Comparative Evaluation of Large Language Models for Translating Radiology Reports into Hindi

Amit Gupta[1]   Ashish Rastogi[1]   Hema Malhotra[1]   Krithika Rangarajan[1]

[1] Department of Radiology, Dr. B.R.A. IRCH, All India Institute of Medical Sciences, New Delhi

Indian J Radiol Imaging

Address for correspondence   Krithika Rangarajan, MD, FRCR, Department of Radiology, Room No 160D, Radiology, Dr. B.R.A. IRCH, All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110029, India (e-mail: krithikarangarajan86@gmail.com).

Abstract

Objective   The aim of this study was to compare the performance of four publicly available large language models (LLMs)—GPT-4o, GPT-4, Gemini, and Claude Opus—in translating radiology reports into simple Hindi.

Materials and Methods   In this retrospective study, 100 computed tomography (CT) scan report impressions were gathered from a tertiary care cancer center. Reference translations of these impressions into simple Hindi were done by a bilingual radiology staff in consultation with a radiologist. Two distinct prompts were used to assess the LLMs' ability to translate these report impressions into simple Hindi. Translated reports were assessed by a radiologist for instances of misinterpretation, omission, and addition of fictitious information. Translation quality was assessed using Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Translation Edit Rate (TER), and character F-score (CHRF) scores. Statistical analyses were performed to compare the LLM performance across prompts.

Results   Nine instances of misinterpretation and two instances of omission of information were found on radiologist evaluation of the total 800 LLM-generated translated report impressions. For prompt 1, Gemini outperformed others in BLEU ($p < 0.001$) and METEOR scores ($p = 0.001$), and was superior to GPT-4o and GPT-4 in TER and CHRF ($p < 0.001$), but comparable to Claude ($p = 0.501$ for TER and $p = 0.90$ for CHRF). For prompt 2, GPT-4o outperformed all others ($p < 0.001$) in all metrics. Prompt 2 yielded better BLEU, METEOR, and CHRF scores ($p < 0.001$), while prompt 1 had a better TER score ($p < 0.001$).

Keywords
- ► large language models
- ► ChatGPT
- ► radiology reports
- ► BLEU score

Conclusion   While each LLM's effectiveness varied with prompt wording, all models demonstrated potential in translating and simplifying radiology report impressions.

Introduction

Radiology reports are integral to medical decision-making, providing crucial information for diagnosis, treatment planning, and monitoring disease progression of the patient. Conventionally, these reports were tailored for use only by radiologists and referring providers. However, with the advent of telemedicine and patient portals, access to electronic health records has significantly expanded, altering patients' exposure to these medical data.[1] Radiology reports are often laden with complex jargon that can be difficult for patients to understand, undermining patient-centered care. Such lack of understanding can increase patient anxiety, misunderstandings, and emotional distress, particularly due to abnormal findings.[2] However, in a world with ever-increasing pressure on radiology departments across the world, it is unreasonable and impractical to expect radiologists to be able to effect such communication with patients, in the existing workflow routine.

Leveraging the large language models (LLMs) to translate radiology reports into simpler, more accessible language holds significant potential for enhancing patient comprehension and engagement. Based on natural language processing, such models generate human-like text in response to prompts fed in by the user.[3] By translating complex radiological findings into vernacular languages, LLMs can help bridge the language and literacy gap, ensuring that patients from diverse backgrounds have access to understandable health information. This is particularly vital in multilingual countries, like India, where patient literacy and language proficiency vary widely. This strategy supports the global trend of empowering patients in their health care choices, acknowledging that well-informed patients are more likely to actively participate in their treatment processes.[4] These models have previously been used to facilitate the simplification and translation of medical information for patient comprehension.[5–8] However, to the best of our knowledge, a comparative analysis of these LLMs to translate radiology report impressions into vernacular Hindi has not been explored.

Thus, the purpose of this study was to assess and compare the effectiveness of four publicly available LLMs in translating complex radiology reports into simple Hindi using different prompts.

Materials and Methods

This retrospective study was approved by the institutional review board (Institute Ethics Committee All India Institute of Medical Sciences, New Delhi; meeting dated August 22, 2023). The need for patient consent was waived off by the ethics committee, owing to the use of anonymized radiology reports in the study and noninterference with the routine radiology workflow. Also, none of the LLM-generated reports obtained in this study were given to any of the patients or used in clinical practice.

Computed tomography (CT) scan reports for 100 consecutive oncology patients performed at our tertiary care cancer center between January 1 and 10, 2024, were retrieved from the radiology information system (RIS) of our department. The reports consisted of scans with diverse primary cancers and findings. To keep the information in the reports succinct and avoid less relevant findings, we used only the impression sections of the reports in this study, which routinely include all the important information about the patient's tumor including size and extent of the mass, metastatic involvement, and comparison with previous available imaging.

These 100 report impressions were then translated into simple vernacular Hindi by a bilingual (proficient in Hindi and English) radiology technical staff (3 years of experience in radiology-related research), in consultation with a radiologist (10 years of experience in radiology). The generated manual translations were considered as the reference for assessing LLM outputs. Four LLMs—GPT-4o (*https://www.openai.com/gpt-4o*), GPT-4 (*https://www.openai.com/gpt-4*), Google Gemini (*https://www.google.com/gemini*), and Claude Opus (*https://www.anthropic.com/claude-opus*)—were tested in this study. All the four LLMs were provided with two prompts: prompt 1 was "Translate this radiology report into simple Hindi" and prompt 2 was "Translate this radiology report into simple vernacular Hindi explainable to a 15-year-old." Each prompt was followed by the original report impression and queried once for each LLM.

The primary outcome for the study was the quality of translations by the LLMs. For this, the LLM outputs using the two prompts underwent a dual mode of assessment. First, a thorough evaluation of the translated report impressions was conducted by a certified radiologist (8 years of experience in body imaging), specifically for instances of misinterpretation of information, omission of information, and addition of new findings absent in the original report (hallucinations) by the LLM.

Second, the translated report impressions were assessed by four translation quality metrics: Bilingual Evaluation Understudy (BLEU) score, Metric for Evaluation of Translation with Explicit ORdering (METEOR) score, Translation Edit Rate (TER), and Character F-score (CHRF).[9–12] These metrics were computed using Python 3.0 scripts. Each metric provides unique insights into translation quality:

- BLEU score measures the closeness of machine translation to a reference translation based on phrase matches, and a higher score indicates more similarity to the reference, implying better translation quality.
- METEOR score evaluates translation accuracy by aligning words based on their meaning and structure. Higher scores suggest better understanding and translation of content.
- TER quantifies the editing effort required to change a machine translation into the reference translation. Lower scores indicate fewer edits needed, reflecting higher translation accuracy.
- CHRF focuses on the overlap of character n-grams between the translation and the reference. Higher scores demonstrate better fidelity to the reference text.

Statistical analysis was performed including the descriptive statistics (mean, standard deviation) for each score to
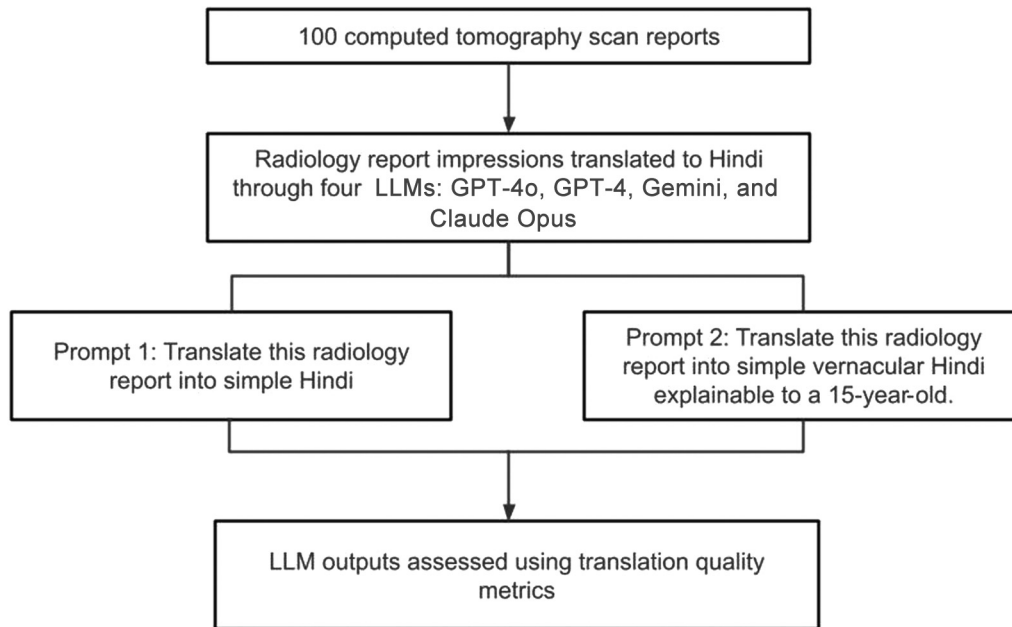
Fig. 1 Study workflow. LLMs, large language models.

summarize the translation quality across all LLMs and prompts. Appropriate statistical tests were used to compare the performance across the different LLMs for each prompt as well as for each LLM across the two prompts. The study workflow is summarized in ►**Fig. 1**.

Results

Radiologist's Evaluation of the LLM-Generated Translations

In the total of 800 translated report impressions (400 for each of the two prompts), there were a total of 9 instances of misinterpretation of the information by the LLMs (1 by GPT-4o, 3 by GPT-4, 3 by Gemini, and 2 for Claude). Examples of the misinterpreted terms and abbreviations include: "adnexae" (interpreted as nearby abdominal structures), "cervical" (referring to cervix, misinterpreted as neck), "OGS" (for osteo-

genic sarcoma, misinterpreted as ovarian sarcoma), "FLR" (for functional liver remnant, misinterpreted as fatty liver ratio), and "BOT" (for base of tongue, misinterpreted as broad ligament). There were two other instances of omission of information from the translated impression, which was present in the original report impression (1 by GPT-4o and 1 by Gemini): omission of "small" size of a hydronephrotic kidney and presence of "pleural effusion" along with metastatic pleural deposits. All the observed mistakes in the LLM outputs were for prompt 2. The analysis of prompt 1 outputs did not reveal any instances of mistakes. Also, there were no instances of hallucinations (addition of fictitious information by the LLMs) in the translated report impressions.

Quantitative Translation Quality Metrics
The mean word counts and the translation quality scores for the LLM-generated translations are detailed in ►**Table 1**.

Table 1 Mean word counts and translation quality metrics calculated for the tested large language models (LLMs) across prompts

Source	Prompt	Mean word count	Mean BLEU score	Mean METEOR score	Mean TER score	Mean CHRF score
Reference		57.86	–	–	–	–
GPT-4o	Prompt 1	51.98	0.098	0.297	0.821	36.353
	Prompt 2	70.33	0.281	0.547	0.804	54.443
GPT-4	Prompt 1	52.20	0.092	0.295	0.837	37.391
	Prompt 2	75.60	0.124	0.41	1.055	44.521
Gemini	Prompt 1	72.20	0.147	0.421	0.934	45.017
	Prompt 2	70.92	0.182	0.456	0.916	47.168
Claude	Prompt 1	64.48	0.070	0.286	1.003	34.992
	Prompt 2	90.22	0.127	0.444	1.173	46.511

Abbreviations: BLEU, Bilingual Evaluation Understudy; METEOR, Metric for Evaluation of Translation with Explicit Ordering; TER, Translation Edit Rate; CHRF, character F-score.

Comparison of the Four LLMs

Preliminary tests for normality and homogeneity of variance assumptions for the BLEU, METEOR, TER, and CHRF scores across the four LLMs indicated that most distributions were non-normal and variances were not homogeneous ($p < 0.05$). Consequently, nonparametric methods (Friedman's test) were used for comparing the models' performance. The Friedman test revealed significant differences in the performance of the four LLMs for both prompts across all metrics (BLEU, METEOR, TER, and CHRF) with p-values less than 0.05. Further exploration through post hoc analysis using the Nemenyi test allowed for detailed pairwise comparisons among the LLMs:

For Prompt 1

- *BLEU scores*: Gemini demonstrated significantly better performance compared with the other LLMs ($p < 0.001$). GPT-4 and GPT-4o showed similar performance to each other with no significant difference ($p = 0.9$).
- *METEOR scores*: Similar to the BLEU scores, Gemini outperformed all other models significantly ($p = 0.001$).

However, there were no significant differences between GPT-4o and Claude ($p = 0.9$), nor between GPT-4o and GPT-4 ($p = 0.855$).

- *TER scores*: GPT-4 showed improved performance compared with GPT-4o and Claude ($p = 0.005$ and 0.001, respectively), while Gemini still had significantly better metrics compared with GPT-4o and GPT-4 ($p = 0.001$). However, the comparison between Gemini and Claude ($p = 0.501$) was not statistically significant.
- *CHRF scores*: Gemini maintained a superior performance ($p = 0.001$) when compared with the other LLMs ($p < 0.001$). Claude, GPT-4, and GPT-4o were again similar in their performance metrics.

The overall trend for prompt 1 indicates that Gemini consistently outperformed the other models across all metrics. GPT-4 and GPT-4o show similar levels of performance, typically less effective than Gemini but more competitive with each other. The heat maps comparing the performance of the four LLMs using prompt 1 for each of the translation metrics are shown in ►**Fig. 2**.
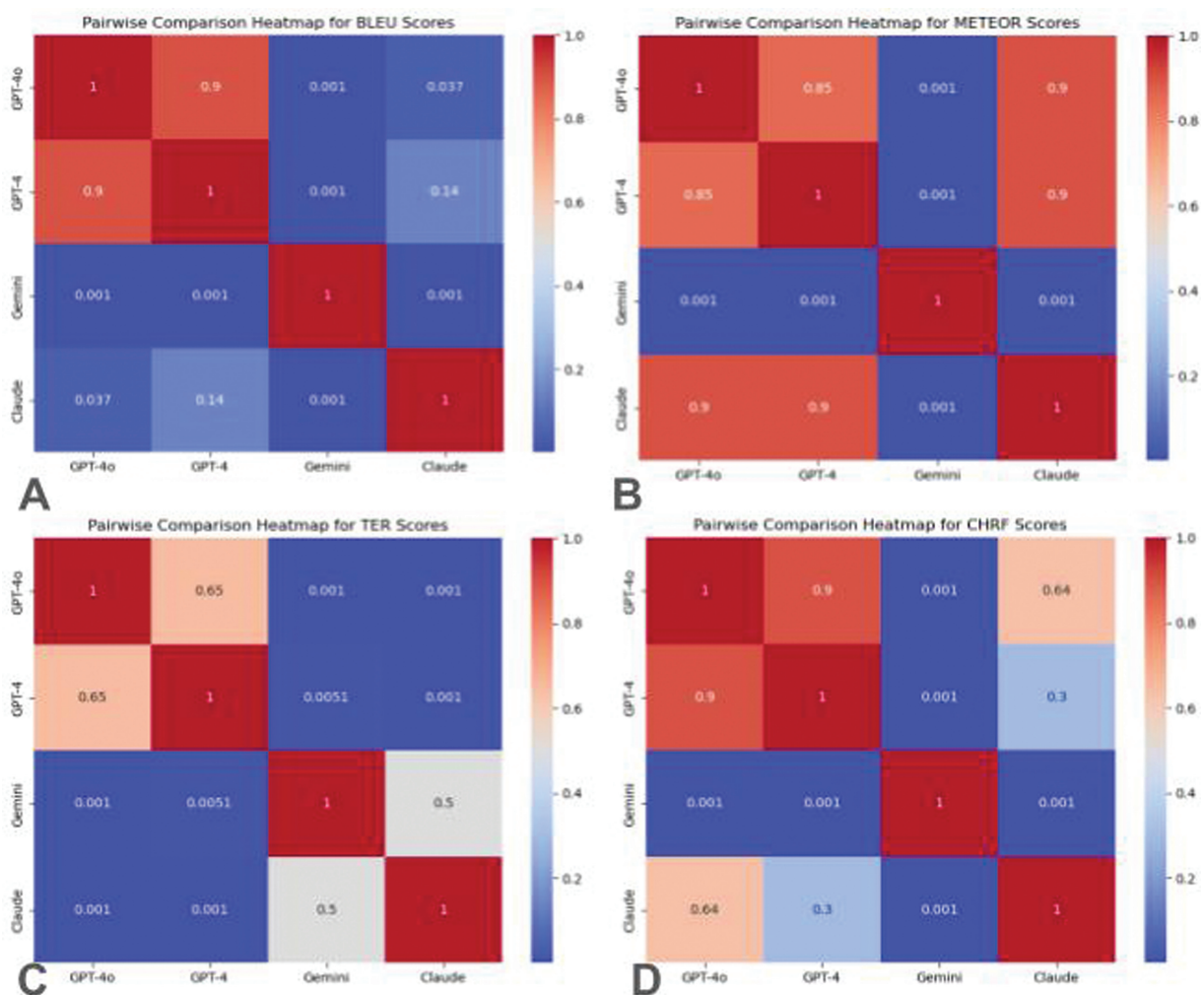

Fig. 2 Heat maps showing *p*-values for comparison of various translation quality metrics among the four large language models (LLMs) for prompt 1: (**A**) Bilingual Evaluation Understudy (BLEU) scores, (**B**) Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores, (**C**) Translation Edit Rate (TER) scores, and (**D**) character F-score (CHRF) scores.

For Prompt 2

- *BLEU scores*: GPT-4o significantly outperformed GPT-4, Gemini, and Claude ($p = 0.001$). GPT-4 and Gemini showed somewhat competitive results ($p = 0.043$), with Gemini also having significantly better scores than Claude ($p = 0.010$). GPT-4 and Claude exhibited similar performances ($p = 0.900$).
- *METEOR scores*: GPT-4o again showed superior performance, significantly outperforming GPT-4, Gemini, and Claude ($p = 0.001, 0.005$, and 0.001, respectively). GPT-4 and Claude had less distinct differences from each other ($p = 0.086$), while Gemini and Claude showed no significant differences ($p = 0.855$).
- *TER scores*: GPT-4o demonstrated a superior performance with the lowest TER score significantly outperforming GPT-4 ($p = 0.001$) and Claude ($p = 0.001$), and performing comparably to Gemini ($p = 0.370$). Gemini also performed significantly better than both GPT-4 ($p = 0.001$) and Claude ($p = 0.001$).
- *CHRF scores*: GPT-4o, with the highest mean CHRF score, demonstrated statistically significant superior perfor-

mance compared with all other models ($p < 0.05$). Gemini and Claude exhibited comparable performance, with no statistically significant difference ($p = 0.90$). While GPT-4's performance was statistically inferior to GPT-4o ($p = 0.001$), it did not significantly differ from Gemini and Claude.

Across the four evaluated metrics, GPT-4o consistently showed a superior performance for prompt 2. GPT-4, Gemini, and Claude demonstrate varied performance across the metrics, with Gemini generally performing better than GPT-4 but similarly to Claude in some metrics. The heat maps comparing the performance of the four LLMs using prompt 2 for each of the translation metrics are shown in ►**Fig. 3**.

Prompt 1 versus Prompt 2

Examples of translated outputs in Hindi across prompts for different LLMs are shown in ►**Table 2**. The results from the Wilcoxon signed-rank test comparing the translation metrics for the two prompts combining the responses from all four LLMs revealed highly significant differences in all
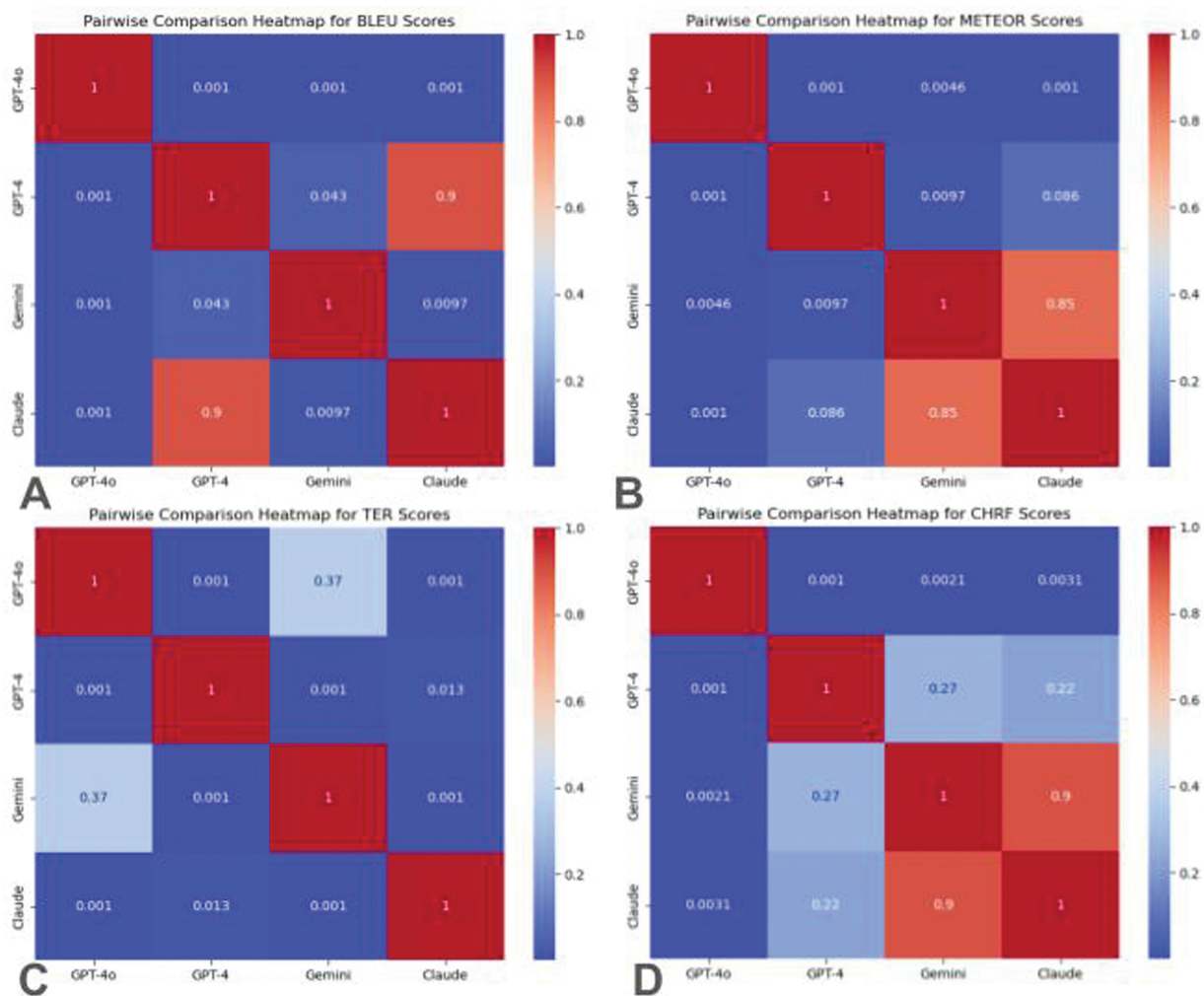
Fig. 3 Heat maps showing *p*-values for comparison of various translation quality metrics among the four large language models (LLMs) for prompt 2: (**A**) Bilingual Evaluation Understudy (BLEU) scores, (**B**) Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores, (**C**) Translation Edit Rate (TER) scores, and (**D**) character F-score (CHRF) scores.

Table 2 Examples of translated outputs in Hindi across prompts and large language models (LLMs)

Prompt	Original report	Reference translation	GPT-4o	GPT-4	Gemini	Claude Opus
Prompt 1	C/o ca lung, central mass in the right hilum with multiple metastatic nodules in the right lung; largest in the right lower lobe and multiple right pleural deposits, pleural effusion. Compared with previous scan, there is no interval change in size of hilar mass and right lower lobe nodule, however there is increase in RUL involvement and size of pleural lesions – progressive disease	दाहिने फेफड़े के मध्य भाग में एक बड़ी गांठ है और इसके आसपास और भी छोटी गांठें हैं, जो फेफड़े के कैंसर के लक्षण हैं। फेफड़े के नीचे वाले हिस्से में सबसे बड़ी गांठ है और फेफड़े के चारों ओर की परत में भी गांठें और पानी जमा हो गया है। पिछले स्कैन की तुलना में, दाहिने निचले हिस्से की गांठ का आकार वैसा ही है, पर दाहिने ऊपरी हिस्से में गांठ बढ़ गई है और फेफड़े के बाहरी परत की गांठें बड़ी हो गई हैं – यानी बीमारी बढ़ रही है।	फेफड़े का कैंसर का मामला, दाएँ हिलम में केंद्रीय गांठ के साथ दाएँ फेफड़े में कई मेटास्टेटिक गांठें; सबसे बड़ी गांठ दाएँ निचले लोब में और कई दाई फुफ्फुस जमाव, फुफ्फुस उत्सर्जन। पिछले स्कैन की तुलना में, दाएँ निचले लोब नोड्यूल की हिलर गांठ के आकार में कोई परिवर्तन नहीं है, लेकिन दाएँ ऊपरी लोब में संलिप्तता और फुफ्फुस घावों के आकार में वृद्धि हुई है – रोग बढ़ रहा है।	फेफड़े के कैंसर का मामला, दाहिने हिलम में केंद्रीय द्रव्यमान के साथ दाहिने फेफड़े में कई मेटास्टेटिक गांठें; सबसे बड़ी गांठ दाहिने निचले लोब में है और दाहिने प्लूरल में कई जमाव और प्लूरल इफ्यूजन हैं। पिछले स्कैन की तुलना में, दाहिने निचले लोब के हिलर द्रव्यमान और गांठ के आकार में कोई अंतराल परिवर्तन नहीं है, हालांकि दाहिने ऊपरी लोब में शामिलता और प्लूरल घावों के आकार में वृद्धि हुई है – बीमारी बढ़ रही है।	फेफड़ों के कैंसर के मामले में, दाहिने फेफड़े के बीच के हिस्से (हिलम) में एक गांठ दिख रही है और साथ ही फेफड़े में कई और गांठें भी हैं। इनमें से सबसे बड़ी गांठ दाहिने फेफड़े के निचले हिस्से (लोब) में है। फेफड़े की झिल्ली में भी कई जगह जमाव है और पानी भी भरा हुआ है। पिछले स्कैन की तुलना में, बीच के हिस्से वाली गांठ और निचले हिस्से की गांठ के आकार में कोई बदलाव नहीं है, लेकिन ऊपरी हिस्से में गांठें बढ़ गई हैं और फेफड़ों की झिल्ली के आसपास जमाव भी बढ़ गया है, जो दर्शाता है कि बीमारी बढ़ रही है।	फेफड़े के कैंसर के मामले में, दाएं हाइलम में एक केंद्रीय द्रव्यमान है और दाएं फेफड़े में कई मेटास्टैटिक नोड्यूल हैं; सबसे बड़ा दाएं निचले लोब में है और दाएं फुफ्फुस पर कई जमाव और प्लूरल एफ्यूजन भी हैं। पिछले स्कैन की तुलना में, दाएं निचले लोब के हाइलर द्रव्यमान या नोड्यूल के आकार में कोई अंतराल परिवर्तन नहीं है, हालांकि दाएं ऊपरी लोब में संलग्नता और प्लूरल घावों के आकार में वृद्धि हुई है – यह प्रगतिशील रोग है।

Prompt	Original report	Reference translation	GPT-4o	GPT-4	Gemini	Claude Opus
Prompt 2	C/o ca lung, central mass in the right hilum with multiple metastatic nodules in the right lung; largest in the right lower lobe and multiple right pleural deposits, pleural effusion. Compared with previous scan, there is no interval change in size of hilar mass and right lower lobe nodule, however there is increase in RUL involvement and size of pleural lesions – progressive disease	दाहिने फेफड़े के मध्य भाग में एक बड़ी गांठ है और इसके आसपास और भी छोटी गांठें हैं, जो फेफड़े के कैंसर के लक्षण हैं। फेफड़े के नीचे वाले हिस्से में सबसे बड़ी गांठ है और फेफड़े के चारों ओर की परत में भी गांठें और पानी जमा हो गया है। पिछले स्कैन की तुलना में, दाहिने निचले हिस्से की गांठ का आकार वैसा ही है, पर दाहिने ऊपरी हिस्से में गांठ बढ़ गई है और फेफड़े के बाहरी परत की गांठें बड़ी हो गई हैं – यानी बीमारी बढ़ रही है।	इस रिपोर्ट में बताया गया है कि फेफड़े का कैंसर है। दाहिने हिलम में एक केंद्रीय गांठ है, और दाहिने फेफड़े में कई मेटास्टेटिक गांठें हैं, जिनमें सबसे बड़ी गांठ दाहिने निचले हिस्से में है। दाहिने फेफड़े के बाहरी परत में भी कई गांठें और पानी जमा हुआ है (प्लूरल इफ्यूजन)। पिछले स्कैन से तुलना करने पर, दाहिने निचले हिस्से की गांठ के आकार में कोई बदलाव नहीं है, लेकिन दाहिने ऊपरी हिस्से में अधिक संक्रमण और प्लूरल गांठों का आकार बढ़ा है – बीमारी प्रगति पर है।	फेफड़े के कैंसर के मामले में, दाहिने हिलम में एक केंद्रीय गांठ है, जिसके साथ दाहिने फेफड़े में कई मेटास्टेटिक नोड्यूल्स हैं; सबसे बड़ा नोड्यूल दाहिने निचले लोब में है और दाहिने प्लीउरा (फेफड़े की बाहरी परत) में कई जमाव और प्लीउरल एफ्यूजन (फेफड़ों में पानी भरना) भी हैं। पिछली स्कैन की तुलना में, दाहिने हिलम की गांठ और दाहिने निचले लोब के नोड्यूल के आकार में कोई बदलाव नहीं है, हालांकि दाहिने ऊपरी लोब में शामिलता और प्लीउरल घावों के आकार में वृद्धि हुई है – यह बीमारी के प्रगतिशील होने का संकेत है।	फेफड़ों के कैंसर की जाँच में पता चला है कि दाहिने फेफड़े के बीच वाले हिस्से में एक गांठ है और दाहिने फेफड़े में ही कई छोटी-छोटी गांठें भी हैं। ये कैंसर फेफड़े की झिल्ली में भी फैल गया है और फेफड़े के आस-पास पानी भी जमा हो गया है। पिछली जाँच से तुलना करने पर, बीच वाली गांठ और निचले हिस्से की गांठ के आकार में कोई बदलाव नहीं आया है। लेकिन, कैंसर फेफड़े के ऊपरी हिस्से में और फैल गया है और झिल्ली पर गांठों का आकार बढ़ गया है, यानी कैंसर बढ़ रहा है।	फेफड़ों के कैंसर का केस है जिसमें दाएं हाइलम (फेफड़ों के केंद्र) में एक मध्य मास (गांठ) है। साथ ही दाएं फेफड़े में कई मेटास्टैटिक नॉड्यूल्स (कैंसर के फैलाव से बनी गांठें) हैं, जिनमें से सबसे बड़ा दाएं फेफड़े के निचले लोब में है। इसके अलावा, प्लूरा (फेफड़ों की बाहरी परत) में भी कई जगह जमाव और प्लूरल इफ्यूजन (तरल पदार्थ का जमाव) मौजूद है। पिछले स्कैन से तुलना करने पर, दाएं फेफड़े के निचले लोब की गांठ और हाइलर मास के आकार में कोई अंतर नहीं आया है। हालाँकि, दाएं फेफड़े के ऊपरी लोब में कैंसर का फैलाव बढ़ा है और प्लूरल गांठों का आकार भी बढ़ा है – बीमारी प्रगतिशील है।

evaluation metrics, with prompt 2 faring significantly better in terms of BLEU, METEOR, and CHRF scores, whereas prompt 1 had significantly better TER scores ($p < 0.001$ for all comparisons).

A comparison of the responses to the two prompts was also done for each LLM. For BLEU scores, prompt 2 significantly outperformed prompt 1 across all models ($p < 0.001$). The METEOR scores showed a similar trend, with prompt 2 performing better for all LLMs ($p = 0.049$ for Gemini and < 0.001 for other LLMs). The TER scores revealed mixed results. Prompt 2 showed better but not significantly different TER scores than prompt 1 for GPT-4o ($p = 0.27$) and Gemini ($p = 0.085$). However, for GPT-4 and Claude ($p < 0.001$), the TER scores were significantly better for prompt 1 than for prompt 2. The CHRF scores consistently favored prompt 2, with significant p-values across all models (< 0.001).

The box plot graphs displaying the different translation metrics across the four LLMs using the two prompts as well as the scores combined for all LLMs for each prompt are shown in ►Fig. 4.

Discussion

In this study, we showed that various LLMs—GPT-4o, GPT-4, Google Gemini, and Claude Opus—can all translate the radiology report impressions to Hindi. However, our findings elucidate significant differences in the performance among these LLMs, as well as according to the specific prompt wording. Notably, Gemini performed comparatively better than other models when the LLMs were provided with a straightforward prompt to translate the reports to Hindi (prompt 1). GPT-4o significantly outperformed the other LLMs when the prompt provided more context, such as requesting translation to simple Hindi explainable to a 15-year-old (prompt 2). The study also brings forth instances of misinterpretation and omission of information in the LLM-generated translated report impressions, signifying the need for expert supervision.

The observed differences in translation metrics across the different LLMs could be attributed to several factors, including variations in training data and preprocessing techniques as well as fundamental differences in LLM architectures and algorithms, that may affect how well the LLMs manage and translate the medical jargon and abbreviations typical in radiology reports.[13,14] Among the myriad of potential prompts available for testing, we used these two prompts to understand how including context in the prompt can impact translation quality. Previous studies have investigated the effect of specific prompt wording and context on the output of LLMs for simplifying radiology reports using various qualitative and quantitative measures.[6,7,15] These studies have reported better performance of GPT models using the prompts where additional context was provided. These
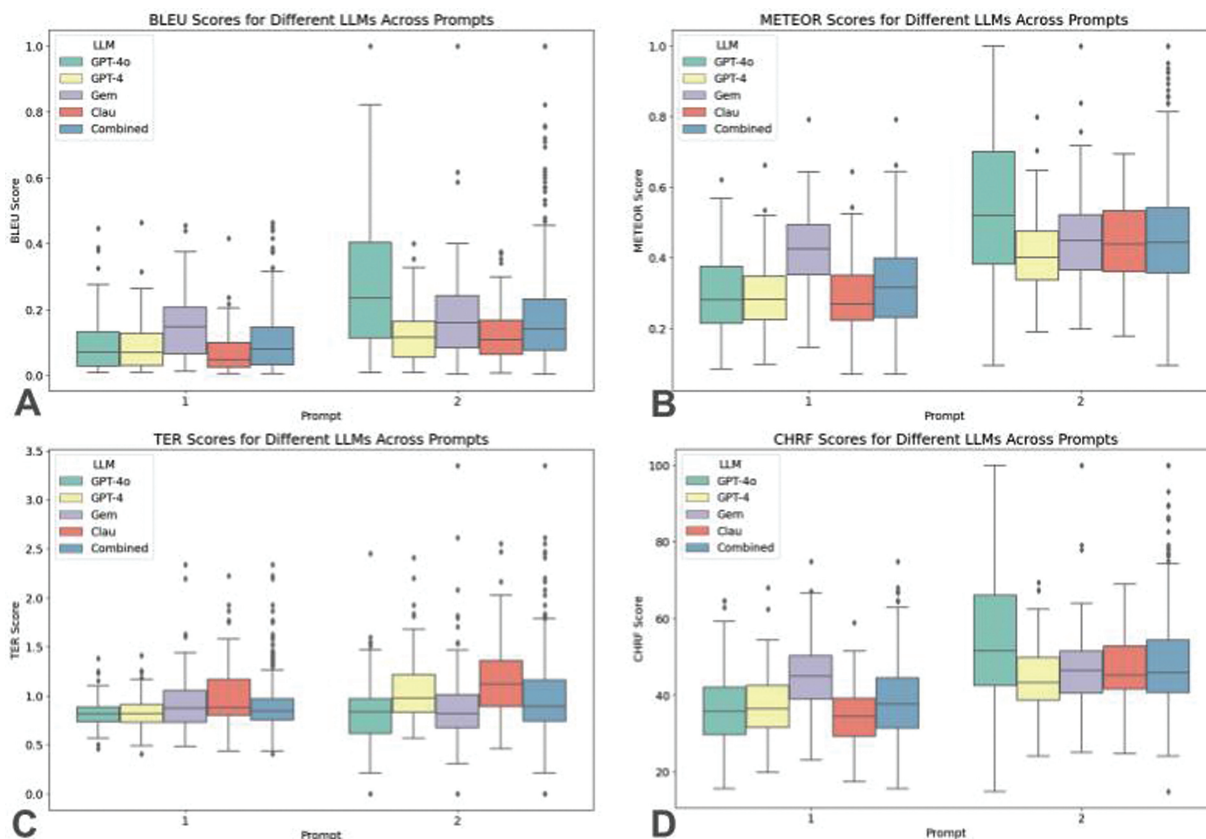

Fig. 4 Box plot graphs showing the various translation quality metrics for each large language model (LLM) as well as combined outputs for prompt 1 and prompt 2: (**A**) Bilingual Evaluation Understudy (BLEU) scores, (**B**) Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores, (**C**) Translation Edit Rate (TER) scores, and (**D**) character F-score (CHRF) scores.

results align with our findings in this study, wherein the prompt with additional context (translation meant for a 15-year-old, prompt 2) performed significantly better than a plain simple prompt asking for Hindi translation (prompt 1), in the majority of the performance metrics. The prompt wording and context related to the specific use case are thus crucial to the performance of these LLMs.

Given that patients are already utilizing these LLMs to clarify medical information, health care providers must recognize the shift in information-sharing dynamics and explore ways to leverage LLMs effectively. The use of LLMs by patients can generate inaccurate and irrelevant outputs.[16] However, our findings indicate that if a radiologist reviews and verifies the LLM output, they can be incorporated into a patient-friendly report translated into vernacular language. This approach can help reduce patient anxiety, misunderstanding, and emotional distress.[5] This is particularly critical in regions with high linguistic diversity, where language barriers can impede patient care. While LLMs show potential in enhancing patients' understanding of their radiology reports by translating them, it is crucial to balance readability with clinical accuracy. Oversimplification and strict translation of complex medical terms could lead to clinical errors, highlighting the indispensable role of health care providers in ensuring effective communication and comprehension.

Our study had a few limitations. The relatively small sample size, the specific context of oncology patients at a tertiary care center, and the focus on a single language pair might limit the generalizability of the study findings. Although there are many currently available LLMs, we could include only four state-of-the-art LLMs for analysis in this study due to practical constraints. Future research could expand both the scope and the duration of the study, include other LLMs, and explore translations into multiple languages to provide a more comprehensive evaluation of the technology.

Conclusion

In conclusion, while the effectiveness of each LLM depended on the specific prompt wording, all four models evaluated (GPT-4o, GPT-4, Gemini, and Claude Opus) were capable of translating the radiology reports. It is important to note that our findings do not endorse any specific LLM; instead, this study demonstrates the potential of LLMs to translate complex medical documents into a simple vernacular language. Proper fine-tuning and customization of each LLM are essential to ensure effective translation while preserving the clinical integrity of the reports. Future research should consider a longitudinal study design and a more diverse dataset to enhance the validity and generalizability of these results.

Ethical Approval and Patient Consent
Ethical approval has been obtained from the institutional review board. Patient consent was not applicable for this study and was waived off by the ethics committee.

References
1 Patil S, Yacoub JH, Geng X, Ascher SM, Filice RW. Radiology reporting in the era of patient-centered care: how can we improve readability? J Digit Imaging 2021;34(02):367–373
2 Bruno B, Steele S, Carbone J, Schneider K, Posk L, Rose SL. Informed or anxious: patient preferences for release of test results of increasing sensitivity on electronic patient portals. Health Technol (Berl) 2022;12(01):59–67
3 Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024;310(01):e232756
4 Itri JN. Patient-centered radiology. Radiographics 2015;35(06): 1835–1846
5 Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34(05):2817–2825
6 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. medRxiv 2023 (e-pub ahead of print). Doi: 10.1101/2023.06.04.23290786
7 Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6(01):9
8 Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. Cureus 2023;15(12):e50881
9 Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. Paper presented at: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; July 6–12, 2002; Philadelphia, PA
10 Lavie A, Denkowski M. The METEOR metric for automatic evaluation of machine translation. Mach Transl 2009;23:105–115
11 Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. Cambridge, MA: Association for Machine Translation in the Americas; 2006:223–231
12 Popović M chrF: character n-gram F-score for automatic MT evaluation. In: Bojar O, Chatterjee R, Federmann C, et al., eds. Proceedings of the Tenth Workshop on Statistical Machine Translation [Internet]. Lisbon, Portugal: Association for Computational Linguistics; 2015:392–395
13 Zhao WX, Zhou K, Li J, et al. A Survey of Large Language Models [Internet]. arXiv; 2023. Accessed June 16, 2024 at: http://arxiv.org/abs/2303.18223
14 Fan L, Li L, Ma Z, Lee S, Yu H, Hemphill L. A bibliometric review of large language models research from 2017 to 2023 [Internet]. arXiv; 2023. Accessed June 16, 2024 at: http://arxiv.org/abs/2304.02020
15 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;101:137–141
16 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(02):e232561