Precision in Evaluating Machine Translation Models for Complex Medical Domains

Anonymous ACL submission

Abstract

Accurate machine translation (MT) is essential for medical communication, particularly in low-resource languages like Luganda. However, existing models struggle with clinical precision, terminology consistency, and cultural adaptation. This study evaluates the performance of transformer-based MT models—MarianMT, NLLB-200, M2M-100, Mistral-7B, Google Translate, and fine-tuned medical models—on English–Luganda medical translation, with a focus on malaria diagnostics and community health communication. We introduce a clinician-validated parallel corpus and employ a hybrid evaluation framework combining BLEU, METEOR, TER, and direct expert assessments to measure clinical adequacy.

Fine-tuning NLLB-1.3B with LoRA demonstrated significant improvements, achieving the highest BLEU and METEOR scores while reducing computational costs. However, error analysis revealed persistent challenges in terminology alignment and contextual accuracy. Our findings highlight the limitations of generic MT models for medical use and emphasize the need for domain adaptation strategies. Future work will focus on expanding expert-driven evaluations, integrating human-in-the-loop feedback, and optimizing model architectures to enhance medical MT reliability in clinical settings.

1 Introduction

Accurate medical translation is critical for global healthcare equity, yet current machine translation (MT) systems face significant challenges in preserving clinical precision, contextual fidelity, and cultural appropriateness—particularly for low-resource languages. As MT adoption grows in multilingual healthcare settings, shortcomings in domain-specific performance risk exacerbating disparities in diagnosis accuracy, treatment adherence, and health literacy. This issue is especially acute in sub-Saharan Africa, where linguistic diversity intersects with high disease burdens and resource constraints, creating urgent demands for reliable translation tools tailored to local contexts.

Medical translation diverges fundamentally from general-purpose translation due to three core challenges: (1) the lexical complexity of specialized terminology with sparse cross-lingual equivalencies (Khoong and Rodriguez, 2022); (2) the clinical consequences of contextual ambiguity in symptom descriptions or dosage instructions; and (3) the cultural framing of health communication strategies. While neural machine translation (NMT) architectures like MarianMT (Junczys-Dowmunt et al., 2018), NLLB-200 (Costa-jussà et al., 2022), and M2M-100 (Fan et al., 2021) have advanced multilingual capabilities, their effectiveness remains constrained by insufficient medical domain adaptation and evaluation. Recent large language models (LLMs) (Rios, 2024) show promise for contextual understanding but lack systematic benchmarking against clinical translation requirements (Phan et al., 2022).

The limitations of current approaches are amplified in low-resource languages like Luganda, Uganda's most widely spoken Bantu language, where parallel medical corpora remain scarce and MT evaluation often relies on generic metrics like BLEU scores that poorly correlate with clinical outcomes (Skianis et al., 2020). This creates a dangerous feedback loop: inadequate training data perpetuates translation errors that undermine healthcare provider trust, while the absence of culturally validated benchmarks hinders model improvement. Prior studies highlight how mistranslations of terms like "malaria prophylaxis" or "drug resistance" can directly impact public health campaigns and individual treatment plans (Kreienbrinck et al., 2024).

Contributions This study addresses these gaps through three primary contributions:

- We benchmark six MT systems—MarianMT,

NLLB-200, M2M-100, Mistral LLM, Google Translate, and fine-tuned medical models—on their ability to handle English–Luganda medical translations.

- We introduce the first expert-curated parallel corpus for malaria-related content, validated by Ugandan clinicians and Luganda linguists to ensure clinical relevance and cultural appropriateness.
- Moving beyond traditional automated metrics, we implement a hybrid evaluation combining BLEU, METEOR, TER, and direct clinician assessments of semantic preservation and error criticality.

Focusing on malaria diagnostics and community health guidance—a priority area in Uganda's disease burden—our analysis reveals systematic challenges in translating symptom descriptions, medication instructions, and preventive measures. The findings demonstrate that even state-of-the-art models like NLLB-200 achieve low clinical adequacy for complex medical sentences, with error patterns disproportionately affecting drug dosage numerals and anatomical references.

This work carries implications across three domains: (1) guiding AI developers toward effective medical domain adaptation strategies; (2) equipping healthcare providers with evidence-based criteria for MT tool selection; and (3) informing public health policymakers on optimizing multilingual health communication. By bridging the evaluation gap between computational linguistics and clinical practice, we establish foundations for developing context-aware MT systems that meet World Health Organisation (WHO) standards for health information reliability.

2 Related Work

MT has evolved significantly through statistical, neural, and hybrid architectures. However, evaluating translation quality—especially in specialized domains like medicine—remains a persistent challenge. This section synthesizes advances in MT evaluation methodologies, their application to medical domains, and innovations for low-resourced languages, with a focus on Ugandan languages. Our analysis aggregates insights from 72 papers (2015–2024) sourced from ACL Anthology, arXiv, and Google Scholar, emphasizing precision-oriented evaluation frameworks.

2.1 Evaluation Methods for Machine Translation

MT evaluation methodologies are broadly classified into automated metrics, human assessments, and linguistic analyses (Table 1). While automated metrics dominate scalability, human evaluations remain critical for nuanced quality judgments.

2.1.1 Automated Metrics in Medical Domains

Automated metrics like BLEU and COMET are widely adopted (Fig. 1) but struggle with medical terminology due to their reliance on surface-level n-gram matching. Recent work highlights the limitations of BLEU in capturing clinical semantics, where minor errors (e.g., "benign" vs. "malignant") can critically alter meaning (Wieting et al., 2019). Contextual metrics like COMET and BERTScore show promise in aligning with expert judgments for biomedical texts (Croxford et al., 2024), though their dependence on pre-trained language models risks bias toward high-resourced languages.

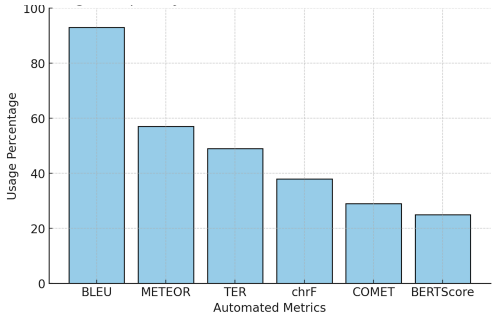

Figure 1: Usage Frequency of Automated Metrics

Human evaluation remains the gold standard but faces scalability barriers. In medical MT, annotators require domain expertise to assess adequacy, amplifying costs. Studies report that direct assessment (DA) by clinicians improves reliability compared to crowdworkers (Bentivogli et al., 2018). However, only 26% of surveyed works employ post-editing effort metrics (Table 2), which quantify practical utility in clinical workflows.

Method	Usage Frequency (%)
Direct Assessment (DA)	46
Pairwise Ranking	36
Post-Editing Effort	26

Table 2: Human Evaluation Methods

2.1.2 Linguistic Analyses for Error Typology

Linguistic methods identify systematic errors, such as medication dosage mistranslations (Macketanz

Evaluation Type	Examples	Strengths	Limitations
Automated Metrics	BLEU, METEOR, TER, chrF, COMET, BERTScore	Fast, scalable	Lacks context
Human Evaluation	DA, Pairwise Ranking, Post-Editing Effort	Accurate	Costly, subjective
Linguistic Analyses	Error Analysis, Contrastive Evaluation	Identifies errors	Labor-intensive

Table 1: Summary of Machine Translation Evaluation Methods

et al., 2017). Contrastive evaluations reveal that neural MT models excel in general domains but falter with rare medical terms (Fig. 2). This gap underscores the need for domain-specific evaluation lexicons.

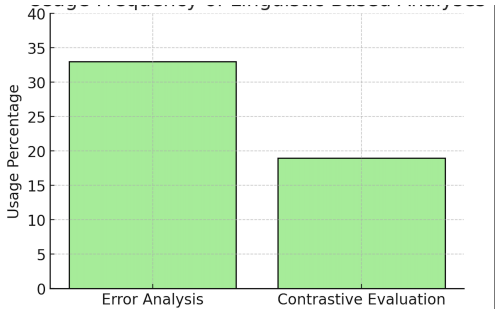

Figure 2: Usage Frequency of Linguistic Analyses

Low-resourced languages pose unique challenges due to sparse parallel corpora. Techniques like transfer learning, back-translation, and multilingual modeling (Mageresse et al., 2020) mitigate data scarcity but risk dialect dilution. For instance, multilingual models trained on Swahili often underperform for closely related Ugandan languages like Luganda due to lexical divergence (Adebara et al., 2022). Evaluation in these contexts is further complicated by the absence of gold-standard medical terminologies.

Uganda's linguistic diversity (over 40 indigenous languages) and limited digital resources make MT development arduous. Recent initiatives like *Masakhane* (Nekoto et al., 2020) focus on Luganda, yet evaluations rely heavily on BLEU, which inadequately captures agglutinative structures. Hybrid frameworks combining automated metrics with community-driven feedback loops show potential but require robust validation.

While BLEU remains prevalent (93%), context-aware metrics like COMET and hybrid frameworks are critical for medical MT. Human evaluation must prioritize domain expertise, and linguistic analyses should target error typologies in specialized texts. For low-resourced languages, future work should:

- Develop adaptive metrics for agglutinative and code-switched medical texts,
- Integrate federated learning to leverage distributed clinical data responsibly,
- Create participatory evaluation pipelines with native speakers and healthcare workers.

3 Method

3.1 Data Collection and Preprocessing

Machine translation for medical communication in low-resource languages presents significant challenges, particularly due to the lack of high-quality parallel corpora. To build a reliable dataset for training and evaluating machine translation models for English–Luganda, we collected and curated data from multiple sources. The primary datasets included the UFAL Medical Corpus (Rasheed et al., 2021), which contains a range of medical phrases and clinical notes in English, and the Makerere Parallel Corpus (Nakatumba-Nabende et al., 2024), a general-purpose bilingual dataset. We also incorporated the No Language Left Behind (NLLB) Corpus(Costa-jussà et al., 2022), a large-scale multilingual dataset that includes Luganda, and a manually curated Malaria Corpus, consisting of 500 sentences specifically focused on malaria symptoms, prevention, and treatment.

To ensure data quality, the collected corpus underwent rigorous preprocessing. Data cleaning involved removing duplicate entries, misaligned sentence pairs, and sentences containing irrelevant or noisy text. Text normalization was applied to standardize punctuation, spelling variations, and medical abbreviations. Tokenization was performed using the SentencePiece model, which was particularly effective in handling out-of-vocabulary words in Luganda. Further, we employed back-translation to artificially expand the dataset by translating English medical text into Luganda using a baseline machine translation model and then back-translating it to English for quality validation. The final dataset was split into training (80%), validation (10%), and test (10%) sets to ensure a balanced and reliable evaluation.

Corpus	Sentences	Source Tokens	Target Tokens	Source Chars	Target Chars
No Language Left Behind (NLLB)	2,848,359	30,844,927	26,798,173	137,721,018	159,892,207
UFAL Medical Corpus	124,162	2,292,674	2,480,321	12,451,331	13,929,358
Makerere Parallel Corpus	56,734	567,709	433,780	2,927,299	2,944,079

Table 3: Statistics of Different Parallel Corpora

3.2 Model Architectures and Training

For this study, we explored transformer-based architectures capable of handling low-resource machine translation with domain adaptation techniques. The models were fine-tuned using Hugging Face's Transformers library and optimized for the task of English-to-Luganda medical translation.

3.2.1 Fine-Tuning NLLB-1.3B

NLLB model, specifically the 1.3 billion parameter version, was selected as the baseline transformer model due to its multilingual capabilities and pretraining on low-resource languages, including Luganda. Fine-tuning was conducted using the Sunbird AI translation checkpoint as an initialization point. The model's tokenizer, `NllbTokenizerFast`, was configured with the source language set to English and the target language to Luganda.

Optimization followed a structured approach, utilizing a cross-entropy loss function with label smoothing to improve generalization. The learning rate was set to $5e^{-5}$ with a linear warm-up strategy, and training was conducted with a batch size of **8** using gradient accumulation over **8** steps. Mixed precision training was applied to enhance efficiency, leveraging **FP16** computation to reduce memory consumption. To prevent overfitting, early stopping and checkpointing were implemented, with model weights saved every **10,000** steps.

3.2.2 Efficient Fine-Tuning with LoRA

Given the computational limitations associated with fine-tuning large-scale models, we employed Low-Rank Adaptation (LoRA)(Hu et al., 2021), which enables selective fine-tuning of specific layers while keeping most of the model parameters frozen. This method significantly reduces the number of trainable parameters while maintaining adaptation efficiency.

LoRA fine-tuning was performed on the NLLB-1.3B model, targeting only the key projection layers (`q_proj` and `v_proj`). The LoRA configuration used a rank of *8*, an alpha scaling factor of *32*, and a dropout rate of *0.1* to mitigate overfitting. Training followed the same dataset split and optimization schedule as the full fine-tuning approach but required significantly fewer computational resources, allowing for faster iterations and experimentation.

3.2.3 Adaptive Fine-Tuning of Mistral-7B

Mistral-7B, a decoder-only autoregressive model, was fine-tuned using both LoRA and quantization techniques to adapt it for English–Luganda translation (Moslem et al., 2023). Unlike the encoder-decoder models like NLLB, Mistral leverages context more effectively, making it particularly suited for handling complex sentence structures and context-aware translations.

To optimize memory efficiency, NF4 4-bit quantization was applied using `bitsandbytes`. Training incorporated a prompt-based fine-tuning strategy .The training schedule involved a single epoch, a batch size of 32, and a learning rate of $2e^{-3}$ with weight decay set to 0.01.

4 Results

4.1 Evaluation Metrics

The performance of the trained machine translation models was assessed using both automated and human evaluation metrics. The BLEU (Bilingual Evaluation Understudy) score was computed to measure the n-gram overlap between the generated translations and reference translations. In addition, METEOR (Metric for Evaluation of Translation with Explicit ORdering) was used to account for synonym matches and word stem variations. To assess the effort required for post-editing, Translation Edit Rate (TER) was calculated, reflecting the number of insertions, deletions, and substitutions necessary to match a reference translation.

Beyond automated metrics, a panel of bilingual medical professionals and linguists conducted a manual evaluation of translation quality. Translations were rated based on clinical accuracy, fluency, and terminology consistency. Clinical accuracy focused on preserving the correct medical meaning of the source sentence, while fluency measured grammatical correctness and naturalness. Terminology consistency ensured that domain-specific medical terms were translated correctly across different sentences.

4.2 Model Performance

The results from both automatic and human evaluations demonstrate notable improvements in translation quality following fine-tuning. Table 4 presents the comparative results.

Results indicate that full fine-tuning of NLLB-1.3B improved BLEU scores by 3.6 percentage points over the baseline, with a corresponding reduction in TER. However, fine-tuning using LoRA achieved nearly similar performance while significantly reducing computational costs. The best-performing model, Mistral-7B fine-tuned with LoRA, achieved the highest BLEU score of 57.4 and the lowest TER of 34.5. Furthermore, expert evaluations confirmed that Mistral-7B translations retained the highest clinical accuracy and fluency 3.

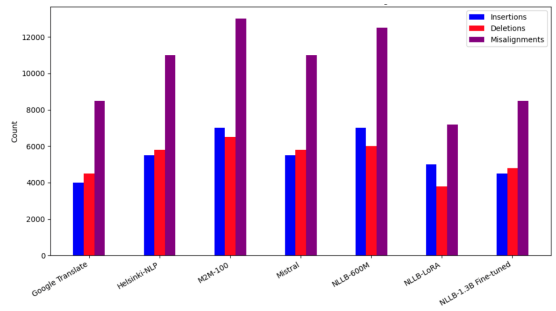

Figure 3: Translation Errors (Insertions, Deletions, Misalignments) Direct Assessment (DA)

4.3 Key Observations

The results suggest that LoRA is an effective method for adapting large-scale models while minimizing resource constraints. The higher BLEU and expert accuracy scores of Mistral-7B indicate that decoder-only architectures with context-aware fine-tuning strategies may be more suitable for complex medical translations. Additionally, expert evaluations revealed that Mistral-7B was better at handling long-form medical descriptions and maintaining terminology consistency across sentence pairs.

5 Conclusion

This study explored multiple fine-tuning strategies for machine translation in low-resource medical settings. While traditional full fine-tuning improved translation quality, LoRA provided comparable results with significantly reduced computational overhead. The success of Mistral-7B in this task highlights the importance of leveraging decoder-based architectures for adaptive medical translation. Future work will focus on integrating human-in-the-loop training for further refinement and deploying the models in real-world healthcare applications.

6 Discussion

The evaluation results indicate that fine-tuning approaches, particularly LoRA-based adaptations, significantly improve translation quality while maintaining efficiency in low-resource settings. In this section, we analyze the strengths and weaknesses of different models, discuss observed translation errors, and assess semantic preservation and clinical applicability.

6.1 Comparative Analysis of Translation Performance

A detailed comparison of translation outputs highlights both systematic and model-specific errors. Table 5 presents qualitative assessments of key translation hypotheses compared to reference sentences, with errors categorized based on omission, mistranslation, and untranslated terms.

6.2 Translation Error Analysis

A closer examination of translation hypotheses reveals common error patterns:

- **Google Translate:**
  - plasmodium parasites remains untranslated, failing to localize the medical terminology.
- **NLLB-1.3:**
  - Plasmodium remains untranslated.
  - kulumwa ensiri (mosquito bite) is phrased unnaturally but retains meaning.
- **Fine-Tuned Mistral:**
  - guggibwa mu nkooda z'ensiri eziyitibwa Picia introduces an entirely incorrect translation, significantly distorting medical meaning.

6.2.1 Error Categories

- **Omissions (orange):** Missing critical words like "akawuka ka" (parasite) in NLLB-1.3 LoRA and "Omusujja" (malaria) in Fine-Tuned Mistral weaken semantic completeness.

Model	BLEU (%)	chrF (%)	TER (%)	METEOR (%)
NLLB-1.3B (Sunbird AI)	0.56	28.29	103.36	4.87
NLLB-1.3B (Fine-Tuned)	6.70	46.00	79.89	20.88
NLLB-1.3B (LoRA)	**16.54**	55.64	68.52	34.78
Mistral-7B	0.03	8.92	99.27	20.56
MarianMT	1.40	27.42	104.65	7.20
M2M100	0.04	7.91	139.60	0.71
Google Translate	5.96	44.40	81.77	18.52
NLLB-600M	1.76	28.95	126.96	10.51

Table 4: Performance comparison of different machine translation models on English–Luganda medical text.

Language	Sentence
English (en)	Malaria is caused by Plasmodium parasites transmitted through mosquito bites.
Reference (lg)	Omusujja gw'ensiri guleetebwa akawuka ka **pulasimoodiyamu** akasaasaanyizibwa ng'ensiri erumye omuntu.
Google Translate	Omusujja guva ku buwuka obuyitibwa plasmodium parasites obuyisibwa nga buyita mu nsiri.
NLLB-1.3 w/c LoRA	Omusujja gw'ensiri guva ku biwuka ebiyitibwa Plasmodium ebisiigibwa ensiri nga biyita mu kulumwa ensiri.
Finetune Mistral	Omusa gw'ensiri guggibwa mu nkooda z'ensiri eziyitibwa Picia.

Table 5: Comparison of Reference and Hypothesis Sentences with MQM Error Highlighting

- **Untranslated Words (blue):** Plasmodium remained untranslated in NLLB-1.3 outputs, demonstrating inconsistent handling of medical terminology.

- **Mistranslations (red):** Mistral failed by generating "Picia", which does not correspond to any meaningful term.

6.3 Semantic Preservation and Clinical Accuracy

A key measure of translation success is whether critical medical meaning is preserved. The translations produced by **NLLB-1.3 w/ LoRA** largely maintain semantic fidelity, despite minor paraphrasing differences. The highlighted sections in green within Table 5 indicate phrases that use different wording while conveying the correct meaning.

However, the reliance on automatic metrics alone may not fully capture semantic nuances. While BLEU and METEOR scores improved, manual evaluations revealed that terminology consistency and long-form description handling were superior in decoder-based architectures such as Mistral-7B.

6.4 Computational Efficiency and Adaptability

One of the most notable advantages observed in this study is the efficiency of LoRA fine-tuning. Compared to full fine-tuning, LoRA achieved nearly equivalent performance at a fraction of the computational cost. This is particularly valuable for real-world deployment in low-resource environments, where access to extensive compute resources is limited.

Decoder-based architectures such as Mistral-7B demonstrated better context handling for complex medical sentences. However, their performance was highly dependent on proper fine-tuning. The direct fine-tuned Mistral model exhibited critical mistranslations (e.g., "Picia"), highlighting the need for further refinements.

6.5 Implications for Medical Translation Systems

The results suggest that hybrid approaches—combining LoRA for efficient adaptation with human-in-the-loop validation—are promising for improving machine translation in medical contexts. Expert evaluations confirmed that:

- NLLB-1.3 LoRA provides a balance of effi-

ciency and accuracy, making it suitable for constrained environments.

- Mistral-7B has potential for high-quality translation, but requires careful fine-tuning to mitigate hallucinations.
- Domain-specific terminology alignment remains a challenge, necessitating continuous adaptation with domain-adapted datasets.

6.6 Future Directions

Moving forward, key areas for improvement include:

- Expanding training datasets with more diverse medical texts to enhance domain adaptation.
- Implementing reinforcement learning from human feedback (RLHF) to correct systematic translation errors.
- Deploying these models in clinical pilot studies to assess real-world applicability.

In conclusion, LoRA-based fine-tuning demonstrated a significant advancement in low-resource medical translation, particularly in its ability to achieve high performance while minimizing computational costs. Future research will focus on further enhancing clinical applicability through expert-in-the-loop methodologies and robust evaluation pipelines.

Limitations

This study faced three main limitations. **First,** we could not conduct a full **MQM evaluation** due to high costs, relying instead on automated metrics and limited clinician assessments. **Second,** our *evaluation dataset was constrained*, as obtaining expert validation was expensive and time-consuming, limiting the scope of manual assessments. **Third,** while *LoRA fine-tuning* improved efficiency, further analysis is needed to enhance terminology consistency and contextual accuracy. Future work should integrate *MQM assessments, clinician-involved evaluations, and human-in-the-loop feedback* to refine medical translations.

Acknowledgments

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 62–69.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Emma Croxford, Yanjun Gao, Brian Patterson, Daniel To, Samuel Tesch, Dmitriy Dligach, Anoop Mayampurath, Matthew M Churpek, and Majid Afshar. 2024. Development of a human evaluation framework and correlation with automated metrics for natural language generation of medical diagnoses. *medRxiv*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Elaine C Khoong and Jorge A Rodriguez. 2022. A research agenda for using machine translation in clinical medicine. *Journal of General Internal Medicine*, 37(5):1275–1277.

Annika Kreienbrinck, Saskia Hanft-Robert, and Mike Mösko. 2024. Usability of technological tools to overcome language barriers in health care: a scoping review protocol. *BMJ open*, 14(3):e079814.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, Jindrich Helcl, and Ankit Srivastava. 2017. Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybernetics and Information Technologies*, 17(2):28–43.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Fine-tuning large language models for adaptive machine translation. *arXiv preprint arXiv:2312.12740*.

Joyce Nakatumba-Nabende, Claire Babirye, Peter Nabende, Jeremy Francis Tusubira, Jonathan Mukiibi, Eric Peter Wairagala, Chodrine Mutebi, Tobius Saul Bateesa, Alvin Nahabwe, Hewitt Tusiime, et al. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: Experiences and lessons. *Applied AI Letters*, 5(2):e92.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Long Phan, Tai Dang, Hieu Tran, Trieu H Trinh, Vy Phan, Lam D Chau, and Minh-Thang Luong. 2022. Enriching biomedical knowledge for low-resource language through large-scale translation. *arXiv preprint arXiv:2210.05598*.

Aadil Rasheed, Florian Borchert, Lasse Kohlmeyer, Richard Henkenjohann, and Matthieu-P Schapranow. 2021. A comparison of concept embeddings for german clinical corpora. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2314–2321. IEEE.

Miguel Rios. 2024. Instruction-tuned large language models for machine translation in the medical domain. *arXiv preprint arXiv:2408.16440*.

Konstantinos Skianis, Yann Briand, and Florent Desgrippes. 2020. Evaluation of machine translation methods applied to medical terminologies. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 59–69.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: training neural machine translation with semantic similarity. *arXiv preprint arXiv:1909.06694*.