REVIEW ARTICLE

Leveraging Large Language Models to Enhance Radiology Report Readability: A Systematic Review

Vasant Patwardhan, MD[a], Divya Balchander, MD[a], David Fussell, MD[a], John Joseph, PhD, MBA[b], Aditya Joshi[a], Hayden Troutt, MPH, CPH[a], Justin Ling, MD, MS[a], Katherine Wei, MD[a], Brent Weinberg, MD, PhD[c], Daniel Chow, MD, MBA[d]

Abstract

Background: Patients increasingly have direct access to their medical record. Radiology reports are complex and difficult for patients to understand and contextualize. One solution is to use large language models (LLMs) to translate reports into patient-accessible language.

Objective: This review summarizes the existing literature on using LLMs for the simplification of patient radiology reports. We also propose guidelines for best practices in future studies.

Evidence acquisition: A systematic review was performed following Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. Studies published and indexed using PubMed, Scopus, and Google Scholar up to February 2025 were included. Inclusion criteria comprised studies that used LLMs for simplification of diagnostic or interventional radiology reports for patients and evaluated readability. Exclusion criteria included non-English publications, abstracts, conference presentations, review articles, retracted articles, and studies that did not focus on report simplification. The Mixed Methods Appraisal tool 2018 was used for bias assessment. Given the diversity of results, studies were categorized based on reporting methods, and qualitative and quantitative findings were presented to summarize key insights.

Evidence synthesis: A total of 2,126 citations were identified and 17 were included in the qualitative analysis. Of these studies, 71% used a single LLM, and 29% of studies used multiple LLMs. The most prevalent LLMs included ChatGPT, Google Bard/Gemini, Bing Chat, Claude, and Microsoft Copilot. All studies that assessed quantitative readability metrics (n = 12) reported improvements. Assessment of simplified reports via qualitative methods demonstrated varied results with physician versus nonphysician raters.

Conclusion and clinical impact: LLMs demonstrate the potential to enhance the accessibility of radiology reports for patients, but the literature is limited by heterogeneity of inputs, models, and evaluation metrics across existing studies. We propose a set of best practice guidelines to standardize future LLM research.

J Am Coll Radiol 2025;■:■-■.

INTRODUCTION

Large language models (LLMs) have been incorporated into the workflow of many professions. One application of LLMs, termed text simplification, seeks to make complex information easier to understand, typically for a lay audience. Text simplification has been used in many industries, from law [1] to the biomedical sciences [2], and has even been used to simplify and translate texts originally written in foreign languages [3]. In health care, LLMs are being leveraged to enhance diagnostics, medical

[a]Department of Radiology, University of California, Orange, California.
[b]Paul Merage School of Business, University of California, Irvine, Irvine, California.
[c]Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA; Division Director, Neuroradiology, Department of Radiology and Imaging Sciences Emory University School of Medicine.
[d]Department of Radiology, University of California, Irvine, Orange, California.; Director, Advanced Analytics and Artificial Intelligence (A3), Institute for Precision Health, Susan & Henry Samueli College of Health Sciences; Codirector, Center for Artificial Intelligence in Diagnostic Medicine, Radiological Sciences School of Medicine.
Corresponding author and reprints: David Fussell, MD, University of California Irvine School of Medicine, Department of Radiological Sciences, Building 1, 101 The City Dr S Orange, CA 92868; e-mail: fusselld@hs.uci.edu.

The authors state that they have no conflict of interest related to the material discussed in this article. All authors are non-partner/non-partnership track/employees.

1546-1440/25/$36.00 ■ https://doi.org/10.1016/j.jacr.2025.09.004

education, scientific writing, and doctor-patient communication [4].

Simplifying complex medical information is especially relevant to radiology, particularly as health care increasingly follows a patient-centric care delivery model. With the advent of the Cures Act in 2021, patients typically have direct and near-immediate access to their radiology reports [5]; however, these reports are written for medical professionals and are often long, technical, and difficult for patients to understand. Patient misinterpretation may lead to confusion, stress, and unnecessary anxiety, straining the patient-physician relationship [6]. Although simplifying radiology reports would produce more accessible language, it is unrealistic for radiologists to create a layperson summary themselves, given increased imaging demand and radiologist workload [7]. Leveraging LLMs to simplify radiology reports has the potential to make these reports more accessible to patients without overburdening radiologists. Enhancing patient literacy may empower them as consumers and lead to improved patient satisfaction.

Although the potential of LLMs to simplify radiology reports for patient consumption is evident, a nuanced understanding of the challenges involved in adapting them for this purpose is required. To that end, this review article summarizes the current literature on using LLMs for radiology report simplification. Further, we use our findings to construct a blueprint for best practices to be adopted by future studies in this area.

EVIDENCE ACQUISITION

A systematic review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. The protocol was not registered before initiation. No external funding was obtained.

Eligibility and Search Parameters

A comprehensive search was conducted across the PubMed, Scopus, and Google Scholar databases up to February 2025. Keywords used for the search included "large language models," "radiology reports," and "simplify," which were modified into MeSH (Medical Subject Headings) terms according to each database. The full search strategy can be found in the e-only Supplemental Materials. A manual search using given parameters was also conducted on Google Scholar (n = 2).

Two independent reviewers (V.P. and D.B.) screened article titles and abstracts for potential inclusion. Full text was then reviewed by independent reviewers (V.P. and D.B.) for inclusion criteria. In the event of disagreement, additional reviewers (A.J. and D.F.) were consulted to reach consensus for inclusion. Inclusion criteria were as follows: (1) The study used one or more LLMs for simplification of reports for patients; (2) the study included evaluation of readability; and (3) the study considered diagnostic or interventional radiology reports. All levels of evidence (I-IV) and study designs were deemed eligible. Non-English publications, abstracts, conference presentations, review articles, retracted articles, and studies that did not focus on report simplification were excluded. A full-text review was conducted to evaluate inclusion criteria.

Outcome Measures

Information about the authors, LLM(s) used, radiographic modality, dataset size, readability and accuracy metrics, and limitations was extracted to a standardized form by two independent authors (V.P. and D.B.). For each study, the number and version(s) of LLMs used, text used for prompting, and output metrics were identified. Output metrics were characterized as either quantitative (eg, Flesch-Kincaid reading level [FKRL]) or qualitative. Extractions were reviewed by a third author (A.J.) to ensure consistency.

The Mixed Methods Appraisal tool 2018 was used for quality assessment given the varied study designs [8]. Assessment was performed by two independent authors (D.B. and A.J.) to determine if listed criteria were met, not met, or incompletely evaluated to create transparency in bias analysis. A comprehensive presentation of each criterion was created to better communicate the quality of included studies. Selection and reporting bias were reduced by adhering to the delineated search strategy and screening protocol. Certainty bias was minimized by including all study designs and levels of evidence.

A formal meta-analysis was not possible because of the heterogeneity of included studies. Instead, all included studies were reviewed for key information, including type of LLM, modality, sample size, AI prompts, and readability metrics. Incomplete summary statistics were listed as such in the final data table to produce transparency in data reporting. No data conversions were performed. Finally, a qualitative descriptive analysis of the key study characteristics in line with systematic review objectives was performed to summarize the results. No heterogeneity or sensitivity analyses were performed.

EVIDENCE SYNTHESIS

A total of 2,126 articles were identified from the three databases and 2 articles were identified via manual search. Of those, 2,072 articles were removed from screening, excluded, and not retrieved. Fifty-six full text articles were retrieved and assessed for eligibility, of which 30 were excluded due to focus on radiology report interpretation or

summarization instead of simplification and 9 were excluded due to lack of data on readability of simplified patient reports. Further information can be found in Figure 1.

Of the 17 studies included in our analysis (Table 1), 8 (47%) used fewer than 100 reports [9-16], 9 (53%) used between 100 and 1,000 reports [17-25], and 0 used over 1,000 reports. Eight (47%) used reports from one imaging modality [11,13,15, 20-24], and 8 (47%) used reports from multiple modalities [9,10,12,14,17-19,25.] These modalities included x-ray, CT, MR, ultrasound, interventional radiology, and mammography. One study did not specify the modality of imaging used [16]. Twelve (71%) studies used a single LLM [9,11,12,14-16,18-21,23,24], and 5 (29%) used multiple LLMs [10,13,17,22,25]. One study analyzed 6 LLMs [22]. The most prevalent LLMs include ChatGPT (OpenAI, San Francisco, CA, USA; n = 15) [9-13,15-20,22-25], Google Bard/Gemini (Google DeepMind, London, UK; n = 5) [10,13,14,17,25], Bing Chat (Microsoft Corporation, Redmond, WA, USA; n = 2) [17,25], Claude (Anthropic, San Francisco, CA, USA; n = 3) [13,21,22], and Microsoft Copilot (Microsoft Corporation, Redmond, WA, USA; n = 1) [10]. Five (29%) evaluated readability qualitatively with only Likert scales [9,12,14,15,23], 8 (47%) evaluated readability quantitatively with only readability metrics [10,11,17-19,21,22,25], and 4 (24%) evaluated readability qualitatively and quantitatively [13,16,20,24]. Fourteen (82%) studies investigated accuracy [9-13,15-18,20-24]. In all studies, accuracy was evaluated qualitatively using Likert scales. Additional information is found in Figure 2 and e-only Supplemental Materials.

Specific inputs used to prompt LLMs were quite variable, but some strategies were commonly employed. For example, nearly all studies included specific simplification commands such as "simplify," "explain," "translate," and "transform."

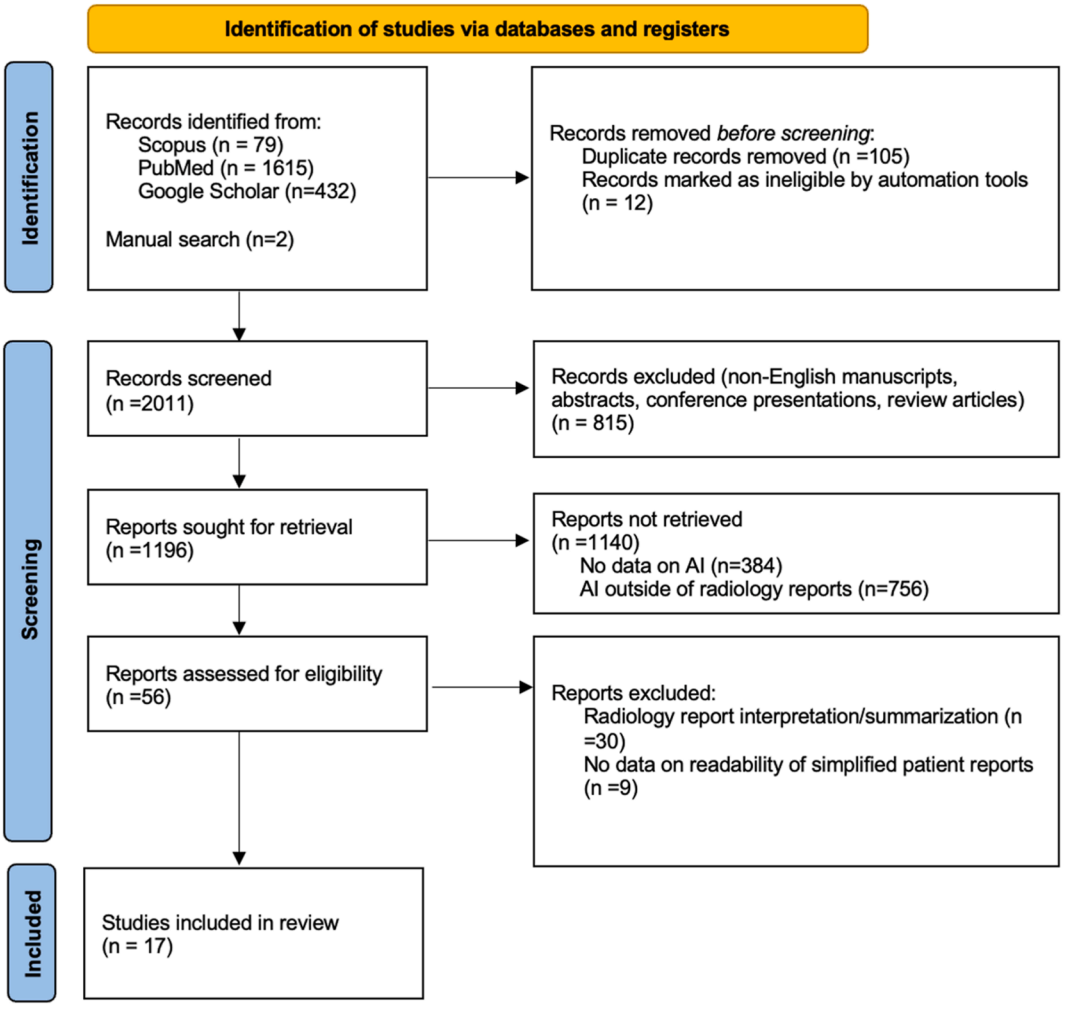

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram. AI = artificial intelligence.

Table 1. Study Characteristics

Author	AI LLM	Modalities Assessed	Report Sample Size (n)
Sarangi et al [9]	ChatGPT 3.5	CT, MRI	9
Tepe and Emekli [10]	ChatGPT-4, BARD, Microsoft Copilot	CT, MRI	30
Chung et al [11]	ChatGPT-3.5	MRI	5
Maroncelli et al [12]	ChatGPT-4o	Mammography, ultrasound, MRI	21
Gupta et al [13]	GPT-4, Gemini, Claude Opus, Llama-3.1-8B, Phi-3.5-mini	CT	50
Berigan et al [14]	Google Bard	Radiograph, ultrasound, CT, MRI	99
Schmidt et al [15]	Chat-GPT	MRI	3
Yang et al [16]	ChatGPT-3.5	Not specified	40
Amin et al [17]	GPT-3.5, GPT-4, Google Bard, and Bing	Radiograph, ultrasound, mammography, CT, MRI	150
Butler et al [18]	AI-LLLM GPT 3.5 (institutional LLM)	Radiograph, CT, MRI	300
Li et al [19]	ChatGPT	Radiograph, ultrasound, CT, MRI	400
Li et al [20]	GPT-4	Interventional radiology	200
Tang et al [21]	Claude v1.3	CT	100
Can et al [22]	GPT 3.5, GPT 4, Claude, Gemini ultra, Mistral-7b, Mistral-8x7b	Interventional radiology	109
Park et al [23]	GPT3.5 Turbo	MRI	685
Tripathi et al [24]	GPT4	Radiograph	500
Doshi et al [25]	GPT-3.5, GPT-4, Google Bard, Microsoft Bing	Radiograph, ultrasound, CT, MRI, mammography	254

AI = artificial intelligence; LLM = large language model.

Fig. 2. (A) Histogram depicting the number of imaging reports analyzed. (B) Bar chart depicting the frequency of each large language model (LLM) investigated. (C) Bar chart depicting how many studies used 1 or >1 LLMs. (D) Bar chart depicting how many studies assessed readability qualitatively versus quantitatively.

Prompt directions to stress readability of the interpretation included "layman's terms," "plain language," and "seventh-grade level." Several studies provided additional modifiers, directing the LLM to use a reassuring and empathetic tone, define medical jargon, provide analogies without omitting details, or suggest actionable recommendations.

Twelve studies total evaluated readability with quantitative metrics (Fig. 3) [10,11,13,16-22,24,25]. The most common readability metrics included the FKRL (n = 12) [10,11,13,16-22,24,25], Flesch Reading Ease score (n = 6) [10,18-20,22,24], automated readability index (n = 6) [13,16,17,21,24,25], Gunning-Fog Index (n = 6) [13,16,17,21,24,25], Coleman-Liau Index (n = 4)1 [3,17,21,25], and Simplified Measure of Gobbledygook readability index (n = 2) [21,22]. Mean length of report and conciseness were also assessed (n = 3) as reflections of readability [19,20,23]. Nine studies total measured readability qualitatively via Likert scale [9,12-16,20,23,24]. Accuracy was assessed by radiologists and other medical professionals, mostly using Likert scales and other institution-specific rating scales or questionnaires (n = 14). A few studies included qualitative analysis from the patient's perspective [12,14-16,20,24]. Several studies provided hallucination rates of simplified reports [16,18,23].

Quality assessment was performed using Mixed Methods Appraisal tool 2018 tool (e-only Supplemental figures). Of the three randomized controlled trials, all were limited by lack of double blinding [13,14,24]. Of the nonrandomized quantitative and descriptive studies, seven studies were limited by their use of fictional reports or randomly chosen radiology reports or report designs that may not represent the target population accurately [9,10,11,15,16,20,22].

READABILITY AND ACCURACY

Current literature finds that LLM translations are more readable and accessible than original radiology reports, while acknowledging that significant challenges remain for text simplification. All studies that assessed quantitative readability metrics (n = 12) reported improvements in readability for the LLM translations. Original reports were found to be difficult to read, with studies reporting readability indices corresponding to high school and college graduate reading levels [13,16,18,21,25]. Translated reports generally demonstrated readability at a middle school level. The lowest reported FKRL score, corresponding to reader grade level, was 5 [11]. The highest Flesch reading ease score, a measure of readability, was 74.3, corresponding to reader grade level 7 [18].

In three of five studies comparing multiple LLMs, versions of ChatGPT outperformed other models including Bard, Bing, Claude, and Gemini [17,22,25]. In the outliers, translations generated by Bard had the best Flesch reading ease score and FKRL scores [10]. and those generated by Claude had highest average improvement in reading scores [13]. Amin et al reported that their physicians expressed greater comfort providing translations generated by ChatGPT compared with those generated by Bing or Bard [17]. Word counts of translated reports were variable, with one study finding that word count decreased [20] and another finding that it increased [23]. Of note, some studies specified a target grade level in their prompts [11,24,25].

In addition to quantitative readability metrics, several studies assessed report translations as assessed by physicians and nonphysicians using Likert scales. Out of the five studies with physician subjective evaluation, two studies reported improvement in simplicity or comprehension via subjective assessment [9,16], and two studies found no significant difference in understandability [12,15]. One study with physician assessment rated the simplified report with the original report as a standard, with average score found to be 4.84 out of 5 [23]. Nonphysician

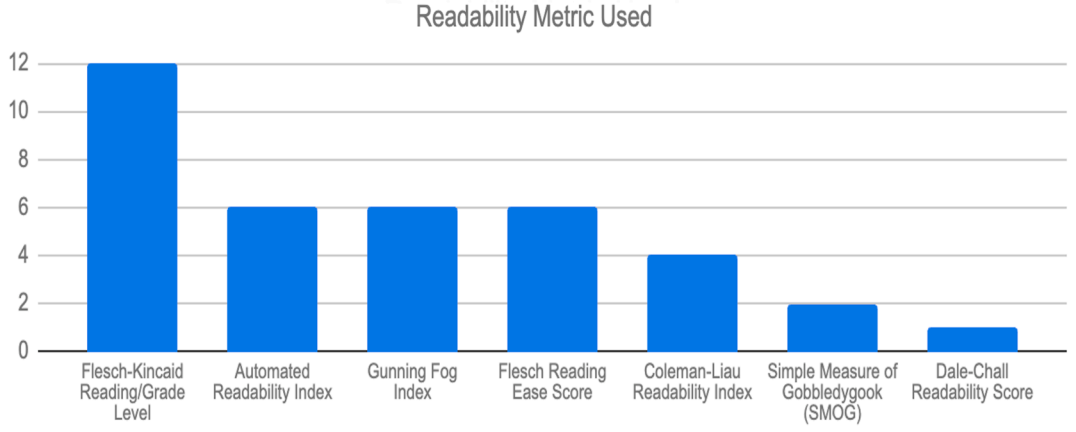

Fig. 3. Bar chart depicting which readability metric(s) studies used.

assessment of simplified reports was conducted by laypersons, patients, and medical and nonmedical volunteers. Nonphysicians reported improved readability in the form of simplicity of word choice and sentence structure, understandability, or comprehension [12-16,20,23,24]. Schmidt et al found that a majority of patients answered "neutral" when asked if simplified reports were as informative as a health care professional [15].

Many studies also assessed the accuracy of simplified reports (n = 14). Many asked radiologists to assess accuracy [9,12,13,15-17,20-23], a few studies used nonradiologist physicians and medical professionals to assess accuracy [11,18,24], and one study did not specify [10]. Assessments were largely via Likert scale or institution-specific questionnaire. Overall, qualitative analyses found translated reports to be accurate. One study found that accuracy improved with medical knowledge prompting [21]. Additionally, ChatGPT [17,22] and Claude [22] had higher ratings in accuracy compared with Bard, Bing, and other models. Among the ChatGPT models tested, one study reported that GPT-3.5 demonstrated improved accuracy [17], and another identified GPT-4 as achieving the highest accuracy [22].

Quantitative measurement of simplification accuracy is difficult given the lack of established, objective tools. Several studies identified in our literature review that did not meet inclusion criteria described the use of intermediate layperson prompting to mimic doctor-patient communication, showing improvement in various precision, recall, semantic similarity, and clinical accuracy metrics among different open- and closed-source LLMs [26,27].

The interaction between readability and accuracy was assessed in one study. The investigators found that readability and accuracy were inversely related, and that translating reports at an eighth-grade level offered the best combination of readability and accuracy [21].

Challenges and Limitations

Our literature review identified several universal challenges when applying artificial intelligence (AI) to generate patient-facing radiology reports. One of the primary issues is the problem of hallucinations (credible yet situationally incorrect statements). In studies that reported hallucination rates (n = 3), numbers ranged from 1.12% [23] to 6% [18]. Errors noted ranged from minor, such as identifying a fracture to be acute versus subacute, to major, such as reporting a distal radius fracture when none was identified [18]. Park et al noted that harmful translations were most often a result of LLMs adding extraneous information not originally present in the radiology report [23]. Other errors include harmful mistranslations [23], which also introduce incorrect medical information, and bias introduction [12] due to limited training datasets. It is clear that hallucinations have the potential to cause psychological harm and undermine trust in both the patient-provider relationship and health care systems. These challenges highlight the intricacies involved in using LLMs for radiology report simplification.

Several limitations are noted in the included studies. Many had small sample sizes with fewer than 100 reports [9-16] or considered only a single modality [11,13,15, 20-24]. Some studies used artificially constructed reports [15,22] or sentences from radiology reports [16]. Many used a single LLM, limiting generalizability of findings [9,11,12,14-16, 18-21,23,24].

A common limitation of these studies is the failure to include patient evaluation of translated reports. Though many demonstrate the feasibility of LLMs to generate such reports, it is important to investigate how these reports are perceived by the intended audience. One study that did investigate patient understanding found that patients generally understood the simplified reports [15]. Of studies that included layperson evaluation, participants were employed and college educated, limiting the generalizability of results to all patient populations [15,16].

DISCUSSION

As the use of LLMs in medicine increases, it is essential to find ways to leverage this technology for radiologists, ordering physicians, and patients. These studies show that LLMs are effective in making radiology reports more readable. However, a unified assessment of the current state of the literature is not without difficulty.

The heterogeneity of inputs, LLMs, and evaluation metrics makes comparisons across different studies challenging. The characteristics of radiology reports used for translation varied considerably, with wide ranges in the number and modality of reports. Most studies used a version of ChatGPT as their LLM; however, many other LLMs were also used. Readability and accuracy were generally evaluated with standardized metrics and Likert scales. The variability in these factors made quantitative evaluation difficult; there is a need for more uniform reporting in future studies.

Beyond heterogeneity of study design, a significant challenge in analyzing LLM studies stems from the rapid rate of change in the field, in which new models are constantly being released. For example, studies of earlier versions of LLMs, such as Chat-GPT 3.5, generally found higher rates of inaccuracy than studies carried out using newer versions. Given the rapid evolution of LLM

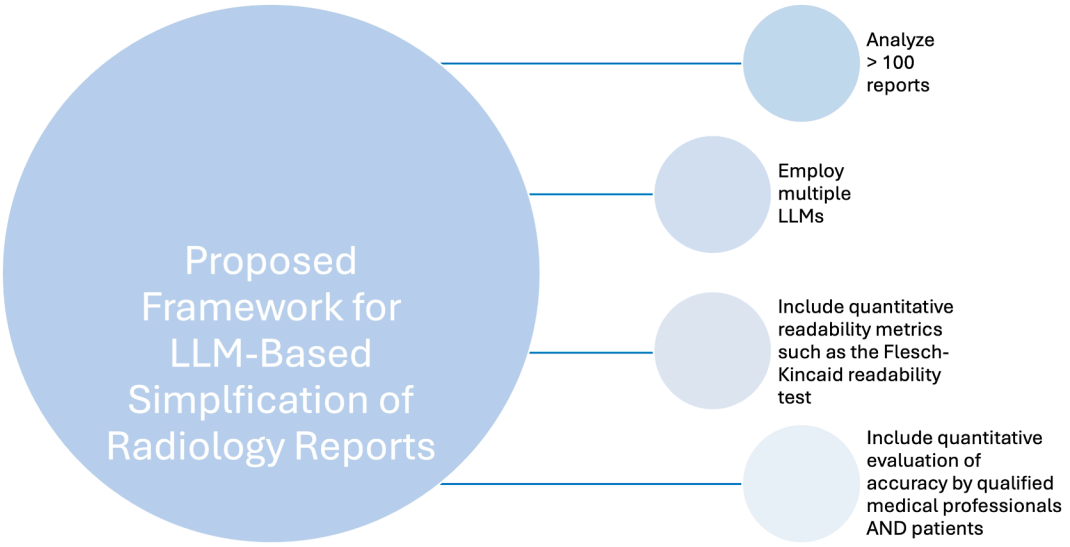

Fig. 4. Proposed framework for evaluating artificial intelligence (AI) utility in simplification of patient radiology reports. Proposed framework for evaluating AI utility in simplification of patient radiology reports. LLM = large language model.

technology, significant lag between the release of models and their subsequent evaluation in literature is inevitable. However, this can be mitigated by having a standardized framework for study design, which would be expected to simplify research execution and speed reporting.

Given the heterogeneity of results, we recommend a framework for evaluating AI utility in simplification of patient reports, grounded in the methodologies most frequently observed in the existing literature (Fig. 4). Regarding input, we recommend that over 100 real clinical reports be included to improve quality while maintaining feasibility. Studies should also employ multiple LLMs to provide insight into which models may be most appropriate for simplifying radiology reports. Quantitative measures such as FKRL tests should be used when evaluating studies for readability. Qualitative readability measures should include perspectives from both physicians and patients of varying education levels. Finally, quantitative evaluation of accuracy by qualified medical professionals can help readers assess functionality and track the rate of errors. Included studies with number of fulfilled criteria are depicted in Figure 5.

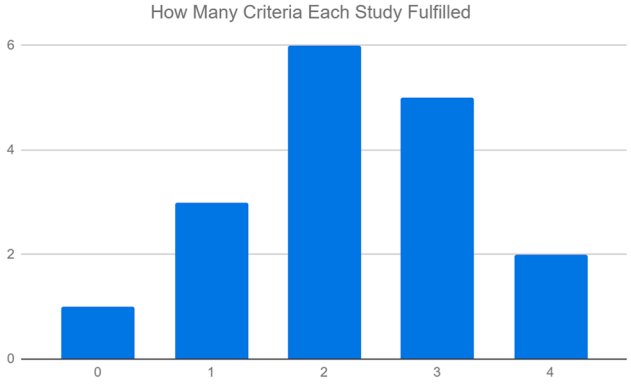

Fig. 5. Bar chart depicting number of proposed framework criteria met by each study.

Before considering standard implementation in practice, further research is essential to clarify the capabilities and limitations of LLMs in radiology report simplification. Important questions remain, including whether simplified reports are actually desired by referring providers, how such reports might alter the physician-patient relationship, and whether simplified interpretations meaningfully improve patient autonomy compared to directly accessing the full report. Furthermore, the relatively high frequency of hallucinations, which have the potential to do real psychological harm by transmitting misinformation, raises significant concerns about the appropriateness of using LLMs for this clinical purpose. Given these outstanding questions, our proposed framework is intended to offer a uniform, streamlined approach for further study of the role of AI in improving the comprehensibility of reports for patients. Going forward, holistic investigation will be essential to ensure that LLMs meaningfully improve the readability of radiology reports and effectively serve the needs of patients and providers.

TAKE-HOME POINTS

- LLMs demonstrate the potential to improve patient care and the patient experience by making radiology reports more accessible.

- However, the current literature is limited by significant heterogeneity in study design.
- In this review, we summarize the current state of the literature and propose a set of guidelines to standardize future research in this area.

REFERENCES

1. Garimella A, Sancheti A, Aggarwal V, Ganesh A, Chhaya N, Kambhatla N. Text simplification for legal domain: {I}nsights and challenges. In: Proceedings of the Natural Legal Language Processing Workshop 2022. Abu Dhabi, UAE: Association for Computational Linguistics; 2022:296-304.
2. Li Z, Belkadi S, Micheletti N, Han L, Shardlow M, Nenadic G. Large language models for biomedical text simplification: promising but not there yet. Published online September 24, 2024. https://doi.org/10.48550/arXiv.2408.03871.
3. Klöser L, Beele M, Schagen JN, Kraft B. German text simplification: finetuning large language models with semi-synthetic data. Published online February 16, 2024. https://doi.org/10.48550/arXiv.2402.10675.
4. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Published online July 8, 2024. https://doi.org/10.48550/arXiv.2401.06775.
5. US Department of Health and Human Services. 21st Century Cures Act: interoperability, information blocking, and the ONC Health IT Certification Program. Available at: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-D/part-171. Published 2020. Accessed March 3, 2025.
6. Nickel B, Barratt A, Copp T, Moynihan R, McCaffery K. Words do matter: a systematic review on how different terminology for the same condition influences management preferences. BMJ Open 2017;7:e014129.
7. McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015;22:1191-8.
8. Hong QN, Pluye P, Fàbregues S, et al. Mixed methods appraisal tool (MMAT), version 2018. J Educ Informat 2018:1-7. 1148552.
9. Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. Cureus 2023;15:e50881.
10. Tepe M, Emekli E. Decoding medical jargon: the use of AI language models (ChatGPT-4, BARD, Microsoft Copilot) in radiology reports. Patient Educ Couns 2024;126:108307.
11. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digit Health 2023;9:20552076231221620.
12. Maroncelli R, Rizzo V, Pasculli M, et al. Probing clarity: AI-generated simplified breast imaging reports for enhanced patient comprehension powered by ChatGPT-4o. Eur Radiol Exp 2024;8:124.
13. Gupta A, Singh S, Malhotra H, et al. Provision of radiology reports simplified with large language models to patients with cancer: impact on patient satisfaction. JCO Clin Cancer Inform 2025;9:e2400166.
14. Berigan K, Short R, Reisman D, et al. The impact of large language model-generated radiology report summaries on patient comprehension: a randomized controlled trial. J Am Coll Radiol 2024;21:1898-903.
15. Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg 2024;144:611-8.
16. Yang Z, Cherian S, Vucetic S. Two-pronged human evaluation of ChatGPT self-correction in radiology report simplification. Published online June 27, 2024. https://doi.org/10.48550/arXiv.2406.18859.
17. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023;309:e232561.
18. Butler JJ, Acosta E, Kuna MC, et al. Decoding radiology reports: artificial intelligence-large language models can improve the readability of hand and wrist orthopedic radiology reports. Hand (N Y). Published online August 13, 2024:15589447241267766. https://doi.org/10.1177/15589447241267766.
19. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging 2023;101:137-41.
20. Li HH, Moon JT, Kumar S, et al. Evaluation of multi-lingual simplifications of IR procedural reports using GPT-4. J Vasc Interv Radiol Published online January 8, 2025;S1051-0443(25):00008-9. https://doi.org/10.1016/j.jvir.2025.01.002.
21. Tang CC, Nagesh S, Fussell DA, et al. Generating colloquial radiology reports with large language models. J Am Med Inform Assoc 2024;31:2660-7.
22. Can E, Uller W, Vogt K, et al. Large language models for simplified interventional radiology reports: a comparative analysis. Acad Radiol. Published online September 30, 2024;S1076-6332(24):00690-1. https://doi.org/10.1016/j.acra.2024.09.041.
23. Park J, Oh K, Han K, Lee YH. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. Sci Rep 2024;14:13218.
24. Tripathi S, Mutter L, Muppuri M, et al. PRECISE Framework: GPT-based text for improved readability, reliability, and understandability of radiology reports for patient-centered care. Published online February 20, 2024. https://doi.org/10.48550/arXiv.2403.00788.
25. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman HP. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. Published online June 5, 2023. https://doi.org/10.1101/2023.06.04.23290786.
26. Butler JJ, Harrington MC, Tong Y, et al. From jargon to clarity: improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. Foot Ankle Surg 2024;30:331-7.
27. Zhao X, Wang T, Rios A. Improving expert radiology report summarization by prompting large language models with a layperson summary. Published online June 20, 2024. https://doi.org/10.48550/arXiv.2406.14500.