Radiology: Artificial Intelligence

AI IN BRIEF

Generative Large Language Models for Detection of Speech Recognition Errors in Radiology Reports

Reuben A. Schmidt, MD, BA • Jarrel C. Y. Seah, MBBS • Ke Cao, PhD, MSc • Lincoln Lim, MBBS (Hons), MCHSM, FRSPH, FRSM, FRANZCR • Wei Lim, MBBS, FRCR, FRANZCR • Justin Yeung, MD, FRACS

From the Department of Medical Imaging, Western Health, Footscray, Australia (R.A.S., L.L., W.L.); Alfred Health, Harrison.ai, Monash University, Clayton, Australia (J.C.Y.S.); Department of Surgery, Western Precinct, University of Melbourne, Melbourne, Australia (K.C., J.Y.); and Department of Surgery, Western Health, Melbourne, Australia (J.Y.). Received June 13, 2023; revision requested July 24; revision received November 8; accepted January 10, 2024. **Address correspondence to** R.A.S. (email: *reuben.schmidt@icloud.com*).

Funding for project development was provided by the Western Health Department of Medical Imaging.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2024; 6(2):e230205 • https://doi.org/10.1148/ryai.230205 • Content code: AI

This study evaluated the ability of generative large language models (LLMs) to detect speech recognition errors in radiology reports. A dataset of 3233 CT and MRI reports was assessed by radiologists for speech recognition errors. Errors were categorized as clinically significant or not clinically significant. Performances of five generative LLMs—GPT-3.5-turbo, GPT-4, text-davinci-003, Llama-v2–70B-chat, and Bard—were compared in detecting these errors, using manual error detection as the reference standard. Prompt engineering was used to optimize model performance. GPT-4 demonstrated high accuracy in detecting clinically significant errors (precision, 76.9%; recall, 100%; F1 score, 86.9%) and not clinically significant errors (precision, 93.9%; recall, 94.7%; F1 score, 94.3%). Text-davinci-003 achieved F1 scores of 72% and 46.6% for clinically significant and not clinically significant errors, respectively. GPT-3.5-turbo obtained 59.1% and 32.2% F1 scores, while Llama-v2–70B-chat scored 72.8% and 47.7%. Bard showed the lowest accuracy, with F1 scores of 47.5% and 20.9%. GPT-4 effectively identified challenging errors of nonsense phrases and internally inconsistent statements. Longer reports, resident dictation, and overnight shifts were associated with higher error rates. In conclusion, advanced generative LLMs show potential for automatic detection of speech recognition errors in radiology reports.

Supplemental material is available for this article.

Accurate radiology report communication is essential for high-quality patient care. However, the use of speech recognition software for radiology reporting has been associated with increased error rates, with speech recognition errors observed in 20% to 60% of dictated reports (1–4). About 2% of reports may contain clinically significant speech recognition errors that risk misinterpretation and may impede care (3,5,6).

Efforts to reduce speech recognition errors include the use of structured reporting templates, which have shown some success but face variable acceptance by radiologists (1,7). More recently, deep learning approaches like neural sequence-to-sequence models (8) and bidirectional autoencoders (9,10) have been investigated for the detection of speech recognition errors. While promising, these methods rely on extensive customized training data and detect errors on only a per-sentence basis.

Recent advances in natural language processing present opportunities to overcome these limitations through the use of generative large language models (LLMs). These models can learn complex linguistic patterns and generate fluent, coherent text. Large context windows allow these models to process the complete textual content of lengthy radiology reports, which may offer advantages over sentence-by-sentence analysis.

Proprietary LLMs represent the state-of-the-art in performance, with OpenAI's GPT-4 showing promise in radiology applications such as transformation of free text to structured reports (11) and appropriate imaging request study protocoling (12,13). However, studies evaluating the use of generative LLMs for decision support in radiology remain lacking.

In this study, we evaluated the performance of five leading generative LLMs—GPT-3.5-turbo, GPT-4, text-davinci-003, Llama-v2–70B-chat, and Bard—for the automatic detection of speech recognition errors in radiology reports. We hypothesized that advanced LLMs can accurately flag such errors to offer automated error identification, potentially improving radiology report accuracy.

Materials and Methods

This study was approved by the Western Health Ethics Panel (HREC/23/WH/94984). Informed consent was waived for this retrospective study of existing radiology reports because all data were de-identified and patient care was not impacted.

Data Collection

The authors for this study analyzed 3233 de-identified radiology reports (2498 CT and 825 MRI). The reports were randomly selected from the picture archiving and communication system of a major tertiary hospital over a 12-month period. Stratified random sampling was used to select reports representing all radiologists (22 attending physicians, 23 residents) and examinations (79 study types covering all body systems). This established a representative sample of our institutional dataset.

Abbreviations

LLM = large language model, OR = odds ratio

Summary

GPT-4 showed high performance compared with other generative large language models for the detection of speech recognition errors in radiology reports, demonstrating the potential of such models to improve report accuracy.

Key Points

- GPT-4 demonstrated high accuracy in detecting clinically significant speech recognition errors (F1 score, 86.9%) and not clinically significant errors (F1 score, 94.3%) in radiology reports.
- GPT-4 effectively flagged challenging errors like internal inconsistencies (12 of 12, 100%) and nonsense phrases (22 of 24, 91.7%), which require assimilation of the entire report context and fluency in radiologic text.
- Increased error rates were associated with longer reports ($P <$ .001), trainee dictation (odds ratio [OR], 2.2; $P = .003$), and overnight shifts (OR, 1.9; $P = .04$), identifying high-yield areas for integrating error detection tools.

Keywords

CT, Large Language Model, Machine Learning, MRI, Natural Language Processing, Radiology Reports, Speech, Unsupervised Learning

Data related to the reports such as report length, imaging modality, study type, patient status (inpatient, outpatient, or emergency department patient), dictating radiologist training level, and shift type (day, evening, or overnight) were acquired from the picture archiving and communication system. Of the 3233 studies, 12 (0.37%) were excluded from the dataset due to missing data.

Reader Evaluation of Radiology Reports

A radiology resident (R.A.S.) (R2 level) performed manual review of the 3233 reports for speech recognition errors. Additional independent evaluation was conducted by two board-certified radiologists (W.L., L.L.) with more than 20 years of experience, who annotated a random subsample of 100 reports for errors.

Specific examples of speech recognition errors are detailed in the Table. Error severity was marked as clinically significant or not clinically significant according to the ontology of Chang et al (3) and as used in recent studies (9,14). Clinically significant errors are considered to change the meaning of the report and risk misinterpretation by the clinician. Examples included nonsense phrases ("The lungs nuclear" instead of "the lungs are clear"), omission of important words ("Intracranial hemorrhage" instead of "No intracranial hemorrhage"), and internally inconsistent statements ("left occipital lesion" referred to as "left parietal lesion" later in the report). The location and severity were recorded for each error in each report.

Data De-Identification

Prior to analysis by the generative LLM, reports were de-identified via a four-step process involving *(a)* automated removal of metadata fields containing identifiable information, *(b)* removal of all report sections except Findings and Conclusion (of most clinical relevance and least likely to contain personally identifiable information), *(c)* automated de-identification of reports with recently published machine learning methods (15), and *(d)* manual inspection.

Generative LLM Analysis

We developed a web application for comparing model outputs. This application provides an interface for entering sample reports, allowing side-by-side comparison of the original and corrected reports (Fig 1). Five generative LLMs were used via application programming interface calls within the application: OpenAI's text-davinci-003, GPT-3.5-turbo, and GPT-4; Meta's Llama-v2–70B-chat; and Google's Bard *(https://platform.openai.com/docs/models/gpt-3-5-turbo, https://openai.com/gpt-4, https://llama.meta.com/, https://bard.google.com)*. The Replicate application programming interface (16) was used to access Llama-v2–70B-chat. The average of three outputs for each model provided the report error classifications.

Prompt Engineering

The textual prompts provided to generative LLMs are known to influence model performance. To optimize prompts, a multistep approach adapted from the AdaTest method (17) was used (Table S1).

First, an extensive prompt-engineering phase was undertaken over 100 iterations for each model. Prompts were systematically modified with four distinct approaches: chain-of-thought (18) (requiring the model to answer in an explanatory step-by-step manner), few-shot learning (19) (providing example input and output as part of the prompt), grounding context (20) (giving the model lexical or contextual information that may be relevant to the subject), and variation in model "temperature" value (designed to allow variation in how deterministic a model's responses are, where 0 is more deterministic and 1 more stochastic). On each iteration, model outputs were assessed for five generated reports containing known errors.

Prompt optimization was validated by comparing model performance on the dataset of 100 reports annotated by three raters, using the base prompt "Correct this radiology report," versus prompts optimized in the engineering phase. The optimized prompts were used for all subsequent analyses.

Statistical Analysis

Statistical analysis was performed with Python version 3.9.6 and the libraries pandas (version 1.5.3), scikit-learn (version 1.2.2), SciPy (version 1.9.1), and statsmodels (version 0.14.0).

We employed a mixed effects logistic regression model to analyze the relationship between errors and other variables including report length, patient status, radiologist training level, and shift type. The reporting radiologist was included as a random effect, controlling for the possibility that radiologists may systematically differ in their error rates and report characteristics. A P value of less than .05 indicated a statistically significant difference. Cohen κ was used to assess interrater reliability for error classification.

Types of Speech Recognition Error

Error Type	Definition	Intended/Spoken Phrase	Clinically Significant Error	Not Clinically Significant Error
Nonsense	Passages/words/phrases that make no sense or have no sensible meaning.	The lungs are clear.	The lungs nuclear.	Lungs are clear stop.
Translational	Translation error that may change the meaning of a phrase/sentence.	There is no opacity in the left lung.	There is an opacity in the left lung.	There is no opacity in their left lung.
Omission	Words not transcribed (omitted) that may change the meaning of a phrase/sentence.	There is no opacity in the left lung.	There is opacity in the left lung.	There is no opacity in left lung.
Homonym error	Misuse of words or phrases that sound the same but are semantically distinct.	The right lung is clear.	The write lung is clear.	Thee right lung is clear.
Grammatical error	Standard grammatical errors including the use of sentence fragments.	The lungs are clear.	The lungs is clear.	The lungs clear.
Template error	Retained statement from a standardized template that contradicts the dictated findings or impression.	There is a hazy opacity obscuring the right hemidiaphragm.	The lungs are clear. There is a hazy opacity obscuring the right hemidiaphragm.	The lungs are clear. Both lungs are clear.
Extraneous statements	Fragments of discussion that are included inadvertently in the radiology report.	The right lung is clear.	The right lung is clear. We may have a problem.	The right lung is clear. The.
Internal inconsistency	Human error deriving from the speech recognition process, related to inconsistencies in describing the side or location of an abnormality or finding. For this error type, the first occurrence was taken to be correct, and any subsequent contradictory occurrence was considered an error.	5 mm left ureteric calculus.	(Previous mention earlier in the report of calculus in left ureter) 5 mm right ureteric calculus.	All internal inconsistency errors are considered clinically significant.

Note.—Adapted, with permission, from reference 1.

Performance of the models was interpreted with the common machine learning metrics of precision (positive predictive value), recall (sensitivity), and F1 score. As the harmonic mean of precision and recall, the F1 score ranged 0 to 1, expressed as 0% to 100%, with higher values indicating better classification performance.

The benchmark used was manual detection of actual errors and their location by the primary researcher. A true positive means the model identified an error at the correct location of an actual error. A false negative means an actual error was missed by the LLM. The models' suggested corrections were not evaluated.

Results

Manual Error Detection

A total of 3233 radiology reports were manually reviewed for errors. The mean report length was 230.07 words. There were 1429 (44.2%) reports that had at least one error, with clinically significant error identified in 106 (3.2%) reports. On the subsample of 100 reports, interrater agreement was strong, with Cohen κ values of 0.79 between the resident and the first consultant, 0.81 between the resident and the second consultant, and 0.88 between the two consultants.

Report length, as measured by word count, showed a significant positive association with the presence of communication errors ($P < .001$). For each additional word in the report, the likelihood of an error increased by 0.1% (coefficient = 0.001; 95% CI: 0.000, 0.001). Radiology residents had significantly higher odds of errors compared with attending radiologists ($P = .003$, odds ratio [OR] = 2.2, 95% CI: 1.3, 3.7). Working the overnight shift was also associated with higher odds of errors ($P = .04$, OR = 1.9, 95% CI: 1.2, 3.4), whereas working evening and day shifts showed no significant association ($P = .46$, OR = 1.04, 95% CI: 1.0, 1.1 for evening shifts and $P = .35$, OR = 1.08, 95% CI: 1.0, 1.2 for day shifts). No significant association was identified between patient status and error presence (emergency department patient status, $P = .50$, OR = 0.02, 95% CI: -0.1, 0.1; outpatient status $P = .33$, OR = 0.03, 95% CI: -0.1, 0.0, with inpatient status used as reference).

Model Error Detection

Optimized prompts increased the models' F1 scores by 5%–15% on the subset of 100 reports assessed by three independent raters. For GPT-3.5-turbo, F1 score increased from 59.1% to 73% for clinically significant errors and 32.2% to 45% for not clinically significant errors. F1 score for GPT-4 increased from 86.9% to 91% for clinically significant errors and from 94.3% to 97% for not clinically significant errors. Further increases were achieved for text-davinci-003 (72% to 82% F1 score on clinically significant errors, 60% to 74.3% F1

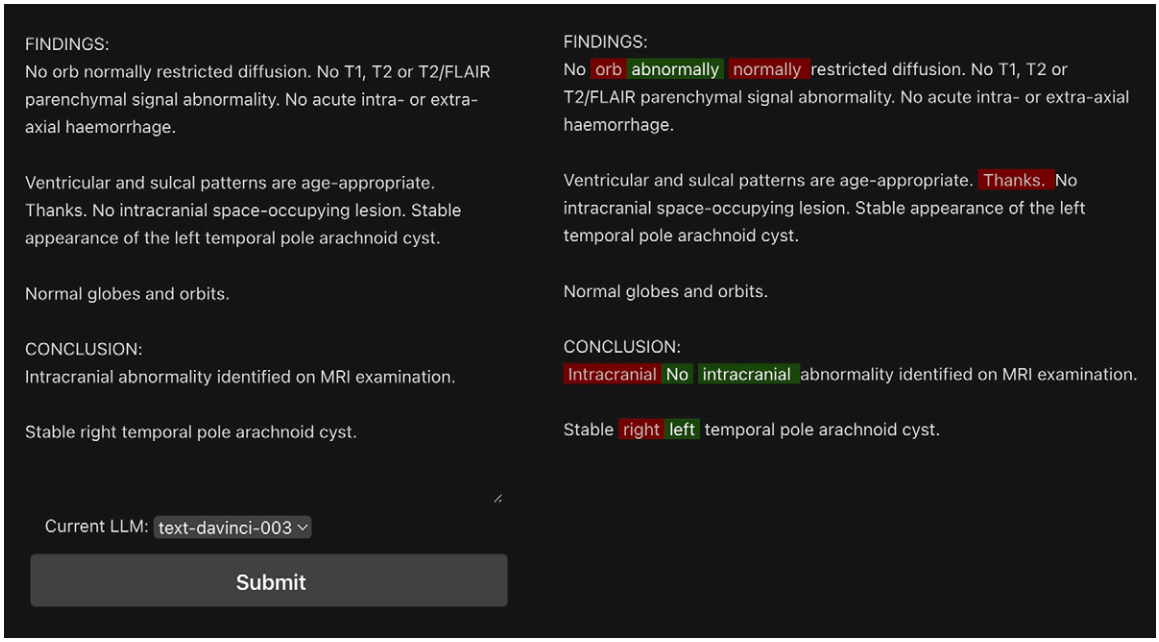

Figure 1: Web interface for review of large language model–detected errors on a sample generated radiology report. Example shown for text-davinci-003 model. In red are the sections that have been removed, in green the model's suggestions. Examples of word omission, nonsense phrases, and internal inconsistency (eg, "right" changed to "left") are included in the sample report for demonstration.

score on not clinically significant errors), Llama-v2–70B-chat (58.8% to 67% F1 score, 31.2% to 41%), and Bard (34.8% to 44% F1 score, 33.2% to 39%).

For clinically significant errors, the optimized GPT-4 achieved 76.9% precision, 100% recall, and 86.9% F1 score. GPT-3.5-turbo (65.1% precision, 54.1% recall, 59.1% F1 score) and text-davinci-003 (75.2% precision, 69.2% recall, 72% F1 score) performed less well. Llama-v2–70B-chat (62.5% precision, 87.3% recall, 72.8% F1 score) and Bard (34.1% precision, 44.1% recall, 38.5% F1 score) showed the lowest performance. Of 12 internal inconsistency errors in the corpus, GPT-4 detected 100% (text-davinci-003, 91.67%; GPT-3.5-turbo, 50.00%; Llama-v2–70B-chat, 33.33%; Bard, 33.33%). Of 24 nonsense errors, GPT-4 detected 91.6% (text-davinci-003, 83.3%; GPT-3.5-turbo, 79.1%; Llama-v2–70B-chat, 79.1%; Bard, 70.8%).

For not clinically significant errors, GPT-4 had 93.9% precision, 94.7% recall, and 94.3% F1 score. Text-davinci-003 (31.3% precision, 91.3% recall, 46.6% F1 score), GPT-3.5-turbo (21.5% precision, 64.1% recall, 32.2% F1 score), Llama-v2–70B-chat (32.3% precision, 91.3% recall, 47.7% F1 score), and Bard (11.9% precision, 84.7% recall, 20.9% F1 score) demonstrated lower accuracy. Sample model outputs are demonstrated in Figure 2, with examples of GPT-4's performance demonstrated in Figure 3.

Discussion

This study provides preliminary evidence that advanced generative LLMs can automatically detect speech recognition errors in radiology reports. GPT-4 demonstrated the highest accuracy in this comparison, achieving an F1 score of 86.9% for clinically significant errors in the dataset of 3233 radiology reports, and 94.3% F1 score for errors that were not clinically significant. Effectiveness was demonstrated in the detection of internal inconsistency and nonsense errors. The former requires assimilation of context across the entire radiology report, beyond the capacity of per-sentence analysis by deep learning models described in recent literature (8,9). The latter requires comprehension of what is considered appropriate for a radiology report, a subject still in contention regarding GPT-4's current abilities (21).

Our radiology center demonstrated error rates aligning with previous studies, with increased error rates associated with longer reports, overnight shifts, and resident dictations. Overnight shifts potentially induce fatigue-related oversights. Inexperience with speech recognition software likely drives higher resident error rates, compounded by confirmation bias of the attending radiologist checking the resident report. This confirmation bias represents an attending radiologist's "inattentional blindness" (22) to discrepancies when anchored to an existing report. These findings reveal areas for selective application of AI-assisted error detection tools to compensate for human limitations.

Despite anonymization efforts, data processed by third-party systems risk compliance violations. On-site private generative LLMs may mitigate this, possibly with locally hosted versions of open-source models such as Llama-v2–70B-chat. This model is a fine-tuned version of the Llama-v2–70B base model and has been designed for optimal chat-based interactions. Llama-v2–70B-chat performed poorly in comparison to the proprietary models evaluated. Further work is needed to assess the utility of fine-tuning the base Llama-v2–70B model, or other open-source models, on domain-specific data to improve performance.

Although theoretical capabilities exist for generative LLMs to synthesize novel radiology reports without human input (14), they are more likely to be decision support tools in the near future. Such tools will still require manual visual inspection of all reports by a radiologist prior to sign-off. In this context, it is

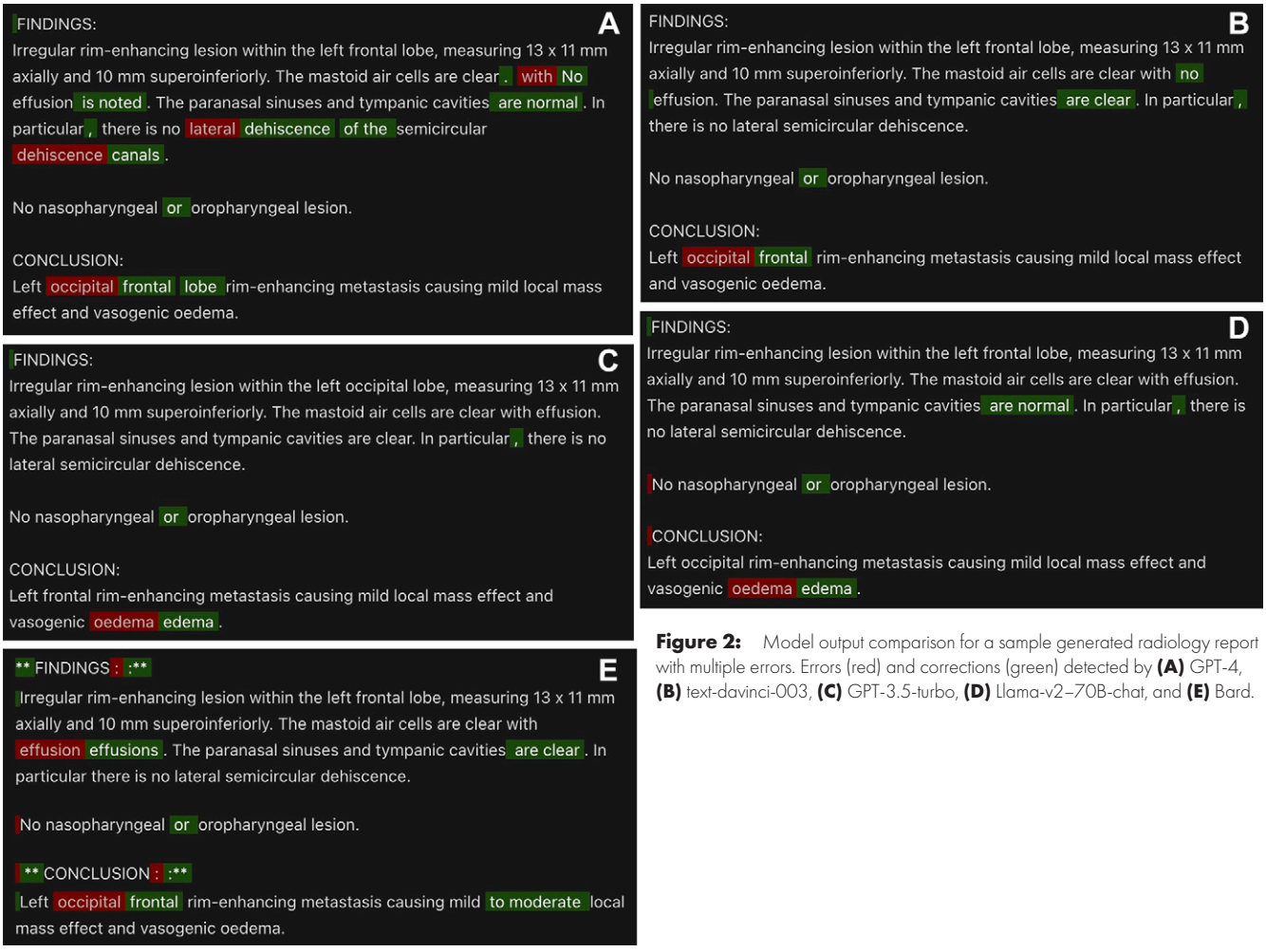

Figure 2: Model output comparison for a sample generated radiology report with multiple errors. Errors (red) and corrections (green) detected by **(A)** GPT-4, **(B)** text-davinci-003, **(C)** GPT-3.5-turbo, **(D)** Llama-v2–70B-chat, and **(E)** Bard.

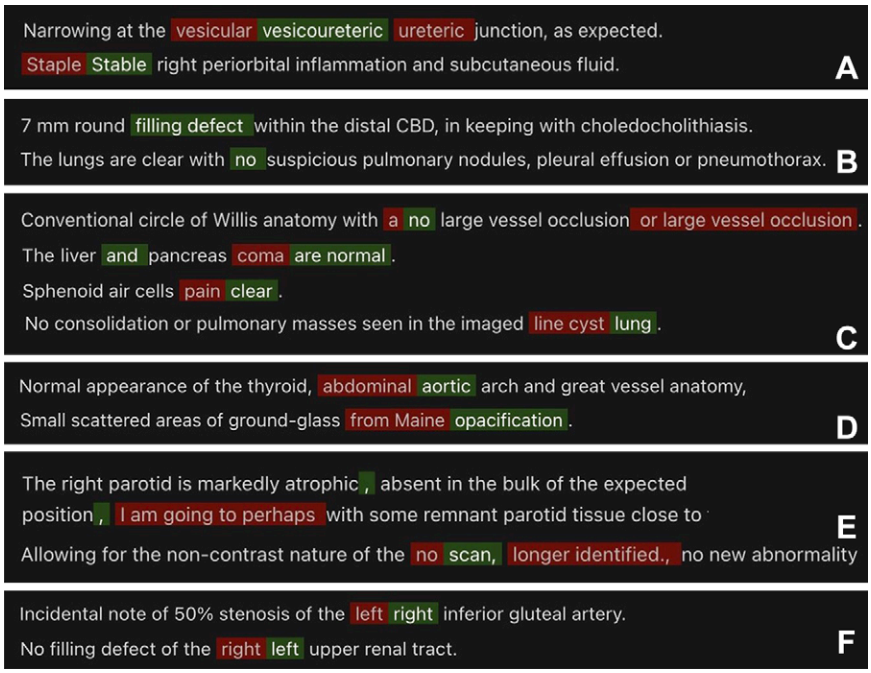

Figure 3: GPT-4 correctly detected multiple types of clinically significant errors (red indicates errors, green indicates correction): **(A)** homonym error, **(B)** omission error, **(C)** nonsense phrase, **(D)** translational error, **(E)** extraneous statements, and **(F)** internal inconsistency error.

worth acknowledging that error detection makes only one part of the creation of an accurate radiology report (23).

This study had several limitations. Data preparation and model evaluation were performed by one radiology resident at a single institution, with assessment performed on a small dataset, raising uncertainty regarding real-world viability. Additionally, while customized prompt optimization aims to maximize each model's capabilities, it introduces nonstandardization that risks biasing comparisons. The prompts were engineered and validated on a limited sample, incurring risk of overfitting. Finally, model capabilities are rapidly evolving, which makes comparisons quickly outdated.

In conclusion, this study demonstrates capabilities of advanced generative LLMs, particularly GPT-4, to automatically detect speech recognition errors in radiology reports. Further research is warranted to validate these findings in larger datasets across multiple institutions. If integrated into the radiology workflow, such models could potentially assist in improving report accuracy.

Author contributions: Guarantor of integrity of entire study, **R.A.S.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agree to ensure any questions related to the work are appropriately resolved, all authors; literature research, **R.A.S., W.L.**; clinical studies, **R.A.S., L.L.**; experimental studies, **R.A.S., L.L.**; statistical analysis, **R.A.S.**; and manuscript editing, **J.C.Y.S., K.C., L.L., W.L., J.Y.**

Disclosures of conflicts of interest: **R.A.S.** No relevant relationships. **J.C.Y.S.** No relevant relationships. **K.C.** No relevant relationships. **L.L.** No relevant relationships. **W.L.** No relevant relationships. **J.Y.** No relevant relationships.

References

1. Hawkins CM, Hall S, Zhang B, Towbin AJ. Creation and implementation of department-wide structured reports: an analysis of the impact on error rate in radiology reports. J Digit Imaging 2014;27(5):581–587.
2. Minn MJ, Zandieh AR, Filice RW. Improving radiology report quality by rapidly notifying radiologist of report errors. J Digit Imaging 2015;28(4):492–498.
3. Chang CA, Strahan R, Jolley D. Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. J Digit Imaging 2011;24(4):724–728.
4. Pezzullo JA, Tung GA, Rogg JM, Davis LM, Brody JM, Mayo-Smith WW. Voice recognition dictation: radiologist as transcriptionist. J Digit Imaging 2008;21(4):384–389.
5. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Stud Health Technol Inform 2015;216:922.
6. McGurk S, Brauer K, Macfarlane TV, Duncan KA. The effect of voice recognition software on comparative error rates in radiology reports. Br J Radiol 2008;81(970):767–770.
7. Ganeshan D, Duong PT, Probyn L, et al. Structured reporting in radiology. Acad Radiol 2018;25(1):66–73.
8. Zech J, Forde J, Titano JJ, Kaji D, Costa A, Oermann EK. Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. Ann Transl Med 2019;7(11):233.
9. Chaudhari GR, Liu T, Chen TL, et al. Application of a domain-specific BERT for detection of speech recognition errors in radiology reports. Radiol Artif Intell 2022;4(4):e210185.
10. Min D, Kim K, Lee JH, Kim Y, Park CM. RRED: a radiology report error detector based on deep learning framework. In: Proceedings of the 4th Clinical Natural Language Processing Workshop, Seattle, Wash. Association for Computational Linguistics, 2022; 41–52.
11. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023;307(4):e230725.
12. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023;20(10):990–997.
13. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. Radiology 2023;307(5):e230877.
14. Yu F, Endo M, Krishnan R, et al. Evaluating progress in automatic chest X-ray radiology report generation. Patterns (N Y) 2023;4(9):100802.
15. Chambon PJ, Wu C, Steinkamp JM, Adleberg J, Cook TS, Langlotz CP. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. J Am Med Inform Assoc 2023;30(2):318–328.
16. replicate/Llama-2-70b-chat. Replicate.com. https://replicate.com/replicate/Llama-2-70b-chat. Accessed August 7, 2023.
17. Ribeiro MT, Lundberg S. Adaptive testing and debugging of NLP models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland. Association for Computational Linguistics, 2022; 3253–3267.
18. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 2022; 24824–24837. https://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
19. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv 2005.14165 [preprint] https://arxiv.org/abs/2005.14165. Published May 28, 2020. Accessed September 13, 2023.
20. Berger E. Grounding LLMs. https://everything.intellectronica.net/p/grounding-llms. Published 2023. Accessed August 28, 2023.
21. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. Radiology 2023;307(5):e231259.
22. Nanapragasam A, Bhatnagar P, Birchall D. Trainee radiologist reports as a source of confirmation bias in radiology. Clin Radiol 2018;73(12):1052–1055.
23. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to create a great radiology report. RadioGraphics 2020;40(6):1658–1670.