Can I Read My X-Ray Report? Towards Accessible Radiology Report in Low-Resource African Context

Aziza Umer Yibrie[1], **Abinew Ali Ayele**[1], **Shamsuddeen Hassan Muhammad**[2],
Seid Muhie Yimam[3],
[1]Bahir Dar University, [2]Bayero University Kano, [3]University of Hamburg

Abstract

Healthcare communication in native languages is a critical unmet need for Amharic-speaking populations in Ethiopia and diaspora communities. This study develops a preliminary framework for translating English radiology reports into Amharic using multilingual machine translation systems (Google Translate, *NLLB-200, M2M100*) and instruction-tuned large language models (*GPT-4.1-mini, Gemini-2.0-Flash*, and others), combined with human-in-the-loop evaluation. A subset of 100 IU X-Ray reports is translated, with 67 reports manually annotated for systematic assessment. Preliminary evaluation shows that Google Translate achieves the highest overall performance (BLEU 46.17, chrF 48.74, ROUGE-L 42.39), while LLMs such as Gemini-2.0-Flash (chrF 27.55) and GPT-4.1-mini (BLEU 13.14) produce fluent Amharic text but require substantial post-editing to ensure correct clinical terminology. Human annotator analysis emphasizes the importance of expert oversight in achieving terminological accuracy and report completeness. This work establishes an initial benchmark, introduces a scalable workflow, and provides a foundation for developing reliable Amharic radiology report translation systems, with potential applicability to other low-resource languages.

1 Introduction

Healthcare communication in native languages remains a fundamental yet unmet need for millions of Ethiopians and Ethiopian migrants globally. Amharic is spoken by over 57 million people as a first or second language, with additional speakers across Ethiopia and in diaspora communities in the United States, Europe, the Middle East, and beyond(Maatoug et al., 2025). Despite this substantial speaker population, Amharic remains virtually absent from medical AI systems and healthcare technology development, creating significant barriers to health equity and clinical communication (Pakray et al., 2025).

Radiology report generation represents a critical healthcare application where language barriers create tangible harm. Automated systems for generating accurate, clinically appropriate radiology reports have advanced significantly in high-resource languages, as shown using the IU X-Ray dataset of over 7k chest X-rays with English reports (Chen et al., 2021). However, Ethiopian patients, whether in local residents or the diaspora community settings, who cannot read radiological reports in their native Amharic language, face substantial barriers to understanding their medical imaging results, participating in clinical decision-making, and accessing quality healthcare (Zheng et al., 2025).

The IU X-Ray(Demner-Fushman et al., 2016) dataset provides the foundation necessary for developing cross-lingual models through transfer learning and machine translation, enabling researchers to leverage existing data rather than undertaking prohibitively resource-intensive new data collection in Amharic (Zhou, 2024). By building on this dataset, it is possible to develop automated systems that generate accurate Amharic-language radiology reports, potentially improving healthcare accessibility and patient engagement for Amharic-speaking populations.

Key Contributions To the best of our knowledge, there is currently no publicly reported system for generating radiology reports in Amharic, highlighting a critical gap in AI-driven healthcare for low-resource African languages. This work makes the following key contributions:
(i) Automated Amharic radiology dataset: We develop the first systematically curated bilingual radiology resource in Amharic.
(ii) Comparative analysis of translation methods: We evaluate multiple machine translation approaches, including Google Translate, NLLB-200,

M2M100, and large language model variants, for producing clinically accurate Amharic reports.
(iii) Radiology report translation annotation tool: We develop a web-based tool that generates initial translations, supports human curation by clinical experts, and standardizes terminology, enabling consistent, high-quality bilingual dataset creation.
(iv) Scalable and replicable workflow: We establish a workflow for dataset creation and translation that can be adapted to other low-resource African languages, supporting reproducible development of bilingual healthcare resources.

2 Related Work

Radiology Report Generation Radiology report generation (RRG) automatically generates free-text clinical descriptions from medical images to reduce radiologist workload and improve efficiency (Sloan et al., 2024). Early encoder-decoder methods with CNNs and RNNs had limited context windows (Nurbanu Aksoy and Nishant Ravikumar and Alejandro F. Frangi, 2023), but the field has evolved toward transformer-based architectures with attention mechanisms, contrastive learning, and reinforcement learning (Dhamanskar and Thacker, 2024). Modern systems integrate multimodal information including patient demographics and clinical history (Wang et al., 2024), with recent advances extending to 3D imaging and longitudinal data incorporation. Despite progress, RRG faces significant challenges including maintaining clinical accuracy to avoid hallucinations (Hamamci et al., 2024), handling imbalanced normal/abnormal findings (Khare et al., 2021), ensuring fairness across patient populations (Pang et al., 2023), and developing appropriate evaluation metrics (Sloan et al., 2024). Large language models offer promising solutions by leveraging pretrained linguistic knowledge fine-tuned for medical tasks (Jiang et al., 2025).

Radiology Report Translation for Low-Resource Languages Radiology reports contain domain-specific terminology and structured descriptions of imaging findings, making their translation into low-resource languages challenging due to limited parallel data and the need to preserve clinical accuracy. Prior work shows that general-purpose translation systems often struggle with radiology-specific expressions, motivating the use of multilingual models designed for low-resource settings. Massively multilingual neural machine translation systems address data scarcity through cross-lingual transfer. The M2M-100 model demonstrates that direct many-to-many translation can improve performance for low-resource languages, though domain-specific medical accuracy remains limited (Fan et al., 2021). The No Language Left Behind (NLLB) model further advances this direction through a sparsely gated mixture-of-experts architecture, achieving an average BLEU improvement of 44 percent over prior systems across more than 40,000 translation directions (Costa-Jussà et al., 2024; Team et al., 2022). These characteristics make NLLB particularly relevant for translating radiology reports into languages such as Amharic. Commercial systems such as Google Translate are also commonly used in practice and provide strong baselines for evaluating applied medical translation despite not being domain-specific.

LLMs for Radiology Report Translation Large language models (LLMs) have recently been applied to translate radiology reports across multiple languages, achieving high scores on BLEU, TER, and character-level metrics while producing outputs that radiologists find clear and consistent, though clinical terminology accuracy can vary (Meddeb et al., 2024; Lee et al., 2025). Comparative studies show that model choice and prompt design significantly affect translation quality, particularly for low-resource languages (Gupta et al., 2025). Beyond direct translation, LLMs can simplify reports for patient comprehension, supporting multilingual healthcare workflows (Khanna et al., 2024). Systematic reviews highlight their contextual understanding and generalization capabilities, while noting the need for rigorous evaluation and domain-specific validation in clinical settings (Shool et al., 2025).

Language Barriers in Ethiopian Healthcare Language barriers in Ethiopian healthcare affect both patient-provider communication and clinical documentation. Although Amharic is widely spoken, much documentation is in English, creating challenges for both healthcare workers and patients (Olani et al., 2023). In rural settings, healthcare professionals often simplify explanations for Amharic-speaking patients, which can result in miscommunication and the omission of important clinical details. These barriers reduce patient comprehension, confidence in diagnoses, and healthcare utilization (Barrio-Ruiz et al., 2024), particularly in radiology, where complex findings require specialized explanation. Such communication gaps contribute to preventable errors, higher treatment costs, and longer hospital stays, ultimately undermining patient outcomes and overall health system performance (Sawadogo et al., 2023; Roudsari et al., 2023).

Healthcare Access for Ethiopian Migrants Ethiopian migrant communities face healthcare challenges due to limited proficiency in host-country languages, which hinders access and comprehension of medical information (Zheng et al., 2025). Language barriers are compounded by cultural differences, unfamiliarity with healthcare systems, and discrimination (Bäumel et al., 2024). Radiology is particularly affected, as imaging reports are typically available only in host-country languages, limiting migrants' understanding of essential diagnostic results in settings(Zheng et al., 2025).

3 Methodology

This study presents an approach for generating Amharic radiology reports from English source texts. This approach combines multiple machine translation (MT) systems with LLMs based translation and human evaluation, aiming to produce clinically accurate, linguistically fluent, and readable reports under low-resource language settings. At this stage, the methodology focuses on initial proof-of-concept experiments using a limited dataset, with plans to scale up to the full IU X-Ray dataset.

Data Source and Selection We used the publicly available IU X-Ray dataset as the primary data source for English radiology reports (Chen et al., 2021). The IU X-Ray dataset is an open-source medical imaging corpus that contains chest X-ray images paired with expert-written radiology reports. It is widely used in medical vision and language research and provides clinically structured reports that are suitable for downstream text generation and translation tasks.

For the initial phase of this study, a subset of 100 radiology reports was randomly sampled from the IU X-Ray dataset. From this subset, 67 reports were randomly sampled for detailed human evaluation by medical professionals due to the cost and clinician availability. The selected reports include both normal and abnormal findings and reflect common chest radiology report structures. This subset is intended to support exploratory analysis and baseline establishment rather than broad clinical generalization. Future work will extend evaluation to larger portions of the dataset to capture broader linguistic and clinical variation.

Machine Translation for Amharic Radiology Reports We selected models spanning different architectures and sizes to capture diverse approaches to Amharic radiology translation, enabling a meaningful comparison of translation quality, clinical adequacy, and usability rather than focusing on model complexity. We translated 100 English radiology reports into Amharic using three machine translation systems: M2M100_418M[1], NLLB-200-distilled-600M[2], and Google Translate. M2M100_418M enables direct translation between multiple languages without pivoting through English, helping preserve meaning in low-resource languages such as Amharic (Fan et al., 2021). NLLB-200-distilled-600M serves as the supervised neural machine translation baseline, leveraging a sparsely gated mixture-of-experts architecture and cross-lingual representations to improve translation quality and preserve domain-specific terminology (Costa-Jussà et al., 2024; Team et al., 2022). Google Translate is included as a widely used commercial system to provide industry-standard performance and a comparative baseline. Figure 1 shows the annotation tool used by experts to validate and correct LLM translations.

4 Experimental Setups and Results

4.1 Setups

We used instruction-tuned large language models (LLMs), including GPT-4o-mini[3] , GPT-4.1-mini[4], LLaMA-3.3-70B[5], Mistral-Large[6], Gemma-3-27B[7], DeepSeek-R1[8], DeepSeek-R1-0528[9], and Gemini-2.0-Flash[10]. These models were chosen for their ability to handle low-resource languages

[1] https://huggingface.co/facebook/m2m100_418M

[2] https://huggingface.co/facebook/nllb-200-distilled-600M

[3] GPT-4o-mini link

[4] https://huggingface.co/openai/gpt-4.1-mini

[5] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

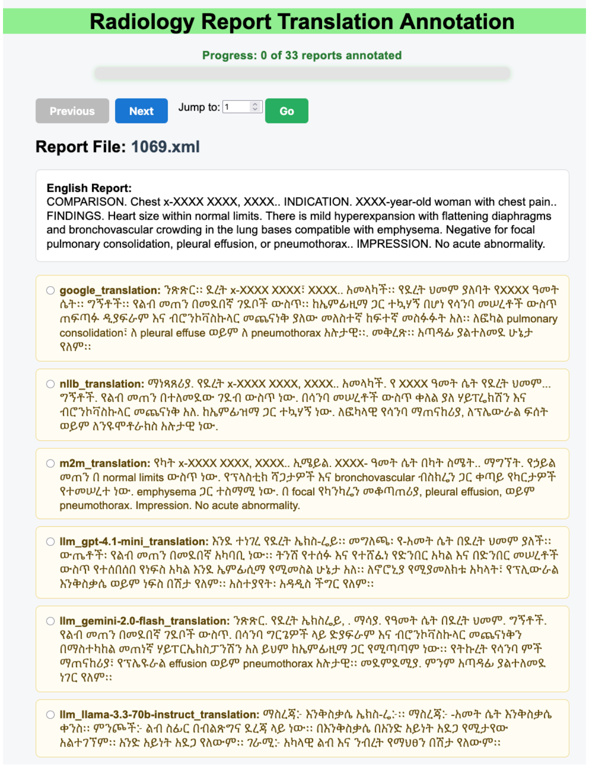
[6] https://mistral.ai/news/mistral-3

[7] https://huggingface.co/google/gemma-3-27b-it

[8] https://huggingface.co/deepseek-ai/DeepSeek-R1

[9] https://console.cloud.google.com/vertex-ai/publishers/deepseek-ai/model-garden/deepseek-r1-0528-maas

[10] https://oumi.ai/docs/en/latest/user_guides/infer/inference_engines.html



Progress: 0 of 33 reports annotated

Report File: 1069.xml

English Report:
COMPARISON. Chest x-XXXX XXXX, XXXX.. INDICATION. XXXX-year-old woman with chest pain.. FINDINGS. Heart size within normal limits. There is mild hyperexpansion with flattening diaphragms and bronchovascular crowding in the lung bases compatible with emphysema. Negative for focal pulmonary consolidation, pleural effusion, or pneumothorax.. IMPRESSION. No acute abnormality.

○ **google_translation:** ንጽጽር:: ደረት x-XXXX XXXX፣ XXXX.. አመላካች:: የደረት ህመም ያለባት የXXXX ዓመት ሴት:: ግኝቶች:: የልብ መጠን በመደበኛ ገደቦች ውስጥ:: ከኤምፊዚማ ጋር ተኳሃኝ በሆነ የሳንባ መሠረቶች ውስጥ ጠፍጣፋ ዲያፍራም እና ብሮንኮቫስኩላር መጨናነቅ ያለው መለስተኛ ከፍተኛ መስፋፋት አለ:: ለፎካል pulmonary consolidation፣ ለ pleural effuse ወይም ለ pneumothorax አሉታዊ::. መቅረጽ:: አጣዳፊ ያልተለመደ ሁኔታ የለም::

○ **nllb_translation:** ማነጻጸሪያ. የደረት x-XXXX XXXX, XXXX.. አመላካች. የ XXXX ዓመት ሴት የደረት ህመም... ግኝቶች. የልብ መጠን በተለመደው ገደብ ውስጥ ነው. በሳንባ መሠረቶች ውስጥ ቀለል ያለ ሃይፐሬክሽን እና ብሮንኮቫስኩላር መጨናነቅ አለ. ከኤምፊዝማ ጋር ተኳሃኝ ነው. ለፎካላዊ የሳንባ ማጠናከሪያ, ለፕሌውራል ፍሰት ወይም ለንዩሞቶራክስ አሉታዊ ነው.

○ **m2m_translation:** የካት x-XXXX XXXX, XXXX.. ኢሜይል. XXXX- ዓመት ሴት በካት ስሜት.. ማግኘት. የኃይል መጠን በ normal limits ውስጥ ነው. የፕላስቲክ ሻጋታዎች እና bronchovascular ብስክሬን ጋር ቀጣይ የካርታዎች የተመሠረተ ነው. emphysema ጋር ተስማሚ ነው. በ focal የካንካሬን መቆጣጠሪያ, pleural effusion, ወይም pneumothorax. Impression. No acute abnormality.

○ **llm_gpt-4.1-mini_translation:** እንደ ተነገረ የደረት ኤክስ-ሬይ:: መግለጫ: የ-አመት ሴት በደረት ህመም ያለች:: ውጤቶች: የልብ መጠን በመደበኛ አካባቢ ነው:: ትንሽ የተሰፉ እና የተሸፈነ የድንበር አካል እና በድንበር መሠረቶች ውስጥ የተሰበሰበ የነፍስ አካል እንደ ኤምፊሲማ የሚመስል ሁኔታ አለ:: ለኖሮኒያ የሚያመለክቱ አካላት፣ የፕሊውራል እንቅስቃሴ ወይም ነፍስ በሽታ የለም:: አስተያየት: አዳዲስ ችግር የለም::

○ **llm_gemini-2.0-flash_translation:** ንጽጽር. የደረት ኤክስሬይ, . ማሳያ. የዓመት ሴት በደረት ህመም. ግኝቶች. የልብ መጠን በመደበኛ ገደቦች ውስጥ. በሳንባ ግርጌዎች ላይ ድያፍራም እና ብሮንኮቫስኩላር መጨናነቅን በማስተካከል መጠነኛ ሃይፐርኤክስፓንሽን አለ ይህም ከኤምፊዚማ ጋር የሚጣጣም ነው:: የትኩረት የሳንባ ምች ማጠናከሪያ፣ የፕሌዩራል effusion ወይም pneumothorax አሉታዊ:: መደምደሚያ. ምንም አጣዳፊ ያልተለመደ ነገር የለም::

○ **llm_llama-3.3-70b-instruct_translation:** ማስረጃ፦ እንቅስቃሴ ኤክስ-ሬ፦:: ማስረጃ፦ -አመት ሴት እንቅስቃሴ ቀንስ:: ምንጮች፦ ልብ ስፊር በብልጽግና ደረጃ ላይ ነው:: በእንቅስቃሴ በአንድ አይነት አደጋ የሚታየው አልተገኘም:: አንድ አይነት አደጋ የለውም:: ገራሚ፦ አካላዊ ልብ እና ንብረት የማህፀን በሽታ የለውም::

Figure 1: Radiology report translation annotation tool. Experts check the LLM translation; if satisfactory, they select the LLM output, otherwise they provide their own translation.

and domain-specific medical terminology. Each LLM generated Amharic translations directly from the English reports.

For instruction-tuned large language models, translations were generated using a single prompt instructing the model to translate the full radiology report from English into Amharic while preserving clinical meaning and report structure. Default decoding parameters provided by each platform were used, with no additional sampling constraints or temperature tuning. No terminology constraints or external medical lexicons were applied, and no post-processing steps were performed beyond basic text normalization. Commercial and multilingual NMT systems were used through their standard inference interfaces.

4.2 Results

Automatic Evaluation Reference translations were constructed through expert validation and correction of system-generated outputs. While machine translation systems, including Google Translate, were used to produce initial drafts, all reference texts were finalized by medical professionals, who either verified the translation or rewrote it to ensure clinical accuracy, terminological correctness, and linguistic fluency in Amharic. Reference translations therefore do not consist of unedited system outputs, and models contributing initial drafts were not used as independent reference generators.

Google Translate achieved the strongest overall performance, with BLEU of 46.17, chrF of 48.74, and ROUGE-L of 42.39, consistently producing translations closely aligned with the reference texts. Among large language models (LLMs), *GPT-4.1-mini* and *Gemini-2.0-Flash* outperformed other models. *Gemini-2.0-Flash* achieved the highest chrF score among LLMs (27.55), indicating improved character-level fidelity, while *GPT-4.1-mini* demonstrated more balanced performance across BLEU and ROUGE-L. These results suggest that instruction-tuned LLMs can generate fluent and coherent Amharic text, although precise alignment with clinical terminology remains limited.

Multilingual neural machine translation models showed mixed performance. *NLLB-200* achieved moderate ROUGE-L scores, indicating better preservation of report structure, whereas *M2M100* exhibited weaker lexical and character-level alignment. Several LLMs, including *DeepSeek-R1*, *DeepSeek-R1-0528*, *LLaMA-3.3-70B*, and *Gemma-3-27B*, performed poorly across automatic metrics despite often producing fluent surface-level Amharic text.

Human Evaluation In addition to automatic metrics, we conducted a human evaluation with two licensed medical doctors. The experts reviewed translations produced by both machine translation systems and LLMs, assessing whether each output accurately preserved the clinical meaning of the source report and was linguistically fluent and appropriate in Amharic. Translations judged satisfactory were accepted as it is; otherwise, the experts provided corrected versions. This process ensured that the translations used for analysis were clinically accurate, fluent, and suitable for the target language context.

Human evaluation focused on overall clinical adequacy and preference, reflecting realistic post-editing and usability judgments by medical professionals. Annotators indicated whether a translation was acceptable for clinical interpretation or whether another system's output was preferred. We note that this binary and preference-based setup does not capture fine-grained error categories such as missing findings, mistranslations, or halluci-

Model	BLEU	chrF	ROUGE-L
DeepSeek-R1	0.52	7.35	0.42
DeepSeek-R1-0528	0.49	7.07	0.43
GPT-4.1-mini	13.14	21.14	20.43
Gemini-2.0-Flash	14.19	27.55	10.68
Gemma-3-27B	2.33	12.36	7.87
Google Translate	**46.17**	**48.74**	**42.39**
LLaMA-3.3-70B	4.05	10.24	11.02
M2M100	1.29	7.31	20.50
NLLB-200	5.89	19.61	28.51

Table 1: Automatic evaluation results for English–Amharic radiology report translation on the 67 human-annotated reports. Higher values indicate better performance for all metrics.

nated content. Developing a more detailed, clinically informed error taxonomy is left for future work.

Human Annotator Model Preference Analysis In addition to automatic metrics, we analyzed human annotator preferences recorded through the custom web-based annotation tool (Figure 1) developed in this work. For each of the 67 human-evaluated reports, annotators select the system output that requires the least correction or the manually revised translations when none of the system outputs are acceptable. Google Translate was the most frequently selected system, serving as the preferred base translation for 30 out of 67 reports. Annotators favored these outputs due to their relatively accurate lexical choices and consistent handling of radiology terminology, which reduced the need for extensive correction. Among LLMs, Gemini-2.0-Flash was selected in 5 cases, making it the most frequently chosen LLM. Annotators noted its fluency and readability, although post-editing was still required to ensure terminological accuracy. The NLLB-200 system was selected directly in one case. In 31 cases, annotators selected "Other," indicating that none of the provided system outputs were sufficiently accurate. In these cases, annotators used the tool to manually rewrite or heavily edit translations, often combining elements from multiple systems. These selections typically corresponded to reports containing complex findings or less frequent clinical terminology. Overall, annotator preferences reveal that human usability and clinical adequacy are not fully captured by automatic metrics. While some systems achieved moderate scores, annotators prioritized terminological correctness, completeness, and adherence to standard radiology report structure.

5 Discussion

The results demonstrate that general-purpose translation systems, particularly Google Translate, provide strong performance for Amharic radiology translation based on both automatic metrics and human evaluation. Google Translate frequently achieved the highest BLEU, chrF, and ROUGE-L scores and was preferred by annotators for most reports, reflecting its ability to capture common radiology terminology and report structure with minimal post-editing.

Instruction-tuned LLMs, including GPT-4.1-mini and Gemini-2.0-Flash, produced fluent and coherent Amharic text, highlighting their potential for low-resource language translation. Gemini-2.0-Flash achieved the highest chrF among LLMs, indicating strong character-level fidelity. However, it often retained specialized medical terms in English rather than translating them, signaling a limitation in domain-specific terminology coverage. Despite this, Gemini-2.0-Flash consistently generated high-quality radiology translations, demonstrating its promise as a preliminary tool in this domain. These findings suggest that further domain adaptation could improve clinical adequacy, particularly in ensuring accurate and complete representation of medical terminology.

Multilingual neural machine translation models, such as NLLB-200, preserved report structure moderately well, demonstrating that cross-lingual transfer and sparsely gated architectures can support structural fidelity in low-resource languages. In contrast, M2M100 showed weaker lexical and character-level alignment, highlighting the limitations of general many-to-many translation without domain adaptation.

Human annotator analysis reinforced the importance of expert oversight. Many reports required substantial post-editing, indicating that neither LLMs nor multilingual NMT models alone currently guarantee terminological correctness or report completeness. This emphasizes the critical role of hybrid approaches that combine automated translation with human-in-the-loop validation, particularly in specialized clinical domains such as radiology.

The workflow developed in this study, including the web-based annotation tool, enabled systematic translation refinement, terminology standardization, and creation of a bilingual reference dataset. This framework establishes a foundation for scalable Amharic radiology translation and could be adapted to other low-resource languages. By providing a structured approach to translation evaluation and refinement, this work contributes both a practical workflow and a high-quality reference dataset for future research. Future work will explore domain adaptation for LLMs, integration of multimodal imaging features, and development of evaluation metrics that more accurately capture clinical adequacy beyond traditional automatic scores.

6 Conclusion

This study presented preliminary development and evaluation of English–Amharic radiology report translation systems using multilingual machine translation, large language models, and human-in-the-loop annotation. A total of 100 radiology reports from the IU X-Ray dataset were translated, with 67 reports manually annotated for systematic evaluation.

Results indicated that Google Translate provided comparatively strong overall performance, while LLMs such as Gemini-2.0-Flash and GPT-4.1-mini demonstrated promise in producing fluent Amharic translations, though consistent clinical terminology alignment remained challenging. Multilingual NMT models such as NLLB-200 preserved report structure but require further adaptation for medical domains.

A key contribution of this ongoing work is the web-based annotation tool, which supports human correction, terminology standardization, and the creation of a bilingual reference dataset. Preliminary findings highlight the continued importance of human oversight, as many reports still require expert rewriting to ensure clinical reliability.

The study establishes an initial benchmark and a workflow for Amharic radiology translation. Future efforts will expand human annotation coverage, incorporate clinician-in-the-loop evaluation, refine LLM prompt and terminology constraints, and explore multimodal approaches integrating imaging features with text.

Ethical Considerations

This ongoing research involved translation of clinical text, carrying inherent risks if outputs were used inappropriately. All translations were performed on de-identified public IU X-Ray reports, and outputs were not used for clinical decision-making.

Human annotators were bilingual individuals with healthcare familiarity, specifically general doctors, but not all were certified radiologists. Outputs were carefully reviewed to ensure terminological accuracy within the scope of this study.

The work also aimed to improve accessibility and equity by providing Amharic translations, supporting health information access for native speakers within Ethiopia and diaspora communities. Automated translations were considered supplementary, requiring expert oversight for clinical applications.

The workflow, including the web-based annotation tool, was designed to be transparent and replicable, supporting responsible development of bilingual medical resources for low-resource languages while ensuring patient privacy and data security. Future work will continue to focus on clinical validation, human-in-the-loop evaluation, and ethical deployment of translation systems in healthcare settings.

Limitation

This study had several limitations. First, the evaluation was based on 67 human-annotated reports, which limited statistical generalization. While sufficient for exploratory analysis, larger annotated datasets are needed for more robust conclusions.

Second, the evaluation relied primarily on automatic metrics, which may not fully reflect clinical adequacy or patient safety. Although annotator preferences and post-editing partially addressed this gap, formal validation by licensed radiologists was not conducted.

Third, the annotation process was performed by bilingual annotators with healthcare familiarity, specifically general doctors, but not uniformly by certified radiologists. This may have affected the precision of certain specialized terms.

Fourth, the annotation tool requires reviewing multiple outputs, imposing cognitive load that may limit scalability to larger datasets. Additionally, displaying model identities in the annotation interface could have introduced bias in human evaluation. Annotators were instructed to focus on translation quality and clinical adequacy, but future studies should implement a blind evaluation setup to minimize potential bias and enhance the robustness of human preference analysis.

Finally, the study focused on text-only translation and did not incorporate visual features from

X-ray images, which are central to fully integrated radiology report generation systems.

References

Carmen Barrio-Ruiz, Regina Ruiz de Viñaspre-Hernandez, Sofia Colaceci, Raul Juarez-Vela, Ivan Santolalla-Arnedo, Angela Durante, and Marco Di Nitto. 2024. Language and Cultural Barriers and Facilitators of Sexual and Reproductive Health Care for Migrant Women in High-Income European Countries: An Integrative Review. *Journal of Midwifery & Women's Health*, 69(1):71–90.

Anika Christin Bäumel, Alexandra Sauter, Andrea Weber, Michael Leitzmann, and Carmen Jochem. 2024. Subjective health status and health literacy of African refugees and asylum seekers in Germany: a cross-sectional survey. *International Journal of Migration, Health and Social Care*, 20(2):261–275.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Marta R. Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2024. Scaling Neural Machine Translation to 200 Languages. *Nature*, 630(8018):841–846.

Dina Demner-Fushman, Manish D Kohli, Michael B Rosenman, Sonya E Shooshan, Luis Rodriguez, Sameer Antani, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Prajakta Dhamanskar and Chintan Thacker. 2024. A detailed analysis of deep learning-based techniques for automated radiology report generation. *International Journal of Electrical and Computer Engineering (IJECE)*, 14(5):5906–5915.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Amit Gupta, Ashish Rastogi, Hema Malhotra, and Krithika Rangarajan. 2025. Comparative Evaluation of Large Language Models for Translating Radiology Reports into Hindi. *The Indian Journal of Radiology & Imaging*, 35(1):88–96.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pages 476–486, Cham. Springer Nature Switzerland.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *The Thirteenth International Conference on Learning Representations*.

Praneet Khanna, Gagandeep Dhillon, Venkata Buddhavarapu, Ram Verma, Rahul Kashyap, and Harpreet Grewal. 2024. Artificial Intelligence in Multilingual Interpretation and Radiology Assessment for Clinical Language Evaluation (AI-MIRACLE). *Journal of Personalized Medicine*, 14(9).

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U. Deva Priyakumar, and C.V. Jawahar. 2021. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036.

Ryan C. Lee, Roham Hadidchi, Michael C. Coard, Yossef Rubinov, Tharun Alamuri, Aliena Liaw, Rahul Chandrupatla, and Tim Q. Duong. 2025. Use of Large Language Models on Radiology Reports: A Scoping Review. *Journal of the American College of Radiology*.

Taha Maatoug, Anissa Ouahchi, Farah Seedat, Anna Deal, Abdedayem Khelifi, Mohamed Douagi, Wejdene Mansour, Ali Mtiraoui, Bouchra Assarag, Ana Requena-Méndez, Dominik Zenner, and Stella Evangelidou. 2025. Healthcare access among sub-Saharan migrants and refugees in Tunisia: an interpretative qualitative study. *BMC Medicine*, 23(1):547.

Aymen Meddeb, Sophia Lüken, Felix Busch, Lisa Adams, Lorenzo Ugga, Emmanouil Koltsakis, Antonios Tzortzakakis, Soumaya Jelassi, Insaf Dkhil, Michail E. Klontzas, Matthaios Triantafyllou, Burak Kocak, Sabahattin Yüzkan, Longjiang Zhang, Bin Hu, and 1 others. 2024. Large language model ability to translate ct and mri free-text radiology reports into multiple languages. *Radiology*, 313(3):e241736.

Nurbanu Aksoy and Nishant Ravikumar and Alejandro F. Frangi. 2023. Radiology report generation using transformers conditioned with non-imaging data. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, volume 12469, page 124690O. International Society for Optics and Photonics, SPIE.

Amanti Baru Olani, Ararso Baru Olani, Takele Birhanu Muleta, Dame Habtamu Rikitu, and Kusa Gemeda Disassa. 2023. Impacts of language barriers on healthcare access and quality among Afaan Oromoo-speaking patients in Addis Ababa, Ethiopia. *BMC Health Services Research*, 23.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Ting Pang, Peigao Li, and Lijie Zhao. 2023. A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22.

Robab Latifnejad Roudsari, Farangis Sharifi, and Fatemeh Goudarzi. 2023. Barriers to the participation of men in reproductive health care: a systematic review and meta-synthesis. *BMC Public Health*, 23(1):818.

Pengdewendé Maurice Sawadogo, Drissa Sia, Yentéma Onadja, Idrissa Beogo, Gabriel Sangli, Nathalie Sawadogo, Assé Gnambani, Gaëtan Bassinga, Stephanie Robins, and Éric Tchouaket Nguemeleu. 2023. Barriers and facilitators of access to sexual and reproductive health services among migrant, internally displaced, asylum seeking and refugee women: A scoping review. *PLOS ONE*, 18.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A Systematic Review of Large Language Model (LLM) Evaluations in Clinical Medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated Radiology Report Generation: A Review of Recent Advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Xinyi Wang, Grazziela Patrocinio Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. 2024. A Survey of Deep Learning-based Radiology Report Generation Using Multimodal Data. *ArXiv*, abs/2405.12833.

Wanqing Zheng, Sara Akram, and Ruqin He. 2025. Exploring health care experiences and access challenges: A qualitative study of African migrants in Guangzhou, China. *African journal of reproductive health*, 29 8:40–50.

Zhong Zhou. 2024. Massively Multilingual Text Translation For Low-Resource Languages. *ArXiv*, abs/2401.16582.