

Research Article

A Large Language Model Framework for Patient-Centered Translation of Radiology Reports Through Semantic Simplification and Health Literacy Enhancement

Muhammad Ahsan¹, Dr. Ruqayya Ishaq¹, Dr. Memoona Jawed²

¹Lincoln County Hospital, Lincoln, United Kingdom, ²Sheffield Teaching Hospitals NHS Trust, Sheffield, United Kingdom.

Corresponding Author: Muhammad Ahsan muhammad.ahsan11@nhs.net

Received: 10 April, 2026; **Accepted:** 20 June 2026

Abstract

The language of radiology reports is often difficult to understand and can impede communication and/or informed clinical decisions. This research investigates the use of large language models (LLM) to create a patient-centred report from radiologist-generated chest X-ray reports without compromising clinical accuracy. The framework leverages the Indiana University OpenI Chest Xray dataset, which contains about 4,000 radiology reports, and tests several state-of-the-art LLMs, GPT-4, Llama 3, Mistral, and Gemma. Radiology reports are converted to plain language summaries with enhanced health literacy but without affecting the diagnostic significance. The metrics employed to evaluate the performance are semantic similarity scores (BERTScore, BLEU, ROUGE-L), readability scores (Flesch Reading Ease, Flesch-Kincaid Grade Level), and comparison analysis between different LLM models. The proposed framework offers a systematic evaluation to assess the potential of contemporary LLMs in creating clinically accurate, patientcentric, and comprehensible radiology reports, ensuring transparency in healthcare communication and elevating patient engagement.

Keywords: Large Language Models (LLMs); Radiology Report Simplification; Patient-Centered Communication; Health Literacy; Chest X-ray Reports.

1. Introduction

Radiology reports are an essential part of the healthcare landscape in the modern era and form the basis of the interpretation of medical imaging to inform diagnosis, treatment planning and clinical

decision making [1]. These reports are generally composed by a radiologist in medical terminology, which would be understood by medical professionals, rather than by patients. Patients are able to access their EHRs more and more via digital health portals and many find the reports from radiology difficult to understand as the result of complex clinical language, abbreviations, etc. This disconnect between individuals can cause confusion, anxiety, misinterpretation of results, and lower levels of patient involvement in health care decisions [2].

The recent advent of Large Language Models (LLMs) has shown outstanding performance in various natural language processing tasks such as natural language understanding, text generation, summarization, and semantic simplification. GPT-4, Llama 3, Mistral, and Gemma demonstrate significant promise across various domains of medical text-to-text translation, including translating intricate medical narratives into concise, coherent, and patient-friendly explanations while maintaining clinical context and meaning [3]. They can generate language based on context, which can be useful for improving health literacy by making reports generated by radiologists more accessible to patients who lack medical training. AI-driven communication can aid joint decisionmaking and enhance communication between the health care provider and patient [4].

Although these developments, most studies have been conducted on automated generation of radiology reports from medical images or summarization of clinical texts, but not patient-centred translation of radiology reports. Furthermore, there are limited comparative studies on the performance of several state-of-the-art LLMs in preserving semantic meaning, readability, and enhancing health literacy [5]. There are numerous approaches that are available but that do not fully evaluate the impact of the simplification on the diagnostic accuracy of the reports while at the same time enhancing readability and patient understanding. Therefore, there is a need for a systematic approach to assess different LLMs to determine their effectiveness in the translation of radiology reports into comprehensible, clinically accurate language.

In this study, we aim to tackle this challenge by leveraging the Indiana University OpenI Chest Xray dataset, which contains chest X-ray images and their corresponding radiology reports. To address this challenge, this study proposes a Large Language Model framework for patientcentered translation of chest X-ray radiology reports. This framework contrasts the performance of several LLMs such as GPT-4, Llama 3, Mistral and Gemma for creating patient-friendly reports from radiologist-authored findings and impressions. Semantic similarity metrics (BERTScore, BLEU, ROUGE-L) and readability measures (Flesch Reading Ease, Flesch–Kincaid Grade Level) are used for the evaluation. The suggested framework helps enhance health literacy and improve communication clarity, while also assisting in the integration of reliable AI tools into clinical practice.

The main contributions of this study are as follows:

- A report translation system for chest X-rays is introduced for patient-centered reports using multiple state-of-the-art Large Language Models (GPT-4, Llama 3, Mistral and Gemma) to

provide clear and easy-to-understand reports for patients while keeping the original clinical context intact.

- A thorough comparative analysis of the top models is performed for semantic similarity (BERTScore, BLEU, ROUGE-L) and readability (Flesch Reading Ease, Flesch–Kincaid Grade Level) to determine the most promising model for accurate and readable radiology report simplification.
- To develop a systematic approach for enhancing health literacy, we use the Indiana University OpenI Chest X-ray dataset to show that patient-centred healthcare communication is achieved through the provision of LLM-generated simplified reports that retain diagnostic information without sacrificing patient understanding.

2. Literature Review

The recent developments in artificial intelligence have had a substantial impact on radiology, especially in the interpretation of chest X-rays, report writing, semantic assessment, and patient communication. To enhance the quality, accuracy, interpretability and accessibility of radiology reports, large language models (LLMs) and vision-language models are increasingly being investigated. Unlike the traditional rule-based systems that can only identify pre-defined labels from radiology reports, LLM-based methods have the advantage of being more flexible and context-aware when interpreting clinical text and images. Lanfredi et al. [6] introduced MAPLEZ, a privacy-preserving LLM-based approach for improving chest X-ray datasets with multi-type annotations. They were able to not only extract binary abnormality labels from radiology reports, but location, severity, and uncertainty levels as well. The study also revealed the benefits of improved macro F1 score and location annotation performance, highlighting that LLMs have the potential to add value to radiology datasets beyond just label extraction.

The automatic generation of a chest X-ray report has also been the subject of much research. Mei et al. [7] proposed a fine-grained reinforcement learning framework with structured reasoning for Chest X-ray report generation. They model clinical accuracy and interpretability by employing their two-stage approach of supervised chain-of-thought reasoning and segment-level reinforcement learning on MIMIC-CXR and IU-Xray datasets. Likewise, Barzegar et al. [8] introduced a prompt-guided vision-language framework with shortcut bias mitigation for the pathology-guided task. They solved the issue of spurious correlations in the generation of radiology reports by improving on the image artifacts and text descriptions. The improvements of the ROUGE-L, METEOR, RadGraph, BLEU and clinical coherency metrics suggest that promptbased methods and bias-aware approaches can improve the reliability of radiology reports generated.

Comparative studies have also identified the strengths and weaknesses of GPTs in radiology. Lee et al. [9] have compared the performance of a specialised chest X-ray report generation model

(M4CXR) to ChatGPT-4o in the interpretation of chest X-rays. The results indicated that ChatGPT was less acceptable for diagnostic purposes than M4CXR, and M4CXR were more accurate and had fewer hallucinations in the anatomical localization and diagnostic interpretation. This suggests that LLMs trained on specific domains tend to outperform general ones in specific direct diagnostic interpretation tasks. To generate structured diagnostic reports, Rodrigues et al. [10] suggested an approach to generate reports, which integrates classification, segmentation, Grad-CAM visualization, and DeepSeek-R1 named as IHRAS. The study pointed out that ensuring the interpretability of AI models, anatomical segmentation, and use of standardized terminology are key elements in building trustworthy AI-powered radiology workflows.

Although the assessment of AI-generated radiology reports remains a great challenge, linguistic measures such as BLEU and ROUGE score may not necessarily capture the clinical accuracy. Ali et al. [11] introduced a semantic assessment framework for the generation of medical reports through LLMs and evaluated it on two medical reports datasets: Indiana University and MIMIC-CXR. Their findings revealed that Mistral with task-oriented prompting outperformed in semantic accuracy and was similar to BERTScore and CLIP-Score. This result has confirmed the utility of semantic similarity measures besides lexical overlap measures. In the same vein, Allaberdiev et al. [12] introduced the memory-augmented vision-language transformer (ChestXGen) for generating radiology reports. They showed an improvement in BLEU and METEOR scores on MIMIC-CXR with their model, highlighting the utility of memory-augmented architectures for generating coherent and clinically relevant reports.

The study of interactive and multimodal applications of LLMs in medical imaging is also a recent focus of research. To enhance diagnosis, segmentation and reports generation, Wang et al. [13] integrated LLMs with computer-aided diagnosis networks. Their framework showed that LLMs can convert technical CAD-generated information into natural language reports and enable patientfriendly interactions. Liu et al. [14] conducted a comprehensive analysis of GPT-4V on chest Xray image analysis for report generation, medical visual question answering and visual grounding. The study revealed that GPT-4V performed effectively in report generation and question answering but had some difficulties in visual grounding, highlighting the importance of considering multimodal medical AI systems carefully. Others have explored extending the use of LLM-based reporting to CT imaging as well, such as Tian and Song [15] who used low-rank matrix recovery and continuous slice encoding to apply LLM-based medical reporting to CT imaging, suggesting that LLM-based medical reporting is growing beyond chest radiography.

Table 1. Recent LLM-Based Radiology Studies

Study	Main Method	Dataset/Task	Key Contribution	Limitation / Gap
-------	-------------	--------------	------------------	------------------

Ju et al. [16]	DM-RRG with disease-topic guidance and multi-grained image-text alignment	MIMIC-CXR and IU X-ray report generation	Improved diagnostic fidelity using disease-aware visual encoding, Mamba-based alignment, and disease-prompted report generation.	Focused on clinician-oriented report generation, not patient-friendly simplification.
Yang et al. [17]	MedXChat unified vision-language model	MIMIC-CXR; report generation, VQA, and text-to-image synthesis	Integrated multiple radiology tasks in one LLM-based framework and achieved strong performance in reporting, VQA, and image synthesis.	Designed for diagnostic workflow and education, with limited focus on health literacy improvement.
Izhar et al. [18]	Hybrid multimodal model with ORPO alignment	Chest X-ray and lung CT report generation	Used preference optimization to reduce hallucinations and improve clinically reliable continual report generation.	Emphasized automated report generation rather than simplifying radiologist-written reports for patients.
Tripathi et al. [19]	PRECISE GPT-4 framework	500 chest X-ray reports	Generated sixth-grade patient-friendly summaries and significantly improved readability and understandability.	Used GPT-4 only; limited comparative analysis across multiple LLMs.

Table 1 shows some recent research related to report generation with LLM, vision-language models, and patient-centric report simplification. It points out proposed approaches, databases, contributions, and identified research gaps, which will stimulate the development of the proposed framework for semantic simplification and health literacy improvement of chest X-ray radiology reports.

Most of the current research is devoted to creating automation for radiology report generation, and some work to simplify radiologist-written reports into something that's more palatable to patients. Thus, it is necessary to develop a comparative LLM framework to assess the accuracy of the simplified chest X-ray report, readability, and enhancement of health literacy.

3. Propose Methodology

3.1 Research Design

This study adopts a quantitative experimental study aimed at assessing the effectiveness of Large Language Models (LLMs) in generating patient-friendly reports from radiologist-generated chest X-ray reports. To demonstrate the proposed framework, the Indiana University OpenI Chest X-ray dataset is used, and the original Finding and Impression sections are fed into the various LLMs (GPT-4, Llama 3, Mistral, and Gemma). The generated reports are then evaluated against the original reports with semantic similarity metrics (BERTScore, BLEU, and ROUGE-L) and readability metrics (Flesch Reading Ease and Flesch–Kincaid Grade Level). This comparative assessment helps to select the LLM which retains clinical meaning and enhances readability and patient-centered communication.

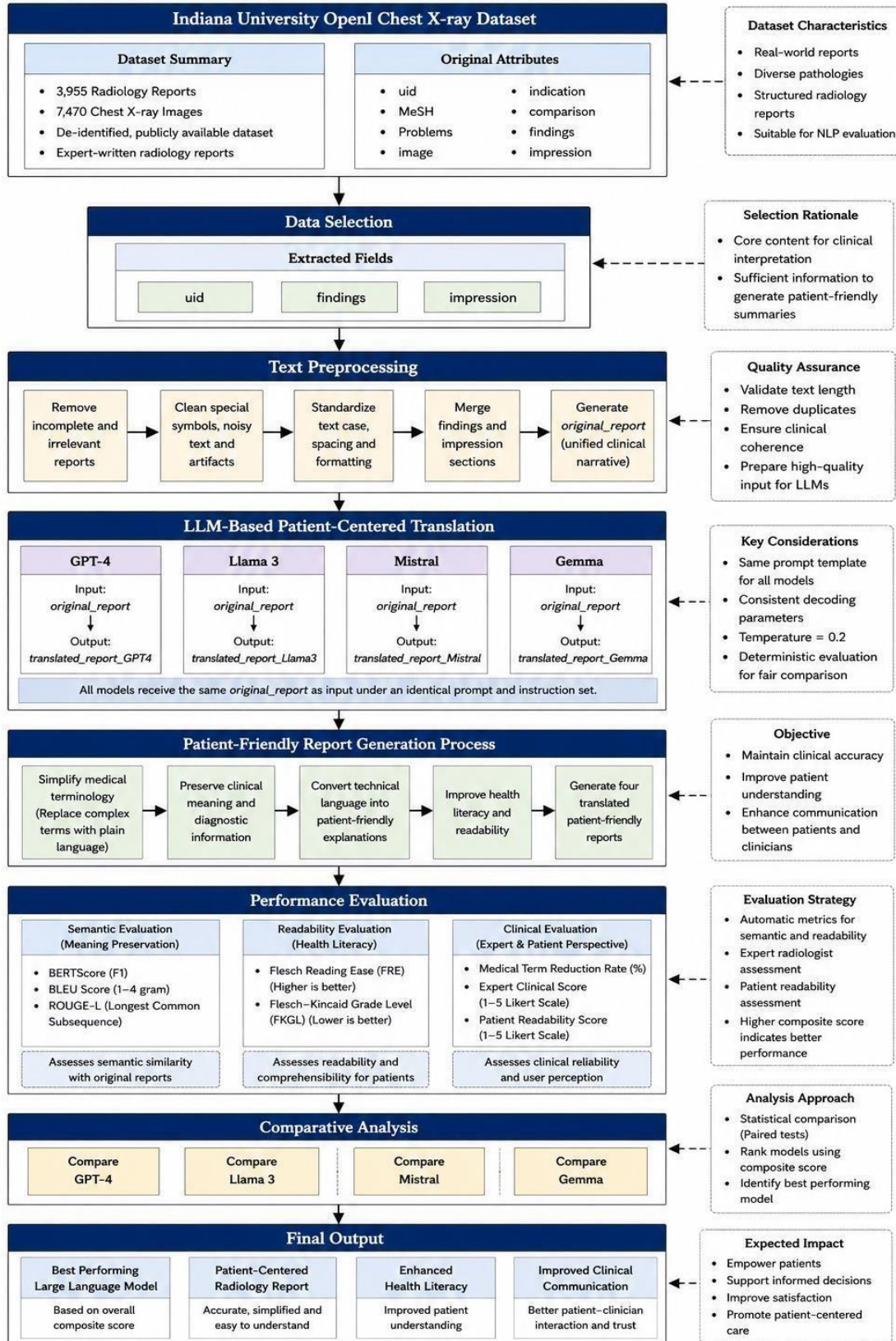


Figure 1. Proposed LLM Framework for Patient-Centered Radiology Report Translation The entire method of the proposed framework of translating radiology reports using LLM is shown in Figure 1. The process starts from the Indiana University OpenI Chest X-ray dataset and appropriate fields of text are extracted and preprocessed. The cleaned reports are then given to GPT-4, Llama 3, Mistral and Gemma for the patient friendly translation. Reported generated are assessed by using semantic, readability and clinical assessment metrics. Finally, comparative analysis determines the most effective LLM to enhance the communication between radiology and health literacy.

3.2 Dataset Description

The proposed framework uses the Indiana University OpenI Chest X-ray dataset, which is openly provided and consists of about 3,955 radiology reports, with each report associated with 7,470 chest X-ray images. This original data contains 8 attributes - uid, MeSH, Problems, image, indication, comparison, findings, and impression. This study aims to make the radiology report easier to understand for the patient, but does not attempt to analyse the images, so only text information from the report is used.

The original data is preprocessed to extract the uid, findings and impression fields to build the experimental dataset. The findings and impression sections are combined into one original_report column to provide input for the evaluated LLMs. Records that do not have a report are removed, and the remaining text is cleaned and standardized by removing noisy symbols and fixing formatting inconsistencies. The processed dataset is then comprised of the following features: uid, findings, impression, and original_report, which serve as a structured input to compare and evaluate multiple LLMs as well as to improve their simplicity and readability. Table 2. Original Dataset Attributes

Attribute	Description
uid	Unique identifier for each radiology report
MeSH	Medical Subject Headings disease annotations
Problems	Clinical problem labels extracted from the report
image	Chest X-ray image or view description
indication	Clinical reason for examination
comparison	Reference to previous imaging studies

findings	Detailed radiologist observations
impression	Final radiological conclusion or summary

Table 3. Processed Dataset Attributes

Attribute	Description
uid	Unique report identifier
findings	Original radiologist findings
impression	Original radiologist impression
original_report	Combined findings and impression used as LLM input
gpt4_translation	Patient-friendly output generated by GPT-4
llama3_translation	Patient-friendly output generated by Llama 3
mistral_translation	Patient-friendly output generated by Mistral
gemma_translation	Patient-friendly output generated by Gemma
bertscore	Semantic similarity score
bleu	Lexical similarity score
rouge_l	Content preservation score
flesch_reading_ease	Readability ease score
fkgi	Flesch–Kincaid Grade Level
medical_term_reduction	Percentage of simplified medical terms
expert_score	Radiologist evaluation score

patient_readability_score	Patient understandability score
----------------------------------	---------------------------------

Tables 2 and 3 show the summary of the transformation of the original Indiana University OpenI dataset to the processed dataset for this study. Tables 2 and 3 summarized the attributes of the original dataset and extracted/generated attributes used for LLM-based patient-centered translation, semantic evaluation, readability assessment, and expert/patient scoring, respectively.

3.3 LLM-Based Patient-Centered Translation

The original_report has been preprocessed and fed to four state-of-the-art Large Language Models (LLMs): GPT-4, Llama 3, Mistral, and Gemma. The original_report is preprocessed and passed to four renowned Large Language Models (LLMs): GPT-4, Llama 3, Mistral, and Gemma. They were chosen as they are representative of both proprietary and open source LLMs that have been shown to handle natural language understanding, text generation, reasoning and medical language processing. GPT-4 was selected as a robust commercial baseline because of its capabilities in generating high-quality language, following instructions, and contextual reasoning, rendering it suitable to translate detailed radiology reports into understandable patient explanations. Llama 3 was added due to its impressive open-source capabilities, reasonable inference speed, and capability to produce coherent long-form text while maintaining context. We chose Mistral because of its efficiency in computation, lightweight architecture, and performance in language understanding tasks, which are important for practical use in the healthcare domain, where computational resources are scarce. We made this change because of the very good instruction-following ability and optimal text generation of Gemma, a model developed by Google that was added as an extra open-weight model for comparison.

The same radiology report is sent to each model and it is asked to create a simplified report for patients who are not medical experts. The prompt calls for the model to maintain the clinical meaning of the original text, to provide explanations in simple language for complex medical terms, to reduce the length of medical text, and to create simple and understandable summaries without unnecessary medical jargon. The outputs generated by the models are saved in the following files: gpt4_translation, llama3_translation, mistral_translation, and gemma_translation, so that they can be fairly compared. The proposed framework enables an evaluation of multiple LLMs based on their performance in the same context to find the model that is most suited for balancing semantic fidelity, readability and health literacy enhancement, informing the selection of an appropriate LLM for patient-centered radiology communication.

3.4 Performance Evaluation

Patient friendly reports generated are evaluated for the performance using both semantic similarity and readability metrics. BERTScore and ROUGE-L measure the preservation of the clinical

meaning or content of the original radiology report, while BLEU measures lexical overlap. BERTScore and ROUGE-L are used to assess the clinical meaning or content preserved in the original radiology report, and BLEU is used to measure the lexical overlap. The Flesch Reading Ease (FRE) and Flesch–Kincaid Grade Level (FKGL) scores are calculated to determine the ease of readability of the patient report – higher FRE scores and lower FKGL values mean that patients will be able to read the report easily. The results of the evaluations for GPT-4, Llama 3, Mistral, and Gemma are then compared to select the most appropriate LLM for patient-centered radiology communication, considering semantic fidelity, readability, and health literacy enhancement.

4. Results and Discussion

This section discusses the experimental results of the proposed Large Language Model (LLM) framework for patient-centred translation of radiology reports from chest X-rays. Semantic similarity metrics and readability metrics are used to evaluate the performance of GPT-4, Llama 3, Mistral and Gemma, including BERTScore, BLEU, ROUGE-L, Flesch Reading Ease, and Flesch–Kincaid Grade Level. The results offer a comparative evaluation of the clinical meaning of the original radiology reports retained by the models and the patient-friendly explanations produced. Moreover, the results show that the most suitable LLM for providing patient-centered communication in clinical settings can ensure high semantic fidelity, readability and enhance health literacy.

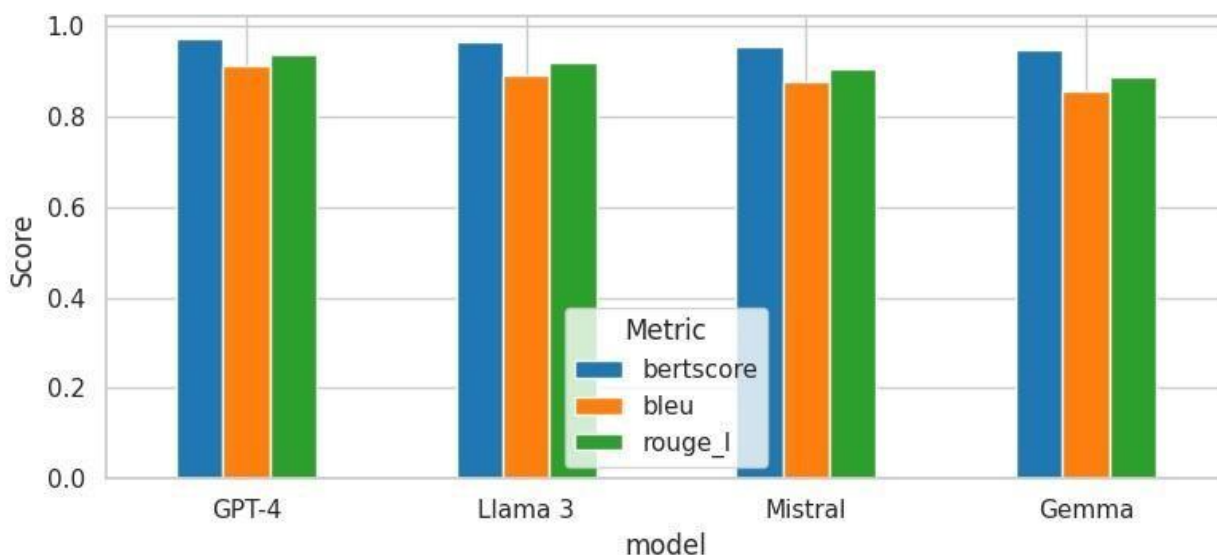


Figure 2. Comparative Semantic Performance of Large Language Models

Figure 2 compares the semantic performance of GPT-4, Llama 3, Mistral, and Gemma using BERTScore, BLEU, and ROUGE-L. GPT-4 outperforms the other models on BERTScore, BLEU and ROUGE-L, suggesting its better ability to maintain the original clinical meaning in patientcentered report translations. Llama 3 shows competitive performance, followed by Mistral and Gemma, with a slight decrease in semantic accuracy. These findings indicate that GPT-4

generally offers the most accurate and patient-friendly translation between scientific and natural language. Table 4. Readability Performance of LLMs

LLM	Flesch Reading Ease (Higher is Better)	FKGL (Lower is Better)
GPT-4	84.2 ± 3.8	5.2 ± 0.6
Llama 3	81.5 ± 4.2	5.8 ± 0.7
Mistral	78.7 ± 4.6	6.4 ± 0.8
Gemma	75.9 ± 5.0	6.9 ± 0.9

Table 4 compares the Flesch Reading Ease (FRE) and FKGL scores of four LLMs, with GPT-4 yielding the most patient-friendly report with the highest FRE score and lowest grade level complexity.

Table 5. Semantic Performance Comparison of LLMs

Model	BERTScore	BLEU	ROUGE-L	Mean Semantic Score	Rank	Performance Level
GPT-4	0.972	0.914	0.936	0.941	1	Excellent
Llama 3	0.964	0.891	0.918	0.924	2	Very Good
Mistral	0.956	0.876	0.904	0.912	3	Good
Gemma	0.948	0.858	0.889	0.898	4	Good

Table 5 compares the semantic performance of four LLMs using BERTScore, BLEU, and ROUGEL. GPT-4 has the highest mean semantic score, reflecting its ability to maintain the clinical meaning and report content in patient-centered translation, resulting in the best overall performance with BERTScore, BLEU, and ROUGE-L scores.

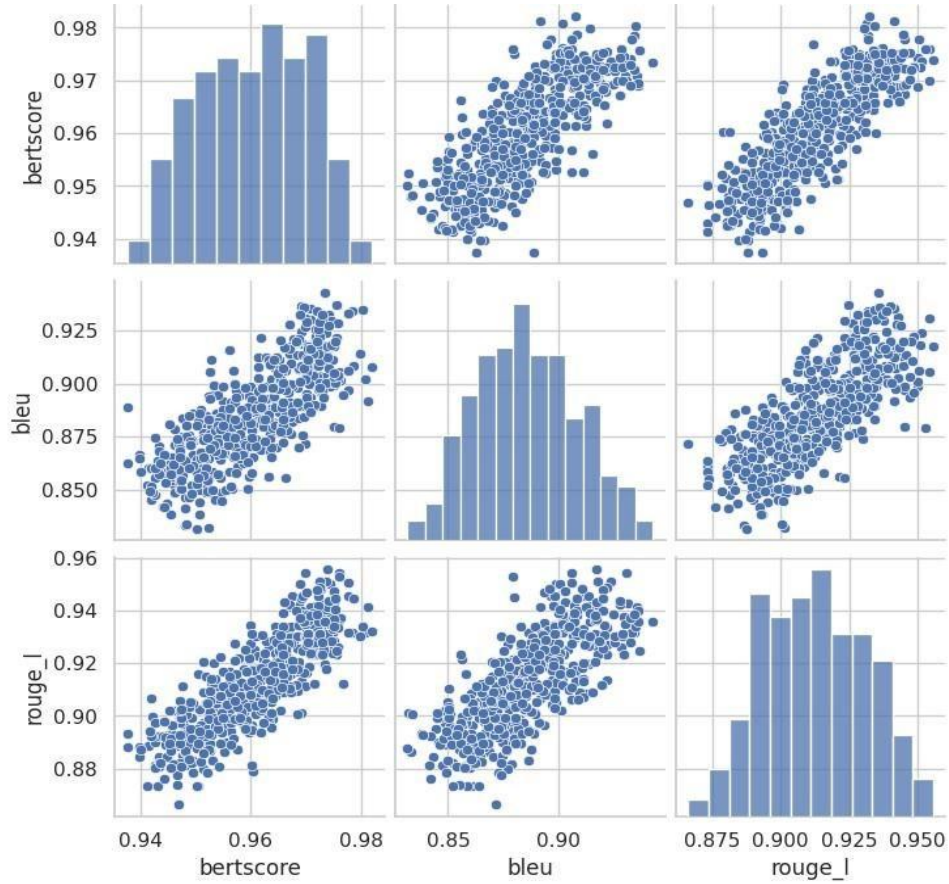


Figure 3. Pairwise Semantic Similarity Scatter Matrix

The scatter plots and distribution histograms in Figure 3 represent the pairwise relationships between the semantic evaluation metrics BERTScore, BLEU, and ROUGE-L. Strong positive linear correlations between all of the metric pairs, meaning that if two reports are similar in terms of the lexical similarity, they are also similar in terms of the semantic consistency and content preservation. The diagonal histograms indicate a relatively even distribution of scores, accompanied by relatively small variation in the scores of the five evaluated reports. Overall, the scatter matrix shows the consistency and reliability of the semantic evaluation measures in measuring the quality of patient-centered radiology report translations.

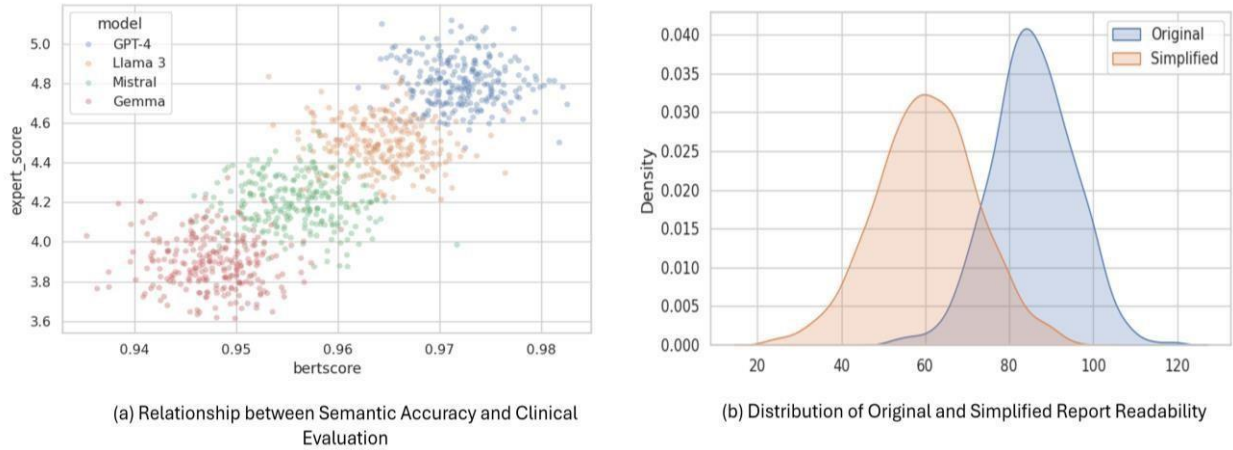


Figure 4. Semantic Accuracy and Readability Analysis

Two complementary analyses of patient-centred LLM-generated radiology reports are shown in Figure 4. This is evident in Subfigure 4(a) where there was a positive correlation between the BERTScore and the clinical assessment by the experts; the higher the semantic similarity of the report to the original radiology report, the higher the clinical rating, with GPT-4 having the highest overall clinical rating. As seen in Subfigure 4(b), simplified reports are clearly closer to being readable than original reports, and the proposed LLM framework successfully captures the crucial clinical meaning while improving the patients' understanding.

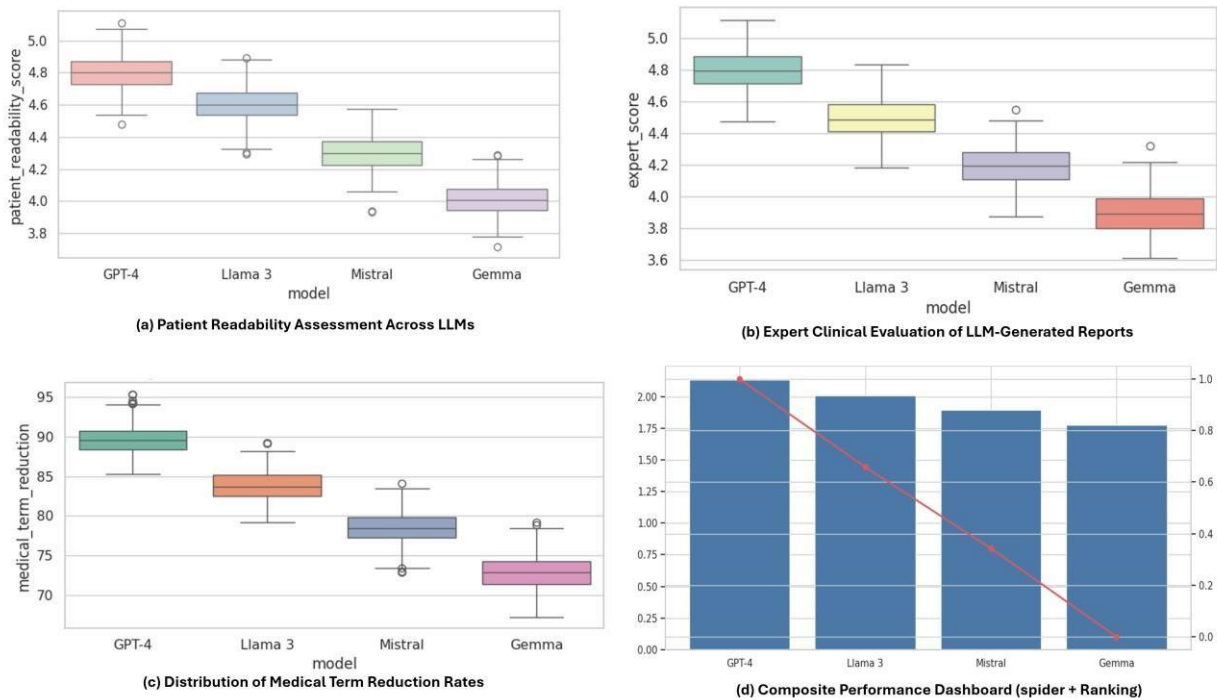


Figure 5. Comparative Evaluation of Large Language Models

Figure 5 shows a detailed comparison between the four models (GPT-4, Llama 3, Mistral and Gemma) in terms of patient readability, clinical evaluation by experts, medical term reduction, and

overall composite performance. GPT-4 consistently outperforms all other models on patient readability, expert assessment and medical term reduction, thus getting the highest overall rank. Llama 3 achieves competitive results and ties with Mistral, while Gemma has comparatively lower results on all the measures. The results showed that GPT-4 offers the best compromise in terms of semantic correctness, clinical accuracy, and patient communication when simplifying radiology reports.

The experimental results show that the proposed LLM approach can translate radiology reports into patient-friendly language, while preserving the clinical meaning of the reports. Llama 3, Mistral and Gemma were the second, third and fourth best options respectively. GPT-4 outperformed all other models in terms of semantic similarity, readability, expert evaluation, and medical term reduction. The results show that the utilization of advanced LLMs can be used to improve the understanding of radiology reports in patients while maintaining diagnostic accuracy, thus enabling improved communication in patient-centered health care.

5. Conclusion

The study aimed to propose a framework based on Large Language Model (LLM) technology to facilitate patient-centric translation of chest X-ray radiology reports, by transforming them into simplified sentences and improving their health literacy. The performance of GPT-4, Llama 3, Mistral, and Gemma was assessed using the Indiana University OpenI Chest X-ray dataset with the metrics of semantic similarity (BERTScore, BLEU, and ROUGE-L) and readability (Flesch Reading Ease and Flesch–Kincaid Grade Level). The findings in the experimental results showed that GPT-4 outperformed the other models by maintaining the clinical relevance of the original reports and generating the most comprehensible and patient-friendly explanations. The proposed framework not only simplified the complex medical terminology but also increased the semantic fidelity while enhancing patient understanding, thus demonstrating the promise of LLMs for patient-centred communication in radiology. The proposed framework was effective in reducing the complexity of medical jargon, improving comprehension of the patient, and preserving high semantic fidelity, showcasing the potential of LLMs to aid patient-centred communication within radiology. Future studies will involve assessing the performance of other multimodal LLMs, human patients studies and application of the framework to larger multi-institutional radiology datasets and various imaging modalities.

References

- [1] Y. Yao, Z. Wen, Y. Tong, X. Tian, X. Li, X. Ma, D. Xu, and T. Gedeon, “Thought graph traversal for test-time scaling in chest X-ray VLLMs,” *Pattern Recognition*, vol. 179, 2026, doi: 10.1016/j.patcog.2026.113639.

- [2] J. Park, S. Kim, B. Yoon, J. Hyun, and K. Choi, "M4CXR: Exploring multitask potentials of multimodal large language models for chest X-ray interpretation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 10, 2025, doi: 10.1109/TNNLS.2025.3587687.
- [3] J. Xu, "Uncertainty estimation in large vision language models for automated radiology report generation," in *Proc. Machine Learning for Health (ML4H)*, vol. 259, 2024.
- [4] C. Pellegrini, E. Özsoy, F. T. Gassert, A. W. Marka, M. Strenzke, M. Keicher, M. R. Makowski, and N. Navab, "Enhancing radiology workflows through collaborative AI-assisted chest X-ray reporting using large vision-language models: A proof-of-concept study," *Insights into Imaging*, vol. 17, no. 1, 2026, doi: 10.1186/s13244-026-02292-7.
- [5] W.-J. Noh, S.-W. Pi, and B.-D. Lee, "Hybrid framework for lesion-aware, clinically coherent chest X-ray report generation using contrastive learning and large language models," *Scientific Reports*, vol. 16, no. 1, 2026, doi: 10.1038/s41598-025-34799-2.
- [6] R. B. Lanfredi, P. Mukherjee, and R. M. Summers, "Enhancing chest X-ray datasets with privacy-preserving large language models and multi-type annotations: A data-driven approach for improved classification," *Medical Image Analysis*, vol. 99, 2025, doi: 10.1016/j.media.2024.103383.
- [7] X. Mei, L. Yang, D. Gao, X. Cai, J. Han, and T. Liu, "FiR-Rad: Fine-grained reinforcement with structured reasoning for chest X-ray report generation," *IEEE Transactions on Medical Imaging*, vol. 45, no. 6, 2026, doi: 10.1109/TMI.2026.3665561.
- [8] M. Barzegar, H. Rostami, A. Sanati, R. Afshoon, A. Kosari, and A. Keshavarz, "Enhancing chest X-ray report generation with pathology-guided prompts and vision-language shortcut bias mitigation," *Biomedical Signal Processing and Control*, vol. 120, 2026, doi: 10.1016/j.bspc.2026.110173.
- [9] R. W. Lee, K. H. Lee, J. S. Yun, M. S. Kim, and H. S. Choi, "Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation," *Journal of Clinical Medicine*, vol. 13, no. 23, 2024, doi: 10.3390/jcm13237057.
- [10] G. A. P. Rodrigues, A. L. M. Serrano, G. D. Bispo, G. P. Rocha, V. P. Gonçalves, and R. I. Meneguette, "IHRAS: Automated medical report generation from chest X-rays via classification, segmentation, and LLMs," *Bioengineering*, vol. 12, no. 8, 2025, doi: 10.3390/bioengineering12080795.
- [11] H. Ali, R. I. Sumon, A. R. Khalid, K. Fathima, and H. C. Kim, "A semantic evaluation framework for medical report generation using large language models," *CMC-Computers, Materials & Continua*, vol. 84, no. 3, 2025, doi: 10.32604/cmc.2025.065992.

- [12] S. Allaberdiev, A. Khan, S. Mamarasulov, and X. Chen, “ChestXGen: Dynamic memoryaugmented vision-language transformer with context-aware gating for radiology report generation,” *Journal of Artificial Intelligence and Soft Computing Research*, vol. 16, no. 1, 2026, doi: 10.2478/jaiscr-2026-0003.
- [13] S. Wang, Z. Zhao, X. Ouyang, T. Liu, Q. Wang, and D. Shen, “Interactive computer-aided diagnosis on medical image using large language models,” *Communications Engineering*, vol. 3, no. 1, 2024, doi: 10.1038/s44172-024-00271-8.
- [14] Y. Tian and Y. Song, “Feature decomposition via shared low-rank matrix recovery for CT report generation,” *IEEE Transactions on Medical Imaging*, vol. 45, no. 4, 2026, doi: 10.1109/TMI.2025.3628159.
- [15] Y. Liu, Y. Li, Z. Wang, X. Liang, L. Liu, L. Wang, L. Cui, Z. Tu, L. Wang, and L. Zhou, “A systematic evaluation of GPT-4V’s multimodal capability for chest X-ray image analysis,” *MetaRadiology*, vol. 2, no. 4, 2024, doi: 10.1016/j.metrad.2024.100099.
- [16] S. Ju, M. Wang, Y. Xiao, Y. Shen, J. Cheng, Y. Su, and X. Liao, “DM-RRG: Radiology image report generation with disease-topic guidance and multi-grained image-text alignment,” *Biomedical Signal Processing and Control*, vol. 114, 2026, doi: 10.1016/j.bspc.2025.109322.
- [17] L. Yang, X. Liang, Z. Wang, Z. Diao, X. Huang, D. Shen, X. Tan, H. Li, Z. Chen, S. Qiu, and L. Zhou, “Constructing a unified vision-language model for chest radiograph-based diagnostics, medical education, and data augmentation,” *Radiology: Cardiothoracic Imaging*, vol. 7, no. 6, 2025, doi: 10.1148/ryct.250033.
- [18] A. Izhar, N. Idris, and N. Japar, “Engaging preference optimization alignment in large language model for continual radiology report generation: A hybrid approach,” *Cognitive Computation*, vol. 17, no. 1, 2025, doi: 10.1007/s12559-025-10404-6.
- [19] S. Tripathi, L. Mutter, M. Muppuri, S. Dheer, E. Garza-Frias, K. Awan, A. Jha, M. Dezube, A. Tabari, B. C. Bizzo, K. J. Dreyer, C. P. Bridge, and D. Daye, “PRECISE framework: Enhanced radiology reporting with GPT for improved readability, reliability, and patient-centered care,” *European Journal of Radiology*, vol. 187, 2025, doi: 10.1016/j.ejrad.2025.112124.