Advancing Clinical Translation in Nepali through Fine-Tuned Multilingual Models

Benyamin Ahmadnia, Sumaiya Shaikh, Bibek Poudel, Shazan Ansar, and **Sahar Hooshmand**

Department of Computer Science
California State University, Dominguez Hills, Carson, USA
bahmadniayebosari@csudh.edu, sshaikh10@toromail.csudh.edu,
bpoudel1@toromail.csudh.edu, smohammed8@toromail.csudh.edu,
hooshmand@csudh.edu

Abstract

Low-resource Neural Machine Translation (NMT) remains a major challenge, particularly in high-stakes domains such as healthcare. This paper presents a domain-adapted pipeline for English–Nepali medical translation leveraging two state-of-the-art multilingual Large Language Models (LLMs): mBART and NLLB-200. A high-quality, domain-specific parallel corpus is curated, and both models are fine-tuned using PyTorch frameworks. Translation fidelity is assessed through a multi-metric evaluation strategy that combines BLEU, CHRF++, METEOR, BERTScore, COMET, and perplexity. Our experimental results show that NLLB-200 consistently outperforms mBART across surface-level and semantic metrics, achieving higher accuracy and lower hallucination rates in clinical settings. In addition, error profiling and ethical assessments are conducted to highlight challenges such as term omissions and cultural bias. This work underscores the viability of large-scale multilingual models in enhancing medical translation for low-resource languages and proposes actionable paths toward safer and more equitable MT deployment in healthcare.

1 Introduction

Neural Machine Translation (NMT) has brought significant advancements to the field of MT, offering more fluent and accurate translations than traditional statistical methods (Goyle et al., 2023). However, the success of NMT systems heavily relies on the availability of large-scale parallel corpora, which remain scarce for many of the world's languages. Nepali, a low-resource language, exemplifies this challenge, particularly in specialized domains such as healthcare (Ranathunga et al., 2021; Elmadani and Buys, 2024).

Medical translation introduces unique complexities: terminology is highly domain-specific, context-sensitive, and error-prone. Inaccuracies in translating clinical terms can have severe consequences, including misdiagnosis and improper treatment. However, current NMT systems face challenges in domain-specific text translation within low-resource settings because they lack adequate high-quality medical parallel corpora and diverse linguistic training data (Ranathunga et al., 2021; Wang et al., 2021).

Recent advances in Large Language Models (LLMs), especially multilingual transformer-based models, have shown great promise in improving translation quality for low-resource languages and specialized domains. In this paper, we investigate the use of LLMs to enhance English–Nepali NMT in the medical domain. Building on architectures such as mBART and NLLB-200, we construct a domain-adapted pipeline using PyTorch and evaluate translation quality in both general-purpose and domain-specific metrics.

Our contributions are threefold:

- A Nepali–English parallel corpus tailored to the medical domain is curated by combining diverse sources with domain-specific relevance.
- The performance of mBART and NLLB-200 is fine-tuned and compared using a unified experimental framework that incorporates metrics, capturing both lexical fidelity and semantic preservation.
- A detailed analysis is conducted on translation errors, model hallucinations, and domain-specific term accuracy, with attention to ethical concerns such as safety, fairness, and deployment constraints.

By demonstrating the effectiveness of multilingual LLMs in translating low-resource and high-risk content, this work contributes a step toward

Proceedings of Recent Advances in Natural Language Processing,pages 48–56
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-006

safer and more inclusive NMT systems for global health applications.

2 Related Work

NMT for low-resource languages, especially in specialized domains such as medicine, presents enduring challenges due to the limited availability of parallel corpora. Prior work has explored techniques such as transfer learning (Zoph et al., 2016), back-translation (Edunov et al., 2018), round-trip (Ahmadnia and Dorr, 2019; Ahmadnia et al., 2019, 2018), and multilingual training (Ranathunga et al., 2021; Elmadani and Buys, 2024) to augment low-resource datasets. Although these methods offer improvements, they often fall short in domain-specific scenarios where precise terminology and contextual nuance are critical.

Recent advances in multilingual LLMs, such as mBART (Liu et al., 2020) and NLLB-200 (Team et al., 2022), trained on massive multilingual corpora, have demonstrated notable gains in low-resource translation performance. However, most existing studies apply these models in general-domain contexts, with limited investigation of domain adaptation for high-stakes fields such as medicine.

In the medical domain, research has emphasized the importance of adapting general models using small in-domain corpora and leveraging domain-specific ontologies (Ranathunga et al., 2021). However, the direct application of LLM to Nepali medical translation has not been explored sufficiently. Our work fills this gap by conducting a focused comparative evaluation of mBART and NLLB-200 in the context of English–Nepali medical translation, specifically measuring their effectiveness using domain-relevant metrics.

Benchmarking efforts for Nepali remain sparse. Existing work has highlighted the lack of comprehensive evaluation pipelines tailored to low-resource languages and has called for specialized metric design and robust model comparisons (Wang et al., 2021). In contrast, our study provides a unified PyTorch-based framework with an extensive suite of evaluation metrics—including both surface-level (BLEU and CHRF++) and semantic metrics (BERTScore, COMET, and perplexity)—to better capture domain-specific translation quality.

Furthermore, while domain adaptation through fine-tuning in small corpora has been shown to be effective, our results demonstrate that LLMs like NLLB-200 can significantly outperform smaller multilingual baselines when fine-tuned even with limited medical domain data. Unlike previous work, we perform detailed error analysis and term-level confusion profiling to understand hallucination patterns, omission rates, and reliability in clinical contexts.

Finally, studies focusing on South Asian languages such as Nepali have identified unique morphological and syntactic challenges that complicate machine translation (Guzmán et al., 2019). Our contribution extends this line of inquiry by presenting model-specific evaluations of how these challenges manifest in the medical domain and how modern multilingual LLMs mitigate or fail under these conditions.

3 Mathematical Background

This section outlines the core mathematical principles relevant to our implementation of mBART and NLLB-200 for English–Nepali medical translation, focusing on the transformer architecture and domain adaptation objectives that underpin our system.

3.1 Transformer Architecture

Both mBART and NLLB-200 are built upon the Transformer model (Vaswani et al., 2023), which replaces recurrence with multi-head self-attention to efficiently model long-range dependencies in sequences. These models follow an encoder–decoder structure where each layer uses scaled dot-product attention to compute contextual representations.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (1)$$

where Q, K, and V are the query, key, and value matrices derived from input sequences, and d_k is the dimensionality of the key vectors.

Multi-head attention allows the model to jointly attend to information from different representation subspaces:

$$\begin{aligned} &\text{MultiHead}(Q, K, V) \\ &\quad = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O\text{head}_i \\ &\qquad = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2) \end{aligned}$$

This formulation enables a better capture of syntactic and semantic features, especially important in morphologically rich languages such as Nepali.

3.2 Domain Adaptation Objective

Domain adaptation is central to our work. We fine-tune general-purpose LLMs on a small in-domain corpus to specialize them for medical translation. The adaptation process involves minimizing the negative log-likelihood of the target medical text conditioned on the source:

$$\theta^* = \arg\min_{\theta} \sum_{(x,y)\in\mathcal{D}_t} -\log P_\theta(y|x) \qquad (3)$$

where $\mathcal{D}_t$ is the domain-specific dataset, and θ are the model parameters. This process enables the model to better internalize the terminology, syntax, and semantics specific to the medical domain.

To support low-resource adaptation, we also consider the trade-off between task-specific loss and domain divergence using a weighted composite objective:

$$\mathcal{L}_{\text{adapt}} = \lambda\mathcal{L}_{\text{task}} + (1-\lambda)\mathcal{L}_{\text{domain}} \qquad (4)$$

This framework allows controlled adaptation by balancing translation accuracy with domain generalization[1].

3.3 Positional Encoding

Since the Transformer lacks recurrence, positional encodings are added to inject information about token order. These encodings follow sinusoidal patterns to allow the model to learn relative and absolute positions:

$$\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \qquad (5)$$

$$\text{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \qquad (6)$$

Although positional encoding is standard, its effectiveness in Nepali–English translation, where syntactic structure differs significantly between languages, remains critical in maintaining translation fidelity.

4 Methodology

This section describes the complete process of developing, fine-tuning, and evaluating multilingual LLMs (mBART and NLLB-200) for English–Nepali medical translation. The pipeline comprises data preparation, model configuration, training setup, and metric-based evaluation.

4.1 Data Collection and Preprocessing

Nepali was selected as the low-resource target language due to its limited parallel corpora, particularly in the medical domain. We collected bilingual (English–Nepali) data from publicly available sources such as “NepaliHealth” and curated it to ensure linguistic diversity, medical relevance, and quality. The dataset included FAQs, news articles, and instructional medical texts.

The raw data underwent a rigorous cleaning process that included:

- Removal of special characters and non-linguistic artifacts,
- Normalization of Nepali Unicode encoding,
- Formatting into JSONL structure: { “en”: “...”, “ne”: “...” }

To address data scarcity, additional monolingual and synthetic parallel data was generated using back-translation through tools such as Google Translate and Gemini, following the method proposed by Yang et al. (2024).

4.2 Model Selection and Configuration

We selected two multilingual pre-trained models for experimentation:

- **NLLB-200**: Initially, the 600M distilled variant was tested for development purposes[2].
- **mBART**: A widely adopted multilingual transformer with proven effectiveness in domain adaptation tasks.

Both models were initialized using pre-trained weights from “Hugging Face”. Tokenization was handled using `AutoTokenizer` and `MBartTokenizer`, with explicit specification of language codes. No structural modifications were made to the model architectures.

4.3 Fine-Tuning Setup

Due to limitations in quantization support, we used full-parameter fine-tuning for both models. Training was carried out on A100 GPUs using PyTorch. Key hyperparameters included:

To enhance efficiency: 1) Mixed-precision (FP16) training was employed, 2) Dynamic

[1] In our implementation, we set $\lambda = 1$ and optimize only the task loss. This equation is presented conceptually to describe potential future extensions.

[2] All final training and evaluation results presented in this paper are based on the full 3.3B-parameter version of NLLB-200 to ensure maximum translation quality.

Hyperparameter	NLLB-200	mBART
Batch Size	16	4
Gradient Accumulatio	1	4
Epochs	20	20
Weight Decay	0.01	0.01
Learning Rate	5e-5	5e-5

Table 1: Training hyperparameters for NLLB-200 and mBART.

padding was managed using Hugging Face's `DataCollatorForSeq2Seq`, and 3) Evaluation was conducted at the end of each epoch.

4.4 Evaluation Metrics

We adopted a diverse set of evaluation metrics to capture both surface overlap and semantic adequacy, especially relevant in clinical translation:

- **BLEU**: Standard lexical overlap metric; included primarily for baseline comparison.
- **CHRF++**: Character-level metric better suited to morphologically rich languages such as Nepali.
- **TER**: Measures the number of edits required to match reference translations.
- **METEOR**: Accounts for synonymy, morphology, and paraphrasing; useful for languages with flexible word order.
- **BERTScore**: Evaluates contextual semantic similarity using XLM-RoBERTa.
- **COMET**: A learned evaluation metric designed to better correlate with human judgments; includes source, hypothesis, and reference.
- **Perplexity**: Used as a fluency and hallucination proxy; Lower perplexity indicates greater confidence and coherence.

Each metric played a distinct role in assessing the suitability of translations for sensitive clinical applications, where literal equivalence is insufficient.

5 Experimental Framework

This section outlines the experimental setup used to evaluate the effectiveness of mBART and NLLB-200 in translating English–Nepali medical texts. We describe the datasets, model initialization, domain-specific training strategy, and runtime environment.

5.1 Role of LLMs in the System

Multilingual LLMs are central to this system. Their contextual representation power, multilingual pre-training, and domain adaptation capacity make them ideal for low-resource, high-risk applications such as medical NMT.

Unlike traditional statistical or phrase-based systems, these models can generalize across domains due to their pre-training on diverse corpora. Fine-tuning allows for specialization without architectural changes, enabling improved translation fluency and terminology consistency across clinical documents.

5.2 Domain-Specific Data Integration

Both mBART and NLLB-200 were initially trained for general-purpose multilingual translation and do not include medical-specific supervision. To bridge this gap, we fine-tuned each model using curated bilingual English–Nepali medical texts. These include:

- Health education materials,
- Doctor–patient dialogue samples,
- Domain-specific FAQs from "NepaliHealth."

To enhance lexical and syntactic robustness, we also incorporated back-translated data using Google Translate and Gemini, following strategies proposed by Yang et al. (2024).

5.3 Dataset Overview

The training dataset consisted of approximately 25K sentence pairs in the medical domain, 10K synthetic pairs from back-translation, and a validation/test set of 2,000 pairs manually reviewed. All data were formatted in "JSONL" with fields `en` and `ne`. Encoding inconsistencies were corrected by normalization. Sentence distributions were inspected to ensure that no subdomain was overrepresented.

5.4 LLM Configuration

The models were instantiated using the Hugging Face Transformers library: 1) mBART that initialized with `facebook/mbart-large-50`, and 2) NLLB-200 that initially tested with the 600M variant; later scaled to the 3.3B model for final evaluations. Language codes were explicitly specified. No modifications were made to the architectures or tokenizers beyond model-specific preprocessing.

5.5 Training Environment

Training was performed on NVIDIA A100 GPUs (80GB memory) using mixed-precision FP16. Dynamic padding was enabled via Hugging Face's `DataCollatorForSeq2Seq`.

5.6 Evaluation Interface

BLEU and CHRF++ were calculated using `sacrebleu`. BERTScore used XLM-RoBERTa, and COMET used multilingual pre-trained weights. Each model output was evaluated at the sentence level and aggregated across the test set. In addition, we implemented confusion analysis to examine the reliability of the model in domain-specific terms and in hallucination cases. The metrics were cross-referenced with semantic metrics (COMET and BERTScore) to detect safe but inaccurate translations.

6 Results Analysis and Discussion

We evaluate mBART and NLLB-200 in English–Nepali medical translation using surface-level and semantic metrics. This section presents quantitative results, sentence-level analyses, and model-specific error profiles, with a focus on translation reliability in clinical contexts.

6.1 Quantitative Evaluation

All results in this section are reported using the final 3.3B-parameter version of NLLB-200, unless otherwise noted. Table 2 presents a comprehensive comparison between NLLB-200 and mBART on several metrics. NLLB-200 outperforms mBART on all fronts, with significant margins in semantic alignment (COMET and BERTScore) and fluency (perplexity).

These results confirm NLLB-200's stronger preservation of both lexical content and semantic meaning under domain-specific constraints. In particular, lower perplexity indicates better model calibration and reduced hallucination.

Metric	**NLLB-200**	**mBART**
BLEU	65.70	60.12
CHRF++	82.81	78.34
METEOR	0.780	0.742
BERTScore (F1)	0.968	0.956
COMET	0.830	0.800
Perplexity	1.84	3.25

Table 2: Comparison of translation performance across evaluation metrics.

Human references and prediction generated by the Nepali translation model are provided in Figure 1. These examples qualitatively evaluate the translation performances of a given NLLB model in the medical domain. From what can be seen, the model usually only captures the general meaning, with somewhat minor rephrasings. These variations, which are often semantically acceptable, have been put into perspective where domain-specific nuances or idiomatic expressions may still be improved.

6.2 Lexical Patterns

Figure 2 illustrates the BLEU score distribution across the test set. NLLB-200 shows a denser concentration in the high-BLEU region, indicating fewer low-quality outputs and better handling of lexical irregularities.

In contrast, mBART's BLEU distribution shows a longer tail, suggesting higher lexical variance and occasional drops in fidelity, particularly with uncommon medical expressions.

6.3 Semantic Fidelity

Figure 3 compares sentence-level COMET and BLEU scores. NLLB-200 maintains a higher COMET even for sentences with moderate BLEU, demonstrating robust paraphrastic fluency and semantic retention.

This supports the idea that exact lexical overlap (as measured by BLEU) can misrepresent true translation quality in the medical domain, where synonyms and reformulations are common.

6.4 Medical Term Accuracy

We performed a confusion matrix analysis to classify translations at the term level. Figure 4 shows that NLLB-200 consistently exhibits fewer false positives, especially in symptom and diagnostic terms.

Source (EN)	Reference (NE)	Prediction (NE) [NLLB-200]
What does a stuffy nose mean?	भरिएको नाकलाई के बुझाउँछ?	भरिएको नाकको अर्थ के हो?
Is it safe to consume caffeine during pregnancy?	सूचना प्रविधि गर्भावस्थामा क्याफिन उपभोग गर्न सुरक्षित छ?	गर्भावस्थामा क्याफिन सेवन गर्न सुरक्षित बनाउनुहोस्?
I have tingling in my toes and the tip of my tongue. Can you help?	मेरो खुट्टाको औंला र जिब्रोको टुप्पोमा झमझम छ। के तपाईं मद्दत गर्न सक्नुहुन्छ?	मेरो खुट्टाको औंला र जिब्रोको टुप्पोमा झनझन छ। के तपाईं मद्दत गर्न सक्नुहुन्छ?

Figure 1: Examples of English medical source sentences with human reference translations and NLLB-200 outputs. These examples illustrate NLLB-200's semantic fidelity and fluency in domain-specific contexts, complementing the aggregate metric-based evaluation.

mBART occasionally inserts plausible but incorrect terms (e.g., mistranslation of "anesthesia" as "dizziness"), while NLLB-200 is more conservative, prioritizing precision even at the cost of minor recall loss.

6.5 Fluency and Hallucination

Perplexity scores highlight notable differences in fluency confidence. NLLB-200 achieves a lower average perplexity (1.84 vs 3.25), indicating higher consistency and fewer hallucinated or malformed outputs. The qualitative review confirms this: mBART more frequently mixes source language tokens or drops critical medical terms.

In addition, hallucinated words in the mBART output often hybridize English and Nepali morphemes, suggesting inadequate domain grounding. NLLB-200 occasionally omits low-frequency terms, but rarely generates novel or fictitious phrases, aligning with safer deployment goals in clinical translation.

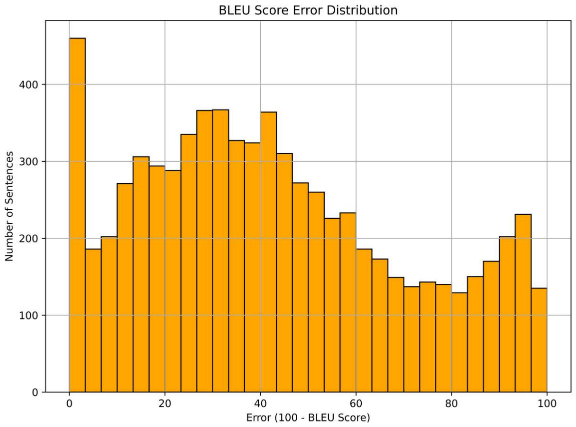

Figure 2: Distribution of sentence-level BLEU scores across the test set, showing that NLLB-200 produces consistently high-quality outputs with fewer low-BLEU outliers compared to mBART.

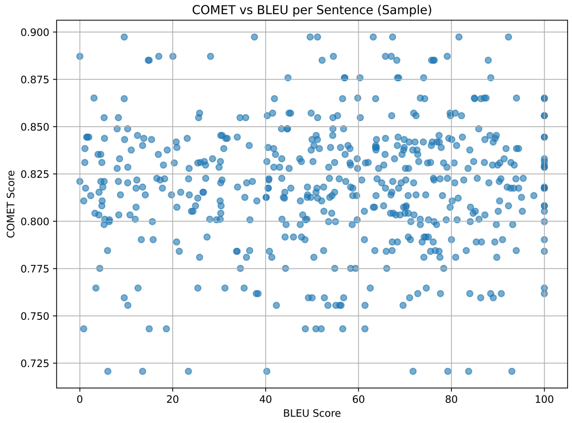

Figure 3: Sentence-level comparison of COMET and BLEU scores for NLLB-200, highlighting cases where high semantic adequacy (COMET) is achieved even when surface-level lexical overlap (BLEU) is moderate.

7 Ethical Considerations

Deploying MT in clinical settings introduces serious ethical and practical responsibilities. In this section, we address potential risks related to translation safety, hallucination, data privacy, computational cost, and fairness.

7.1 Translation Risk

In the medical domain, minor translation errors can lead to critical misinterpretations. Standard metrics such as BLEU and METEOR assign equal weight to all tokens, but in clinical texts, specific terms—e.g., medications, procedures, or diagnoses—carry disproportionately high importance. For instance, mistranslating "complete resection" as "c-section" in a surgical context reverses clinical

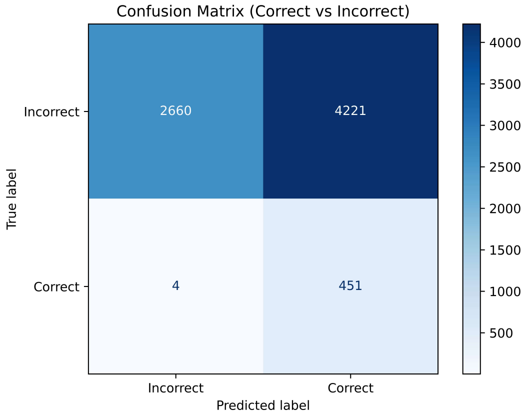

Figure 4: Confusion matrix of domain-specific medical terms translated by mBART and NLLB-200, indicating that NLLB-200 achieves higher accuracy with fewer false substitutions in critical terminology.

intent, yet results in minimal BLEU penalty (Shor et al., 2023).

Our error analysis (see Figures 2, 4) confirms that high lexical scores do not guarantee safe translations. Some model outputs passed surface-level metrics while introducing clinically unacceptable substitutions or omissions. Although NLLB-200 exhibited better reliability, both models showed vulnerability to hallucination—fluent but fabricated terms—which are especially dangerous in healthcare.

To mitigate this, we advocate for human-in-the-loop workflows with uncertainty-driven flagging, where low-confidence segments are highlighted for clinician review. As Gaona et al. (2023) note, even trained post-editors may overlook errors in fluent MT outputs. Interactive review interfaces and term-specific alerts are essential safeguards.

7.2 Computational Cost

Fine-tuning NLLB-200 (3.3B parameters) in medical texts required 18 hours of A100 GPU compute. Such resource demands may be prohibitive in low-resource clinical settings, where infrastructure is limited. In addition, inference through large LLMs may be infeasible in rural clinics lacking stable electricity or Internet.

Compression techniques such as 8-bit quantization and knowledge distillation offer potential solutions but require careful calibration to preserve translation fidelity. Beyond the feasibility of deployment, there are ethical concerns related to environmental impact. As Bender et al. (2021) argue, large-scale model training incurs significant carbon and financial costs that disproportionately burden under-resourced communities. We support transparent reporting of resource use and the development of greener MT workflows, including training on renewable-powered infrastructure and reusing fine-tuned checkpoints where possible.

7.3 Bias and Fairness

Language models can propagate or amplify biases embedded in training data. In clinical NMT, such biases can result in demographic misrepresentations or dialectal exclusion. For example, certain dialects or regional expressions in Nepali may be poorly translated, disadvantaging minority speakers. Although we did not conduct an exhaustive bias audit, future work should include demographic fairness evaluation, dialect sensitivity analysis, and gender representation audits. Bias-aware training strategies and lexicon-level interventions are recommended for deployment-ready systems. We also emphasize that privacy is non-negotiable in clinical Natural Language Processing (NLP). Any cloud-based translation pipeline must ensure end-to-end encryption, no logging of sensitive information, and compliance with relevant healthcare data regulations.

7.4 Human-Centered MT Deployment

Any clinical translation system must be embedded in a socio-technical feedback loop. Inspired by Gaona et al. (2023), we suggest systems that: 1) Surface low-confidence tokens for human correction, 2) Provide term-level explanations or alignments, and 3) Allow feedback from domain experts to be integrated into future model updates. Participatory design is critical: translators, clinicians, and local health workers should be involved in testing and refinement, particularly for underrepresented linguistic communities.

8 Conclusions and Future Work

This paper presented a domain-adapted NMT framework for English–Nepali medical texts, using two state-of-the-art multilingual language models: mBART and NLLB-200. Our experiments show that NLLB-200 significantly outperforms mBART on multiple evaluation metrics, achieving higher scores on BLEU, CHRF++, METEOR, COMET, and BERTScore, along with significantly lower perplexity. These results suggest that NLLB-200 is better suited for clinical translation tasks in low-resource settings, offering improved preservation of medical terminology and contextual fluency.

Our detailed error analysis revealed that traditional surface metrics such as BLEU and METEOR often do not capture clinically important semantic differences. In contrast, COMET and BERTScore provided better correlation with human judgments, though even these metrics occasionally overlooked critical omissions or paraphrastic errors. We observed that mBART tends to hallucinate or substitute medical terms, while NLLB-200 exhibits more conservative translation behavior, with fewer false positives but occasional under-generation of rare terminology. These findings highlight a key trade-off in clinical NMT: balancing lexical precision with semantic adequacy under domain-specific constraints. Although large-scale models such as NLLB-200 show strong promise, their deployment in real-world healthcare contexts remains nontrivial. The computational demands of training and inference, as well as the risks of opaque model behavior, raise important ethical, logistical, and usability concerns. In practical deployment scenarios, clinicians must be able to trust and interpret the model's output. We advocate for the integration of uncertainty estimation, confidence scoring, and human-in-the-loop editing tools into clinical MT interfaces. Visualization of model attention or translation risk zones could further improve explainability and safety.

Future work will explore several promising directions. First, our objective is to compress NLLB-200 via quantization and distillation to enable offline or on-device use in rural clinics. Second, we plan to expand this pipeline to additional South Asian languages and medical subdomains, using the multilingual pre-training of NLLB. Third, we envision building interactive translation systems that incorporate clinician feedback in real time, using corrective supervision to refine the model's behavior dynamically. Finally, we will investigate clinically grounded evaluation frameworks that better account for terminology importance and contextual integrity, including task-specific metrics such as Clinical-BERTScore.

Acknowledgment

The authors thank the Department of Computer Science and the College of Natural and Behavioral Sciences at CSUDH for their support.

References

Benyamin Ahmadnia and Bonnie J. Dorr. 2019. Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, 9(1):268–278.

Benyamin Ahmadnia, Gholamreza Haffari, and Javier Serrano. 2018. Statistical machine translation for bilingually low-resource scenarios: A round-tripping approach. In *Proceedings of the 2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 261–265.

Benyamin Ahmadnia, Gholamreza Haffari, and Javier Serrano. 2019. Round-trip training approach for bilingually low-resource statistical machine translation systems. *International Journal of Artificial Intelligence*, 17(1):167–185.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Association for Computing Machinery*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Khalid N. Elmadani and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158.

Miguel Angel Rios Gaona, Raluca-Maria Chereji, Alina Secara, and Dragos Ciobanu. 2023. Quality analysis of multilingual neural machine translation systems and reference test translations for the English-Romanian language pair in the medical domain. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 355–364.

Vakul Goyle, Parvathy Krishnaswamy, Kannan Girija Ravikumar, Utsa Chattopadhyay, and Kartikay Goyle. 2023. Neural machine translation for low resource languages.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020.

mBART: Multilingual denoising pre-training for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 125–137.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey.

Joel Shor, Ruyue Agnes Bi, Subhashini Venugopalan, Steven Ibara, Roman Goldenberg, and Ehud Rivlin. 2023. Clinical BERTScore: An improved measure of automatic speech recognition performance in clinical settings. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 1–7.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643.

Hao Yang, Min Zhang, and Jiaxin Guo. 2024. From scarcity to sufficiency: Machine translation techniques for low-resource llms enhancement. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 36–41.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.