate of 88.7% (331/373, 95% CI: 85.1, 91.6; $P = 0.006$). DeepSeek-32B's total detection rate, followed by GPT-4o significantly lagged behind the human rate ($P < 0.001$). (Fig. 2F; Table 2).

Table 1: Definitions of each error type and corresponding correct descriptions in CCTA reports

Error category	Error definition	Contextual examples of artificial errors in the Findings Section	Context in the correct description of the Impression Section
Omission	Missing words resulting in incomplete or incorrect phrases	1.main coronary artery 2.calcified plaque	1.left main coronary artery 2.non-calcified plaque
Insertion	The inadvertent addition of inaccurate terms or phrases, such as unsuitable words, erroneous word replacements, insertions, or word misinterpretations	no severe stenosis	severe stenosis
Spelling	Misspelled words, or errors in manual text processing by radiologists, leading to typographical mistakes or incorrect selections of text for removal or modification	□□(moderate)	□□ (severe)
Confusion	Side confusion	right main coronary artery	left main coronary artery
Other	1. punctuation or 2. sequence number errors	1. comma "," 2. (2)	1.period "." 2.(1)

Table 2
Comparison of Error Detection Rates between LLMs and the Radiologist

	Omission		**Insertion**		**Spelling**		**Confusion**		**Other**		**To**
Reader	Detection Rate (%)	*P* Value*	Detection Rate (%)	*P* Value*	Detection Rate (%)	*P* Value*	Detection Rate (%)	*P* Value*	Detection Rate (%)	*P* Value*	De Ra
Radiologist	**88.3** (79.3,93.7) [68/77]	Reference	96.0 (88.9,98.7) [72/75]	Reference	78.6 (67.6,86.6) [55/70]	Reference	**100.0** (95.6,100) [83/83]	Reference	**77.9** (66.7,86.2) [53/68]	Reference	**88** (8 [3
DeepSeek-R1	74.0 (63.3,82.5) [57/77]	.069†‡	**100.0** (95.1,100) [75/75]	.24†‡	**92.9** (84.3,96.9) [65/70]	.048†‡	**100.0** (95.6,100) [83/83]	> .99†‡	30.9 (21.2,42.6) [21/68]	< .001†‡	80 (7 [3
DeepSeek-32B	68.8 (57.8,78.1) [53/77]	.009	98.7 (92.8,99.8) [74/75]	.933	60.0 (48.3,70.7) [42/70]	.051	94.0 (86.7,97.4) [78/83]	.069	27.9 (18.7,39.6) [19/68]	< .001	71 (6 [2
GPT-4o	57.1 (46,67.6) [44/77]	< .001	97.3 (90.8,99.3) [73/75]	> .99	48.6 (37.2,60) [34/70]	< .001	86.7 (77.8,92.4) [72/83]	.003	23.5 (15,34.9) [16/68]	< .001	64 (5 [2

Note. - Data in parentheses are 95% Confidence Interval; data in brackets are numerators/denominators. The Chi-square comparison showed statistically sig between multiple groups (P < 0.05),

and bonferroni correction was used to correct *P* values for multiple comparisons.

* The number of correctly detected errors by LLMs was compared with the radiologist by using Pearson's chi-square tests

The highest Error Detection Rates are highlighted in bold.

† P value < 0.05 against DeepSeek-32B

‡ P value < 0.05 against GPT4o

Out of the 300 error-free reports, DeepSeek-R1, DeepSeek-32B, GPT-4o mistakenly flagged 11 (3.7%), 15(5.0%), and 11 (3.7%) reports, respectively, as containing errors. False positive of all LLMs were significantly higher than those of radiologist (1.7%; $P < 0.001$, respectively, Table 3).

Table 3
Comparison of Incorrectly Flagged Radiology Reports between LLMs and the Radiologist

	Total	
Reader	Rate of Incorrectly Flagged Reports (%, Rate)	*P* Value*
Radiologist	1.7 (0.7,3.8) [5/300]	Reference
DeepSeek-R1	3.7 (2.1,6.5) [11/300]	< .001
DeepSeek-32B	5 (3.1,8.1) [15/300]	< .001
GPT-4o	3.7 (2.1,6.5) [11/300]	< .001

Note. - Data in parentheses are 95% CIs; data in brackets are numerators/denominators. Bonferroni correction was used to correct *P* values for multiple comparisons.

* The number of correctly detected errors by LLMs was compared with the radiologist by using Pearson's chi-square tests

Inter-observer variability

The interrater agreement between radiologist and LLMs was fair (Cohen kappa = 0.47–0.49), indicating that error detection did not follow a discernible pattern (Table 4).

Table 4
Interrater Agreement between LLMs and the Radiologist

Reader	deepseek-R1	DeepSeek-32B	GPT-4o	Attending
Deepseek-R1	...	0.68	0.63	0.49
DeepSeek-32B	0.68	...	0.66	0.47
GPT-4o	0.63	0.66	...	0.48
Radiologist	0.49	0.47	0.48	...

Note. - Data are Cohen's kappa values (0.01–0.20, none to slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81-1.00, almost perfect agreement).

Time Analysis

The mean time per report for DeepSeek-R1, DeepSeek-32B, and GPT-4o, was 72.4 ± 76.7 seconds, 23.0 ± 10.8 seconds, and 6.8 ± 6.2 seconds, respectively, all of which were faster than the radiologist (450.6 ± 128.9 seconds, $P < 0.001$, respectively, Fig. 3A, Table 5).

Table 5
Mean reading time and mean cost per radiology report

	Mean Reading Time			Mean Cost		
Reader	Per Radiology Report (sec)	Cohen *d*	*P* Value	Per Radiology Report ($)	Cohen *d*	*P* Value
Radiologist	450.6 ± 128.9	Reference	Reference	0.692 ± 0.198	Reference	Reference
DeepSeek-R1	72.4 ± 76.7	-3.56	< .001	0.005 ± 0.003	-4.91	< .001
DeepSeek-32B	23.0 ± 10.8	-4.67	< .001	0.004 ± 0.002	-4.92	< .001
GPT-4o	6.8 ± 6.2	-4.86	< .001	0.066 ± 0.023	-4.44	< .001

Note. - The data are presented as means ± SD. Cohen *d* was used as the effect size for the differences in reading time and cost with d values of 0.2, 0.5, and 0.8 constituting small, medium, and large effects, respectively. The average time and cost to process a radiology report by LLMs were compared with the radiologist using paired-sample *t* tests.

Cost Analysis

The mean cost per radiology report for the three models, DeepSeek-R1, DeepSeek-32B, and GPT-4o, was USD 0.005 ± 0.003, USD 0.004 ± 0.002, and USD 0.066 ± 0.023, respectively, all of which were lower than the radiologist (USD 0.692 ± 0.198, $P < 0.001$, Fig. 3B, Table 5).

Discussion

The present study evaluated the performance of DeepSeek-R1, DeepSeek-32B, and GPT-4o in detecting common errors and discrepancies in CCTA reports, assessing their potential to reduce time and cost, with comparisons to a senior radiologist. DeepSeek-R1 demonstrated a total detection rate of 80.7% (95% CI: 76.4, 84.4), outperforming GPT-4o (64.1%; 95% CI: 59.1, 68.8, $P < 0.001$) and DeepSeek-32B (71.3%; 95%CI: 66.5, 75.7, $P < 0.001$). In terms of detecting the five error categories, DeepSeek-R1 showed better or comparable performance in omission, insertion, spelling, and confusion, with detection rates of 74.0%,

100%, 92.9%, and 100%, respectively. The LLMs demonstrated significantly shorter reading times compared to radiologist (all $P < 0.001$). Similarly, the mean cost for processing radiology reports using various LLMs was significantly lower than that of the radiologist (all $P < 0.001$). As a proofreading tool, LLMs can be more cost-effective than human readers, as they are capable of processing large volumes of text quickly, without needing breaks.

GPT-4 is demonstrated to be capable of error detection comparable to that of board-certified radiologists.(17) However, the performance of LLMs varies significantly across different languages, with a notable distinction between English and other languages.(24) All reports in our research are in Chinese, and our results showed that a lower total detection rate of GPT-4o (64.1%) compared to the senior radiologist (88.7%, $P < 0.01$), DeepSeek-R1 (80.7%, $P < 0.001$), and DeepSeek-32B (71.3%, $P < 0.001$). Our findings are consistent with previous studies on the performance of LLMs under Chinese contexts, showing that GPT-4o's performance is inferior to that of DeepSeek R1 or senior radiologists.(23, 29) The reason of performance disparities between DeepSeek and GPT-4o may contributed to not only differences in algorithmic architecture but also language-specific optimizations, such as localized databases and medical terminology.(30) Our findings highlight the importance of multilingual training and the potential of DeepSeek R1 for Chinese healthcare applications, particularly in the field of radiology.

This study demonstrates that various types of errors in CCTA reports can be objectively assessed using Deepseek-R1. Our findings are consistent with those of other studies(17, 31–33) that explore the integration of LLMs into radiologic workflows. Although Deepseek-R1 demonstrated a lower total error rate compared to senior radiologist (80.7% vs. 88.7%), these errors were primarily attributed to minor issues, such as incorrect punctuation usage, which had little impact on the subsequent clinical process. However, in terms of more significant errors that affect report content, such as omission, insertion, spelling mistakes, and confusion, DeepSeek-R1 showed better or comparable performance. This finding aligns with previous literature.(29) Therefore, Deepseek-R1 can still serve as an excellent assistive tool.

Prior studies have indicated that LLMs might overrate the significance of identified revisions, possibly due to challenges in grasping the subtle nuances and contextual information used by expert radiologists.(34) In line with this, our findings indicate that false positive rate of LLMs is higher than that of the radiologist, likely due to the model's insufficient understanding of the relationship between CCTA medical terminology. This lack of comprehension may lead to descriptor misregistration, for example, erroneous labeling of a myocardial bridge as an intramyocardial coronary artery.

Our study has several limitations. First, the use of manually created predefined errors within a controlled experimental setting may not capture the full complexity and nuance of real-world clinical practice, thereby impacting the generalizability of our findings. Furthermore, the experimental design may have induced a Hawthorne effect(35), where participants' awareness of being observed could have increased the error detection rates compared to routine clinical scenarios. Second, the cost observed in our experimental setup does not fully capture the true cost of integrating LLMs into clinical workflows, as it overlooks additional expenses such as infrastructure and data protection.(36)

In conclusion, DeepSeek-R1 demonstrated higher or comparable detection rates compared to those of senior radiologist in the areas of omission, insertion, spelling, and confusion. This suggests its potential for application in clinical settings, offering benefits in terms of efficiency and cost reduction. Nevertheless, human verification remains a critical component of the workflow to ensure accuracy and address legal liabilities. Future research should prioritize refining the model's understanding of clinical context, specialized medical terminology, and subtle punctuation errors.

Abbreviations

CCTA
Coronary computed tomography angiography
LLMs
large language models
GPT-4o
GPT-4 Omni
API
Application Programming Interfaces
CI
Confidence Interval

Declarations

Ethics approval and consent to participate

This single-center, retrospective study was conducted in accordance with the ethical standards of the institutional review board and the Declaration of Helsinki, as revised in 2013. The study was approved by the Institutional Review Board of Shanghai Chest Hospital (Approval No: IS24112), and informed consent was waived due to the retrospective design.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the National Institute of Hospital Administration (YLXX24AIC003); Shanghai Jiao Tong University School of Medicine (WK2424); Interdisciplinary Program of Shanghai Jiao Tong University (YG2024QNA49); the Young Scientists Fund of the National Natural Science Foundation of China (82302188), National Key R&D Program of China (2021YFC2500700), Shanghai Health Research Foundation for Talents (2022YQ060), Shanghai Science and Technology Innovation Action Plan (22Y11911100), and Shanghai Innovative Medical Product Application Demonstration Project (24SF1904000).

Authors' contributions

JC: Writing – original draft, Methodology, Conceptualization. JC and YHW: Writing – review & editing, Methodology, Funding acquisition. LLS: Writing – review & editing, Formal analysis. LZ: Writing – review & editing. HYZ: Writing – review & editing, Supervision, Funding acquisition, Data curation. XXC: Writing – review & editing, Validation, Resources, Project administration. HY: Writing – review & editing, Supervision, Funding acquisition, Data curation. JC, YHW and LLS are co-first authors, and HYZ, XXC and HY are co-corresponding authors.

Acknowledgements

None.

References

1. Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. J Am Coll Radiol. 2008;5(12):1196–9.
2. Onder O, Yarasir Y, Azizova A, Durhan G, Onur MR, Ariyurek OM. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. Insights Imaging. 2021;12(1):51.
3. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol. 2015;22(9):1191–8.
4. Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, et al. Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. JAMA Netw Open. 2018;1(3):e180530.
5. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and Semantic Errors in Radiology Reports Associated With Speech Recognition Software. Stud Health Technol Inf. 2015;216:922.
6. Vosshenrich J, Nesic I, Cyriac J, Boll DT, Merkle EM, Heye T. Revealing the most common reporting errors through data mining of the report proofreading process. Eur Radiol. 2021;31(4):2115–25.
7. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098–102.
8. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y). 2024;5(3):100943.
9. Ufuk F. The Role and Limitations of Large Language Models Such as ChatGPT in Clinical Settings and Medical Journalism. Radiology. 2023;307(3):e230276.
10. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. Radiology. 2024;310(1):e232756.
11. Bluethgen C, Van Veen D, Zakka C, Link KE, Fanous AH, Daneshjou R, et al. Best Practices for Large Language Models in Radiology. Radiology. 2025;315(1):e240528.
12. Li Y, Wang Z, Liu Y, Wang L, Liu L, Zhou L, KARGEN. Knowledge-enhanced Automated Radiology Report Generation Using Large Language Models2024 September 01, 2024:[arXiv:2409.05370 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240905370L
13. Altalla B, Ahmad A, Bitar L, Al-Bssol M, Al Omari A, Sultan I. Radiology Report Annotation Using Generative Large Language Models: Comparative Analysis. Int J Biomed Imaging. 2025;2025:5019035.
14. Sun YH, Khor HG, Wang YZ, Wang ZH, Zhao HL, Zhang Y editors. Continually Tuning a Large Language Model for Multi-domain Radiology Report Generation. 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). 2024 Oct 06–10; Palmeraie Conf Ctr, Marrakesh, MOROCCO2024.
15. Hu D, Zhang S, Liu Q, Zhu X, Liu B. Large Language Models in Summarizing Radiology Report Impressions for Lung Cancer in Chinese: Evaluation Study. J Med Internet Res. 2025;27:e65547.
16. Sun C, Teichman K, Zhou Y, Critelli B, Nauheim D, Keir G, et al. Generative Large Language Models Trained for Detecting Errors in Radiology Reports. Radiology. 2025;315(2):e242575.
17. Gertz RJ, Dratsch T, Bunck AC, Lennartz S, Iuga AI, Hellmich MG, et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. Radiology. 2024;311(1):e232714.

18. Woźnicki P, Laqua C, Fiku I, Hekalo A, Truhn D, Engelhardt S, et al. Automatic structuring of radiology reports with on-premise open-source large language models. Eur Radiol. 2025;35(4):2018–29.
19. Matsuo H, Nishio M, Matsunaga T, Fujimoto K, Murakami T. Exploring Multilingual Large Language Models for Enhanced TNM Classification of Radiology Report in Lung Cancer Staging. Cancers (Basel). 2024;16(21).
20. Lee JE, Park KS, Kim YH, Song HC, Park B, Jeong YJ. Lung Cancer Staging Using Chest CT and FDG PET/CT Free-Text Reports: Comparison Among Three ChatGPT Large Language Models and Six Human Readers of Varying Experience. AJR Am J Roentgenol. 2024;223(6):e2431696.
21. Liu T, Wang J, Hu Y, Li M, Yi J, Chang X et al. HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation2024 December 01, 2024:[arXiv:2412.11070 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv241211070L
22. Mathai TS, Kim B, Stroie OM, Summers RM. Privacy-Preserving Large Language Model for Matching Findings and Tracking Interval Changes in Longitudinal Radiology Reports. J Imaging Inf Med. 2025.
23. Yan Y, Wang K, Feng B, Yao J, Jiang T, Jin Z, et al. The use of large language models in detecting Chinese ultrasound report errors. NPJ Digit Med. 2025;8(1):66.
24. Cozzi A, Pinker K, Hidber A, Zhang T, Bonomo L, Lo Gullo R, et al. BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study. Radiology. 2024;311(1):e232133.
25. Normile D. Chinese firm's large language model makes a splash. Science. 2025;387(6731):238.
26. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature. 2025;638(8050):300–1.
27. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-Redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc. 2025;33(5):1553–6.
28. Wallis S. Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. J Quant Linguistics. 2013;20(3):178–208.
29. Luo PW, Liu JW, Xie X, Jiang JW, Huo XY, Chen ZL, et al. DeepSeek vs ChatGPT: a comparison study of their performance in answering prostate cancer radiotherapy questions in multiple languages. Am J Clin Exp Urol. 2025;13(2):176–85.
30. Shao CY, Li H, Liu XL, Li C, Yang LQ, Zhang YJ, et al. Appropriateness and Comprehensiveness of Using ChatGPT for Perioperative Patient Education in Thoracic Surgery in Different Language Contexts: Survey Study. Interact J Med Res. 2023;12:e46900.
31. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A Context-based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology. 2023;308(1):e230970.
32. Russe MF, Fink A, Ngo H, Tran H, Bamberg F, Reisert M, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep. 2023;13(1):14215.
33. Salam B, Kravchenko D, Nowak S, Sprinkart AM, Weinhold L, Odenthal A, et al. Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand. J Cardiovasc Magn Reson. 2024;26(1):101035.
34. Lyo S, Mohan S, Hassankhani A, Noor A, Dako F, Cook T. From Revisions to Insights: Converting Radiology Report Revisions into Actionable Educational Feedback Using Generative AI Models. J Imaging Inf Med. 2025;38(2):1265–79.
35. Kidwai AS, Abujudeh HH. Radiologist Productivity Increases With Real-Time Monitoring: The Hawthorne Effect. J Am Coll Radiol. 2015;12(11):1151–4.
36. Arshad K, Ardalan S, Schreiweis B, Bergh B. Integrating an AI platform into clinical IT: BPMN processes for clinical AI model development. BMC Med Inf Decis Mak. 2025;25(1):243.

Figures

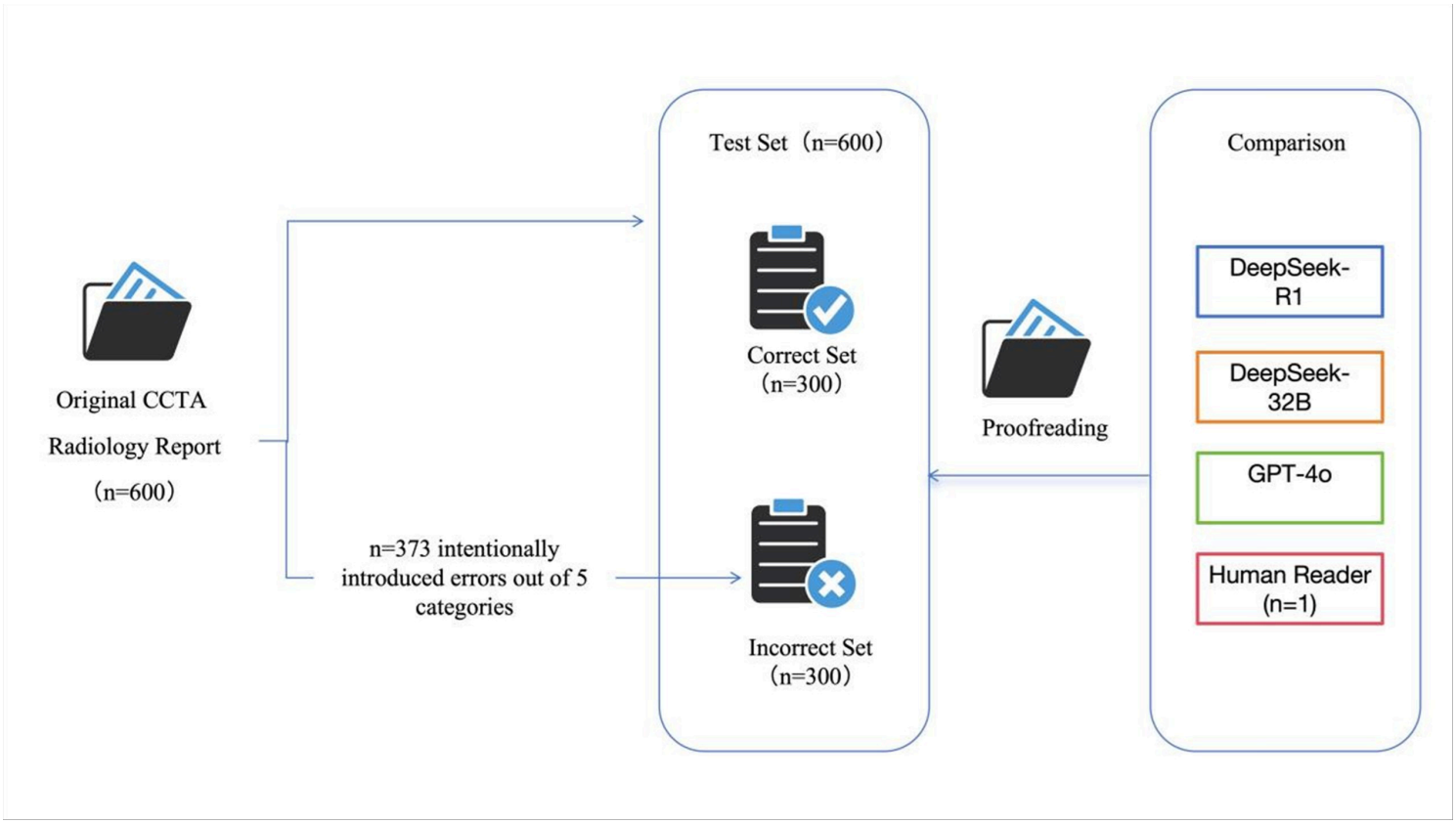


Figure 1

The flowchart of this study. A total of 600 original CTA radiology reports were randomly selected from Radiology Information System (Weining Health Technology Group Co., Ltd., Shanghai, China). These reports were then randomly divided into two groups: test set and incorrect set, with 300 reports in each. In the incorrect set, 373 errors were deliberately introduced, encompassing five categories (omission, insertion, spelling, confusion, and others), with a maximum of 2 errors per report. DeepSeek-R1, DeepSeek-32B, GPT-4o and one senior radiologist was tasked with evaluating each radiology report to identify potential errors, facilitating a comparative analysis of their performance.

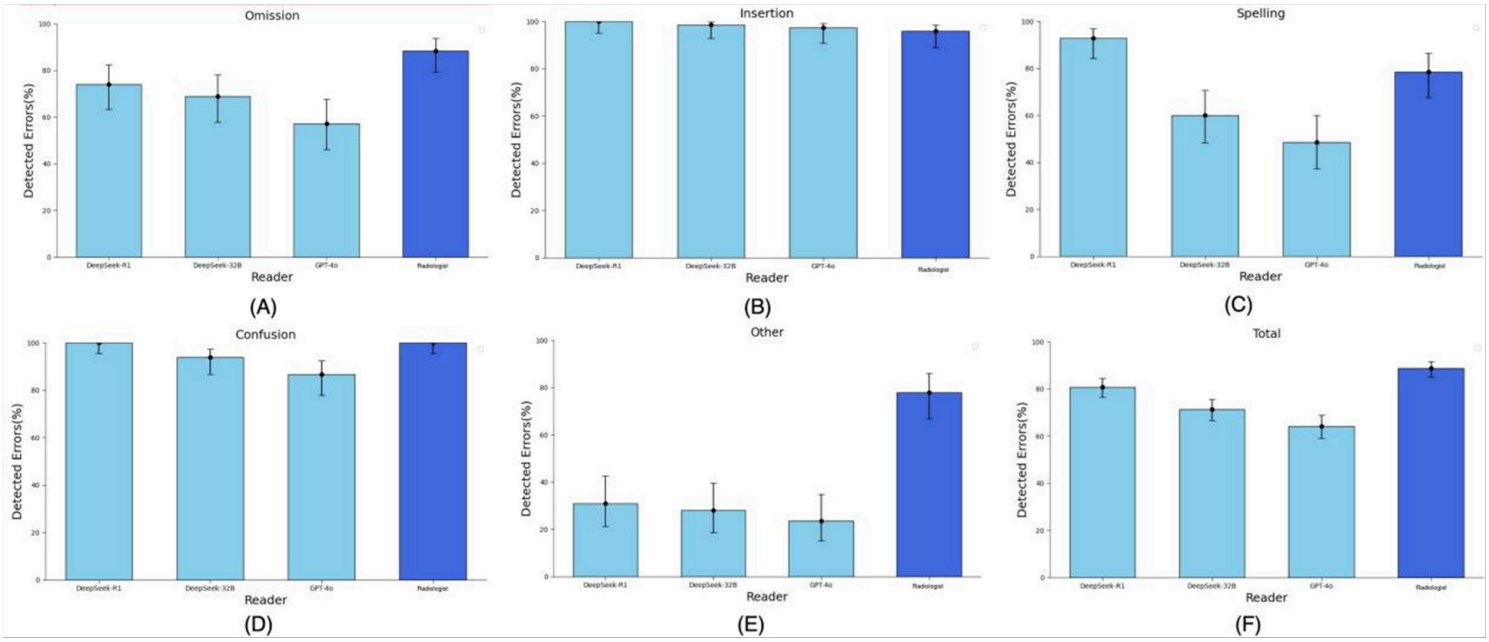


Figure 2

Error detection performance of different models in CCTA report proofreading. The bar chart illustrated the percentage of errors detected by the LLMs and radiologist. The error bars represent the 95% Confidence Intervals.

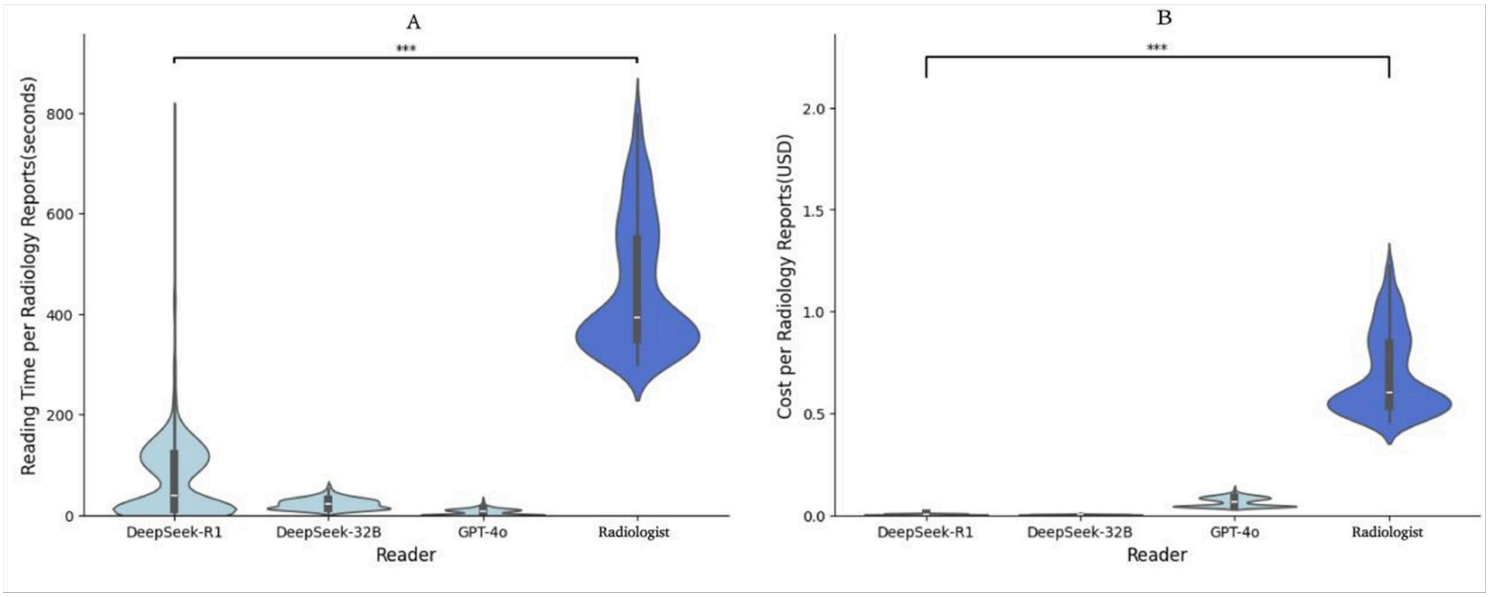


Figure 3

Distribution of reading times and costs per radiology report for different readers. **(A)** Violin plot shows reading time per radiology report in seconds. The white lines are the medians, and the black lines interval are the quartiles. *** $P < 0.001$.**(B)** Violin plot shows cost per radiology report in U.S. dollars. The white lines are the medians, and the black lines interval are the quartiles. *** $P < 0.001$.