# LexiMCH: A Bilingual Medical Knowledge Lexicon for Maternal and Child Healthcare in Low-Resource Languages and Healthcare Environments

**Aziza Umer Yibrie[1], Seid Muhie Yimam[2], Katrin Schöning-Stierand[2], Kaleab Anteneh[3], Rebecca Ashagire[4], Robera Habtamu[4], Rahel Bekele[4], Martin Semmann[2]**
[1]Bahir Dar University, Ethiopia
[2]Hub of Computing & Data Science, University of Hamburg, Germany
[3]BITS College, Ethiopia
[4]Addis Ababa University, Ethiopia

## Abstract

Maternal and child healthcare (MCH) in low-resource contexts faces persistent challenges due to linguistic and cultural barriers to accessing medical information. To address this, we develop a multilingual terminology resource focusing on English and Amharic, using a combination of machine translation, large language models (LLMs), and expert-in-the-loop validation. In this work, we evaluate a subset of 90 terms and definitions across multiple translation models, including Google Translate, *NLLB-200*, *M2M100*, and several LLM variants (*GPT*, *LLaMA*, *Gemma*, *DeepSeek*, Gemini, and *Mistral*). We use BLEU, chrF, and ROUGE-L metrics to assess translation quality for both terms and definitions. Preliminary results indicate variable performance across models, with *DeepSeek-R1* achieving the highest BLEU scores (0.916 for definitions and 0.985 for terms) and LLM-assisted translations generally performing better on definitions than on terms. Ongoing work is extending the evaluation to the full dataset and further refining translation pipelines to produce a comprehensive, open-access, AI-ready resource for maternal and child healthcare in low-resource languages.

## 1 Introduction

Maternal and child healthcare (MCH) continues to face critical global disparities, particularly in low-resource regions where access to accurate and linguistically appropriate medical information remains limited. In 2023, an estimated 260k women died from preventable complications during pregnancy and delivery, with sub-Saharan Africa accounting for approximately 70% (182k) and southern Asia for 17% (43k) of these deaths (Organization, 2025). Beyond clinical constraints, these figures reflect a profound informational and linguistic gap: most medical guidance and digital health tools are produced in English or other dominant languages, leaving communities dependent on oral transmission or poorly localized materials (Tiruneh et al., 2022).

Recent advances in natural language processing (NLP) and artificial intelligence (AI) offer transformative opportunities to improve healthcare delivery. However, the majority medical language resource terminologies, corpora, and pre-trained language models are based on English and built on data from high-resource Western contexts (Névéol et al., 2018; Okafor, 2025). This imbalance not only reduces model performance in non-English settings, but also perpetuates epistemic exclusion of local knowledge systems and patient experiences. Decolonizing AI toolchains thus requires the inclusion of linguistically diverse data and culturally grounded methodologies to ensure equitable access to health information (Mohamed et al., 2020).

In Ethiopia and similar countries, such communication barriers between mothers and healthcare professionals contribute significantly to maternal morbidity. The absence of standardized Amharic medical terminology hampers patient understanding, complicates documentation, and limits the integration of local data into national and global health systems (Tareke et al., 2024). Moreover, current AI-driven health applications are predominantly trained on English datasets and optimized for Western biomedical contexts, making them linguistically inaccessible and culturally incongruent for African settings (Sinha, 2025).

This study explores methods for curating, translating, and validating a standardized bilingual (English–Amharic) terminology resource for MCH. We integrate machine translation, LLMs, and expert-led expert-in-the-loop validation to develop linguistically accurate and culturally grounded medical terminologies for low-resource settings.

The workflow combines automated extraction of English MCH terms and definitions from authoritative public-health and clinical sources with multilingual Machine Translation(MT) and LLM-

based translation pipelines, followed by iterative expert review by Ethiopian clinicians, public health professionals, and linguists. This participatory process ensures semantic precision, contextual relevance, and cultural appropriateness of the resulting Amharic terminology.

The finalized bilingual resource is organized in structured, machine-readable formats that are compatible with APIs and downstream NLP pipelines, enabling integration into digital health applications. By openly releasing the dataset, code, and workflow, this work supports the development of inclusive and equitable AI infrastructures for healthcare. Although this study focuses on English–Amharic maternal and child health terminology, the methodology is designed to scale to additional African languages and to support future multimodal health technologies, including voice, image, and icon-based maternal health tools.

**Contributions** This paper makes three main contributions: (1) We curate and standardize a focused maternal and child health terminology resource comprising 747 English terms and definitions, addressing fragmentation across existing public-health sources; (2) We present a comparative evaluation of machine translation and large language model approaches for translating biomedical terminology into Amharic, a low-resource African language, using both automated metrics and expert validation; and (3) We propose and implement a participatory workflow that integrates local clinical and linguistic expertise throughout dataset creation, resulting in an open, AI-ready bilingual resource for maternal and child healthcare.

## 2 Related work

Research related to this study spans four intersecting areas: biomedical terminology resources, machine translation for low-resource languages, the use of large language models in medical translation, and expert-in-the-loop approaches for healthcare NLP. While each area has seen substantial progress, its intersection remains underexplored in the context of African languages and maternal and child health.

**Biomedical Terminology Resources for Healthcare** Standardized biomedical terminologies are essential for clinical documentation, patient education, and interoperable health information systems. Resources such as SNOMED CT[1], UMLS[2], and ICD[3] offer broad coverage of clinical concepts but are primarily developed in English and optimized for high-resource healthcare settings (Bodenreider, 2004). As a result, support for African languages remains limited, particularly for patient-facing maternal and child health content. These gaps affect both healthcare communication and NLP development. Prior studies show that insufficient biomedical terminology in low-resource languages hinders clinical interaction and constrains healthcare NLP systems (Névéol et al., 2018; Wang et al., 2018). In sub-Saharan Africa, including Ethiopia, language discordance between patients and providers contributes to communication breakdowns and reduced quality of care (Teshale et al., 2025; Olani et al., 2023). Existing global health glossaries are also poorly aligned, leading to fragmented coverage.

**Machine Translation for Low-Resource Languages in Healthcare** Machine translation has long been explored as a means of improving access to healthcare information across languages. Early medical MT systems relied on rule-based and statistical methods, which proved difficult to adapt to morphologically rich and low-resource languages. While neural machine translation has substantially improved translation quality, performance remains uneven for languages with limited parallel data, including Amharic. Recent multilingual MT models, such as *M2M100* (Fan et al., 2021) and No Language Left Behind (*NLLB-200*) (Costa-jussà et al., 2024), leverage large-scale multilingual training to improve cross-lingual transfer for low-resource languages. Despite these advances, biomedical terminology remains challenging due to domain mismatch and the scarcity of specialized medical terms in training corpora (Neubig and Hu, 2018; Wang et al., 2018). In healthcare settings, semantic precision and safety are critical, and machine translation remains challenging. Scoping reviews show that while modern MT systems are increasingly used to lower language barriers, accuracy limitations make them unsuitable for direct use in patient interactions without additional safeguards and human oversight (Merx et al., 2024).

[1] https://www.nlm.nih.gov/research/umls/Snomed/snomed_browsers.html

[2] https://www.nlm.nih.gov/research/umls/index.html

[3] https://icd.who.int/en/

**Large Language Models for Medical and Terminology Translation** Large language models have recently demonstrated strong performance across a wide range of NLP tasks, including translation, summarization, and domain adaptation. Instruction-tuned models such as GPT-4, LLaMA, and Mistral show improved contextual reasoning and fluency compared to traditional MT systems, even in low-resource settings (Touvron et al., 2023; OpenAI et al., 2024). In biomedical NLP, LLMs have been applied to clinical text generation, medical question answering, and terminology normalization, often outperforming task-specific baselines in semantic coherence (Zhou et al., 2025). However, prior work also highlights limitations of LLMs for medical translation. These models may introduce paraphrasing, omit clinically relevant details, or hallucinate content when domain constraints are not explicitly enforced (Busch et al., 2025). As a result, LLM outputs are increasingly used as candidate suggestions rather than final translations, particularly in high-stakes domains such as healthcare. For low-resource languages, empirical evidence suggests that LLMs can complement multilingual MT by improving readability and cultural appropriateness, but they cannot fully replace language-specific expertise (Shool et al., 2025). Systematic evaluations of LLM-assisted translation for African medical terminology remain scarce, especially for maternal and child health.

**Expert in the Loop Approaches in Healthcare NLP** These methodologies are widely recognized as essential for biomedical NLP, where correctness, interpretability, and accountability are paramount (Wang et al., 2021). In low-resource and cross-cultural settings, expert involvement is particularly important to ensure that translated content aligns with local clinical practice and sociocultural norms. Recent research has framed such participatory approaches within a broader decolonial AI perspective, arguing that global health technologies must move beyond extractive data practices and instead center local knowledge systems (Muldoon and Wu, 2023). Studies in African NLP consistently show that systems developed without local linguistic and domain expertise risk reinforcing epistemic inequities and producing tools that are unusable in practice (Adebara et al., 2025). Despite these insights, few published works integrate multilingual MT, LLMs, and structured expert validation into a unified workflow for building open biomedical resources. This gap is particularly evident for maternal and child healthcare, where patient-facing communication requires both technical accuracy and cultural sensitivity.

## 3 Methodology

To address the lack of language-specific maternal healthcare resources, we design a structured pipeline for extracting, processing, and translating maternal and child health terminology into Amharic. As shown in Figure 1, the pipeline integrates web-based data collection, automated term extraction and preprocessing, machine translation and large language model–based translation, and expert human review. This approach results in a culturally and linguistically grounded dataset that supports maternal and child healthcare communication and downstream NLP applications in Amharic.

**Extraction of Maternal Healthcare Terminologies** A Python-based web scraping pipeline was designed to dynamically extract maternal-related medical terminologies and their definitions from publicly available and reputable medical resources on the Web. The pipeline used libraries such as `BeautifulSoup` and `requests` to navigate and parse HTML content from targeted websites. The scraped data, consisting of terms and their corresponding definitions, were structured and stored in a JSON file for downstream processing. This approach ensured scalability and adaptability for expanding the dataset to include additional medical domains or languages.

**Data Processing and Cleaning** To facilitate integration with healthcare systems and further analysis, the JSON output from the web scraping pipeline was converted to structured formats suitable for broader use, such as CSV and Excel. A separate Python script was developed using the `pandas` library to parse the JSON data and transform it into tabular formats. This script included data cleaning steps to handle inconsistencies, such as missing definitions or duplicate entries, ensuring data quality and usability. The resulting files were designed to be compatible with healthcare data management systems and accessible to stakeholders, including healthcare professionals and AI developers, to support co-creative processes in the Ethiopian context.

**Translation to Amharic** To improve linguistic accessibility and ensure that maternal-health ter-

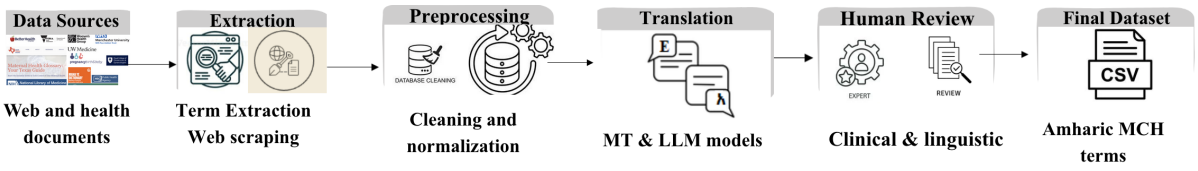


Figure 1: Overview of the pipeline for creating an Amharic maternal and child health terminology dataset. The process includes term extraction from web sources, preprocessing, translation using machine translation and large language models, and expert human review. One hundred terms are validated for evaluation, while the remaining entries are ongoing work.

minology is usable in Amharic, this stage applied a hybrid approach combining machine translation (MT) and large language models (LLMs). The goal is to generate accurate, culturally appropriate and semantically consistent Amharic equivalents for English terms and definitions.

**Machine Translation Baselines** For the initial English to Amharic translation of our preliminary dataset containing 747 maternal and child health terms and definitions, we employed *M2M100_418M*[4], *NLLB-200-distilled-600M*[5], and Google Translate. *M2M100_418M* enables translation from many to many without pivoting through English, maintaining semantic consistency for low-resource languages such as Amharic (Fan et al., 2021). *NLLB-200-distilled-600M* provides a balance of precision and efficiency, preserving strong lexical alignment and domain-specific terminology (Costa-jussà et al., 2024).

**LLM Assisted Translation** To further enhance translation quality, a suite of instruction-tuned large language models (LLMs) was employed, including *GPT-4o-mini, GPT-4.1-mini, LLaMA-3.3-70B*, *Mistral-Large*, *Gemma-3-27B, DeepSeek-R1, DeepSeek-R1-0528*, and *Gemini-2.0-Flash*. These models were chosen for their strengths in low-resource and domain-specific translation, particularly in biomedical and clinical contexts. The LLMs provided candidate translations and refinement suggestions, which were then iteratively evaluated against the English source. This approach ensured semantic fidelity, cultural appropriateness, and clinical accuracy, while enabling rapid parallelization of the English–Amharic maternal and child health terminology.

**Expert in the Loop Validation** For the evaluation, 90 preliminary Amharic translations from both machine translation and LLM outputs are reviewed by a multidisciplinary panel of Ethiopian clinicians, public health specialists, and linguists, while the remaining translations remain a work in progress. Each candidate is evaluated for semantic fidelity, ensuring preservation of clinical meaning and terminological accuracy; cultural relevance, confirming appropriateness for Ethiopian healthcare communication and patient contexts; and linguistic fluency, assessing grammatical correctness, clarity, and stylistic consistency. The experts chose the most appropriate translation or revised candidates to resolve ambiguities. All changes and reviewer comments were recorded in a CSV file, providing a clear audit trail of human decisions. As shown in Figure 2, expert-in-the-loop validation is conducted using a structured annotation interface.

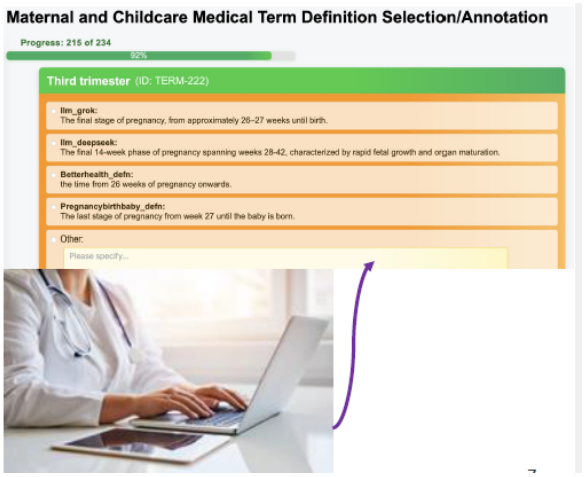


Figure 2: Expert-in-the-loop annotation interface for maternal and child health terminology validation. For each medical term definition, the interface presents multiple candidate Amharic translations generated by machine translation systems and large language models. Domain experts review these candidates and select the most appropriate translation or provide a revised alternative based on semantic fidelity, clinical accuracy, cultural relevance, and linguistic fluency. A progress indicator tracks annotation completion across the terminology set, supporting systematic coverage and quality assurance.

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

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

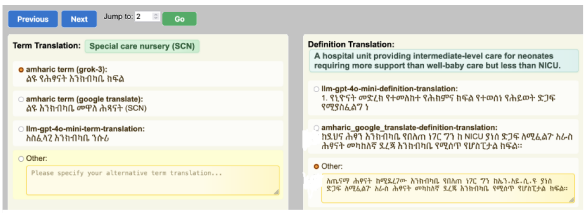


Figure 3: Expert-in-the-loop interface for term and definition translation validation. For each medical concept, the interface presents candidate Amharic translations of the medical term (left panel) and its definition (right panel), generated by machine translation systems and large language models. Experts select the most appropriate translation or provide a revised alternative, ensuring semantic fidelity, clinical accuracy, cultural relevance, and linguistic fluency. This side-by-side design supports consistent alignment between terminology and definitions during expert validation.

Figure 2, illustrates the annotation interface used to validate both term-level and definition-level English–Amharic translations in the maternal and child health dataset. For each medical concept, the left panel presents candidate Amharic translations of the medical term generated by multiple systems, including machine translation and large language models, while the right panel displays corresponding translated definitions. Experts select the most appropriate term and definition or provide a revised alternative when necessary. This side-by-side design enables consistent alignment between terminology and explanatory definitions, supporting semantic accuracy, clinical clarity, and contextual appropriateness for Ethiopian healthcare communication.

**Evaluation Metrics** Translation quality was assessed using automated metrics to ensure semantic and lexical accuracy. This included BLEU, which measures n-gram overlap between candidate and reference translations; chrF, capturing character-level similarity, particularly useful for morphologically rich languages such as Amharic; and ROUGE-L, assessing longest common subsequence recall. These metrics provided a quantitative evaluation of translation performance, ensuring that the final bilingual resource maintained terminological accuracy, linguistic clarity, and suitability for AI-driven maternal and child healthcare applications.

## 4 Result

A total of 90 maternal and child health (MCH) terms were compiled, with a subset used for model evaluation. Eleven translation systems were compared, including Google Translate, *DeepSeek-R1*, *DeepSeek-R1-0528*, *GPT-4.1-mini*, *GPT-4o-mini*, *LLaMA-3.3-70B*, *Gemma-3-27B*, *Mistral-Large*, *NLLB-200*, *M2M100*, and *Gemini-2.0-Flash*. Each model was evaluated using BLEU, CHRF, and ROUGE-L metrics for both term and definition translations.

### 4.1 Term Translation Result

Table 1 reports results for isolated term translation, a more challenging task due to the absence of contextual information. Performance varies considerably across models, with *DeepSeek-R1* achieving a markedly higher BLEU score than all other systems, indicating strong capability in translating standalone terms. Models including *Gemini-2.0-Flash*, *GPT-4o-mini*, and *Mistral-Large* achieve moderate BLEU scores, while several systems such as Google Translate, *Gemma-3-27B*, *M2M100*, and *NLLB-200* obtain zero BLEU scores, underscoring difficulties with short, context-free inputs. CHRF scores are consistently lower than those observed for definitions, reflecting reduced character level overlap in single word or short phrase translations. As with the definition task, ROUGE-L remains zero for all models.

| Model | BLEU | CHRF |
|---|---|---|
| DeepSeek-R1 | 0.9853 | 3.4814 |
| DeepSeek-R1-0528 | 0.1240 | 2.9403 |
| GPT-4.1-mini | 0.1240 | 2.6257 |
| GPT-4o-mini | 0.4298 | 2.0877 |
| Gemini-2.0-Flash | 0.5077 | 2.3860 |
| Gemma-3-27B | 0.0000 | 2.5298 |
| Google-Translate | 0.0000 | 2.4672 |
| LLaMA-3.3-70B | 0.1774 | 2.6392 |
| M2M100 | 0.0000 | 1.5622 |
| Mistral-Large | 0.3014 | 1.9819 |
| NLLB-200 | 0.0000 | 2.3161 |

Table 1: Term Translation Results

### 4.2 Definition Translation Result

Table 2 summarizes the performance of the model on definition level translation, where inputs consist of longer, context rich sentences. Overall, most models achieve relatively high BLEU and CHRF scores, indicating strong semantic preservation and character level alignment in this setting. Google Translate records the highest BLEU score, suggesting superior accuracy for translating definitions,

while *DeepSeek-R1, GPT-4.1-mini, Gemini-2.0-Flash*, and *Gemma-3-27B* demonstrate comparable and robust performance.

Lower performing systems such as *Mistral-Large* and *M2M100* exhibit substantially reduced BLEU and CHRF values, highlighting challenges in handling longer structured text. ROUGE-L scores are uniformly zero in all models, likely reflecting the limitations of this metric in the current evaluation configuration.

| Model | BLEU | CHRF |
|---|---|---|
| DeepSeek-R1 | 0.9162 | **7.9067** |
| DeepSeek-R1-0528 | 0.8474 | 7.3892 |
| GPT-4.1-mini | 0.8729 | 8.1016 |
| GPT-4o-mini | 0.5593 | 4.1978 |
| Gemini-2.0-Flash | 0.8760 | 7.2057 |
| Gemma-3-27B | 0.8752 | 6.8630 |
| Google-Translate | **0.9554** | 7.7860 |
| LLaMA-3.3-70B | 0.6524 | 7.5908 |
| M2M100 | 0.6493 | 2.7052 |
| Mistral-Large | 0.5429 | 3.6368 |
| NLLB-200 | 0.7583 | 7.0932 |

Table 2: Definition Translation Results

## 5 Discussion

The evaluation of translation systems for maternal and child health (MCH) terminology reveals systematic differences between term-level and definition-level translation, driven primarily by input length, availability of contextual information, and evaluation metric behavior. These findings highlight the interaction between model characteristics and evaluation constraints in low-resource biomedical translation.

Isolated term translation remains particularly challenging due to the absence of contextual cues. Single-word or short-phrase inputs provide limited semantic grounding, forcing models to rely heavily on lexical memorization, domain knowledge, or learned terminology mappings. As shown in Table 1, *DeepSeek-R1* achieves the highest BLEU score and comparatively higher chrF than other systems. This result likely reflects the model's ability to reproduce exact or near-exact lexical forms for short inputs, which are disproportionately rewarded by n-gram–based metrics when sequences are very short. Consequently, high BLEU scores in this setting should be interpreted as indicating strong lexical overlap rather than comprehensive terminological adequacy. Several systems, including Google Translate, *M2M100*, and *NLLB-200*, obtain zero BLEU scores for term translation. This outcome does not necessarily indicate incorrect translations, but instead reflects valid lexical variation in Amharic, such as descriptive paraphrases, alternative morphological realizations, or synonym choices that diverge from the single reference translation (Wieting et al., 2019; Reiter, 2018). In contrast, chrF scores remain non-zero for these models, suggesting partial character-level overlap even in the absence of exact word-level matches. ROUGE-L scores are uniformly zero across systems, which is expected given the extremely short sequence lengths and the limited applicability of longest common subsequence measures to isolated terms.

In contrast, definition-level translation benefits substantially from longer, context-rich inputs. Definitions provide syntactic structure and semantic redundancy, enabling models to better infer meaning and maintain coherence. As shown in Table 2, Google Translate achieves the highest BLEU score, while several LLM-based systems, including *DeepSeek-R1*, *GPT-4.1-mini*, *Gemini-2.0-Flash*, and *Gemma-3-27B*, demonstrate comparable performance. In this setting, BLEU and chrF scores are more stable and interpretable, as longer sequences reduce the impact of exact-match effects and better reflect overall semantic preservation.

The comparatively lower performance of some models, such as *Mistral-Large* and *M2M100*, may be attributed to domain mismatch and limited exposure to specialized biomedical definitions during training. While multilingual MT systems such as *NLLB-200* and *M2M100* maintain reasonable semantic fidelity, they lag behind top-performing systems in capturing precise biomedical phrasing, which is critical for maternal and child health communication. As in term-level evaluation, ROUGE-L remains uninformative, reinforcing its limited suitability for biomedical translation evaluation in low-resource settings under the current configuration.

Overall, the results indicate that translation performance is strongly input-dependent. Term translation prioritizes lexical precision and benefits from models capable of reproducing standardized terminology, whereas definition translation leverages contextual reasoning and favors systems optimized for longer-form text. These findings support a hybrid translation strategy in which specialized LLMs assist with isolated terminology, high-performing

MT or LLM systems handle context-rich definitions, and expert human validation serves as the final quality control layer to ensure clinical safety and linguistic appropriateness.

## 6 Conclusion

This study presented a multilingual, participatory workflow for constructing a standardized maternal and child health terminology resource for low-resource settings, with an initial focus on English–Amharic translation. The approach combined machine translation, large language models, and expert expert-in-the-loop validation to produce an open, machine-readable bilingual dataset for healthcare NLP applications. Evaluation across eleven translation systems showed that translation performance was strongly input dependent: definition-level translation benefited from contextual information, with Google Translate and several LLMs achieving strong results, while isolated term translation remained challenging and favored models with stronger lexical precision. Future work was identified to extend evaluation to the full dataset, refine evaluation methods for short biomedical text, and expand the resource to additional low-resource languages and multimodal healthcare applications.

## Ethical Considerations

This work engages directly with maternal and child health, a high-stakes domain where mistranslation or semantic ambiguity may have serious consequences. To mitigate these risks, all Amharic translations generated by machine translation systems and LLMs were reviewed and validated by a multidisciplinary panel of Ethiopian clinicians, public-health experts, and linguists. This expert-in-the-loop process ensured that clinical meaning, cultural appropriateness, and linguistic fluency were preserved. The study adopts a decolonial AI perspective by centering local expertise, acknowledging epistemic inequalities in global health data, and committing to open access and data sovereignty principles.

## Limitation

This study has several limitations. First, the quantitative evaluation relies on automated metrics (BLEU, chrF, ROUGE-L), which may not fully capture semantic or cultural adequacy, particularly for short technical terms and morphologically rich languages like Amharic. Second, model evaluation was conducted on a subset of 90 terms and definitions, and full dataset evaluation is ongoing. Third, the resource has not yet been deployed in live clinical or patient facing contexts, so real world usability and impact remain untested. Finally, the current work focuses on text-based resources, and further research, and needed to extend the resource to multimodal formats such as speech, images, and video for low literacy populations.

## References


Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. Where Are We? Evaluating LLM Performance on African Languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32704–32731, Vienna, Austria.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R. Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, Daniel Truhn, Renato Cuocolo, Lisa C. Adams, and Keno K. Bressem. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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.

Raphaël Merx, Christine Phillips, and Hanna Suominen. 2024. Machine Translation Technology in Health: A Scoping Review. *Studies in Health Technology and Informatics*, 318:78–83.

Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684.

James Muldoon and Boxi A. Wu. 2023. Artificial intelligence in the colonial matrix of power. *Philosophy & Technology*, 36(4):80.

Graham Neubig and Junjie Hu. 2018. Rapid Adaptation of Neural Machine Translation to New Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Ugochi Okafor. 2025. Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 221–229, Vienna, Austria. Association for Computational Linguistics.

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(1):39.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

World Health Organization. 2025. Maternal Mortality Fact Sheet 2023. `https://www.who.int/news-room/fact-sheets/detail/maternal-mortality`.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

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.

Chaitali Sinha. 2025. Global South-led responsible AI solutions to strengthen health systems: an emergent research landscape. *Oxford Open Digital Health*, 3:oqaf016.

Amare Abera Tareke, Edom Getnet Melak, Bezawit Ketsela Mengistu, Jafar Hussen, and Asressie Molla. 2024. Association between maternal dietary diversity during pregnancy and birth outcomes: evidence from a systematic review and meta-analysis. *BMC Nutrition*, 10(1):151.

Manaye Yihune Teshale, Agegnehu Bante, Abebe Gedefaw Belete, Rik Crutzen, Mark Spigt, and Sarah E. Stutterheim. 2025. Barriers and facilitators to maternal healthcare in East Africa: a systematic review and qualitative synthesis of perspectives from women, their families, healthcare providers, and key stakeholders. *BMC Pregnancy and Childbirth*, 25(1):111.

Getayeneh Tilahun Tiruneh, Muluken Demissie, Alemayehu Worku, and Yemane Berhane. 2022. Predictors of maternal and newborn health service utilization across the continuum of care in Ethiopia: A multilevel analysis. *PLOS ONE*, 17(2):e0264612.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Juexiao Zhou, Haoyang Li, Siyuan Chen, Zhangtianyi Chen, Zhongyi Han, and Xin Gao. 2025. Large language models in biomedicine and healthcare. *npj Artificial Intelligence*, 1(1):44.