ViKhoMT: A Vietnamese–K'Ho Neural Machine Translation Dataset and Evaluation for Community Health Communication

Tram Truong, Vinh Nguyen, Dang Van Thin, Ngan Luu-Thuy Nguyen
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
anhvinh.1805@gmail.com, tramtkn.18@grad.uit.edu.vn
{thindv, ngannlt}@uit.edu.vn

Abstract

The Vietnamese government is prioritizing the socio-economic development and societal integration of ethnic minorities, including the K'Ho people. However, the lack of digital resources creates significant communication barriers, particularly in the critical domain of community health. To address this gap, we introduce ViKhoMT, a new, professionally curated Vietnamese-K'Ho parallel dataset containing approximately 10,000 sentence pairs focused on community health communication. To demonstrate the dataset's quality and establish performance benchmarks, we conducted comprehensive evaluations by fine-tuning several pre-trained Neural Machine Translation (NMT) models. Our experiments show that a system based on the M2M100 architecture achieves BLEU scores of 60.5 for K'Ho-to-Vietnamese and 56.4 for Vietnamese-to-K'Ho, respectively. We release our dataset to the research community for free research purposes to support future studies and the development of practical translation tools for the K'Ho community. The dataset is publicly available at https://github.com/NgocTram2711/ViKhoMT.

Keywords: Vietnamese–K'Ho, Machine Translation, Low-resource Language, Vietnamese Language, K'Ho Language.

1. Introduction

Machine Translation (MT), a core field of Natural Language Processing, has seen remarkable advancements in recent years, driven by the dominance of the Neural Machine Translation (NMT) paradigm (Tan et al., 2020). Unlike earlier statistical methods (Koehn et al., 2003), NMT leverages deep learning to model the complex relationship between languages, aiming to break down communication barriers and enable seamless cross-lingual interaction (Bahdanau et al., 2016). The main entities in MT are the source language text and the target language text, and the system's goal is to produce a translation that is both fluent and adequate. As a result, MT systems have become indispensable tools in global commerce, diplomacy, and information access.

Given the increasing focus on digital inclusion, applying MT methodologies to low-resource languages has become a critical research frontier (Meta AI, 2022). In Vietnam, a country with 54 distinct ethnic groups, this effort is gaining significant momentum. For instance, recent work has demonstrated the potential of NMT for Southeast Asian languages, with one study on the Vietnamese-Khmer language pair achieving a notable BLEU score of 55.37 (Thai Nguyen Quoc and Van, 2022), and another recent study on Vietnamese-Bahnaric also reporting strong results (Nguyen et al., 2025). However, significant gaps remain for many other ethnic minority languages. The K'Ho language, spoken by approximately 200,000 people, still suffers from a severe scarcity of high-quality domain-specific digital resources. This impedes the community's access to vital information, particularly in the healthcare sector, where the materials are predominantly Vietnamese. Motivated by this critical need, our contribution in this work is presented below:

1. Our primary effort was the construction of ViKhoMT, a new domain-specific Vietnamese-K'Ho parallel corpus containing approximately 10,000 sentence pairs. This foundational dataset was meticulously compiled by combining data from diverse sources, including OCR-processed documents and manually translated health articles.

2. Building upon this new resource, we conducted a systematic evaluation of modern pre-trained models to establish the first strong performance benchmarks for this language pair. This process provides a critical reference point for any subsequent research. The outcome is a practical, high-quality translation system that directly addresses the critical communication gap in healthcare for the K'Ho community, highlighting the tangible societal impact of applying NMT to serve underserved languages.

The remainder of this paper is organized as follows. Section 2 reviews related work in the field. Section 3 presents a detailed process for creating our ViKhoMT dataset. In Section 4, we describe our methodology, from the selection of pre-trained

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8725–8739
11–16 May 2026.

models to data augmentation techniques. Section 5 reports on our experiments, including comparisons between different models and analyses of scenarios with and without back-translation and data filtering. Finally, Section 6 concludes the paper by summarizing our key findings and contributions.

2. Related Works

Pioneering studies in Vietnamese–K'Ho machine translation often focused on non-neural methods for the weather forecasting domain. For instance, one of the earliest works applied an Example-Based Machine Translation (EBMT) approach, but was constrained by its dataset of only 212 sentence pairs (Nguyen and Dinh, 2016). Another study utilized a Statistics Machine Translation (SMT) based approach for the same domain(Hiep et al., 2018). More recently, a (Nguyen et al., 2023) study utilized OpenNMT (Klein et al., 2017) with a larger, multi-domain dataset of 16,217 sentence pairs. Despite the larger corpus, this work lacked a specific focus on a single domain and reported only accuracy rather than widely adopted MT metrics such as BLEU (Papineni et al., 2002), which limited the comparability of its results with other NMT research.

The broader NMT field has since been revolutionized by the Transformer architecture (Vaswani et al., 2017) (Cho et al., 2014). However, training Transformers from scratch requires large-scale parallel corpora, and they often struggle to generalize on limited data, as is the case for low-resource languages.

The current standard approach to overcome data scarcity is to fine-tune pre-trained multilingual translation models (Zoph et al., 2016). Instead of learning from scratch, this method leverages the vast linguistic knowledge already encoded in the model's parameters. In particular, the development of many-to-many translation models like M2M100 (Fan et al., 2021) marked a significant breakthrough, enabling direct translation between 100 language pairs without pivoting through English. This line of research was further advanced by the NLLB-200 project (Meta AI, 2022), which scaled multilingual translation to 200 languages with a dedicated focus on improving quality for low-resource languages.

More recently, Large Language Models (LLMs) such as GPT-4 have shown strong translation capabilities for high-resource languages. However, recent large-scale evaluations reveal significant limitations for low-resource settings. (Robinson et al., 2023) evaluated ChatGPT across 204 languages and found that it underperformed traditional MT systems for 84.1% of languages, with low-resource languages being the most affected. Similarly, (Zhu et al., 2024) showed that while GPT-4 surpassed the NLLB baseline in about 40% of translation directions, it still lagged behind for low-resource pairs. For an extremely low-resource language comparable to K'Ho, (Merx et al., 2024) tested RAG-augmented LLM prompting for Mambai (also 200,000 speakers) and achieved a maximum BLEU of only 21.2, far below what fine-tuned NMT models can achieve. These findings suggest that fine-tuning dedicated multilingual NMT models remains the most effective paradigm for languages with limited but curated parallel data.

Regardless of model choice, a persistent challenge for low-resource NMT is the limited size of available parallel corpora. To expand limited training datasets, back-translation (Sennrich et al., 2016) (Lample et al., 2018) (Poncelas et al., 2018) is one of the most effective and widely adopted data augmentation techniques. The technique uses a preliminary "seed" model to translate monolingual target-language data back into the source language, creating synthetic parallel data that exposes the model to greater lexical and structural diversity.

A challenge with back-translation is that the resulting synthetic data can be noisy or contain translation errors. To ensure quality, data filtering methods have been proposed. The core idea is that a sentence translated from language A to B, and then back from B to A, should retain its original meaning. As demonstrated by(Imankulova et al., 2017), back-translating the synthetic data and measuring the similarity between the original and the round-trip sentence with a sentence-level BLEU metric can filter out low-scoring pairs, retaining only the highest-quality synthetic data for training.

3. Datasets

3.1. The ViKhoMT Parallel Corpus

The foundation of this research is ViKhoMT, a new parallel corpus meticulously compiled over a six-month period, specifically for the K'Ho-Vietnamese language pair within the community health domain. The final dataset consists of 10,027 sentence pairs, aggregated from three primary sources.

Data Extracted via OCR We collected approximately 4,000 sentence pairs from the bilingual online newspaper Báo ảnh Dân tộc và Miền núi. This monthly publication includes versions translated into 12 ethnic minority languages, including K'Ho. Our data were primarily extracted from the Health and Education sections. Text extraction was performed using Google OCR.

Manually Translated Data To ensure high quality and domain expertise, we manually translated 200 articles related to health. The translation was

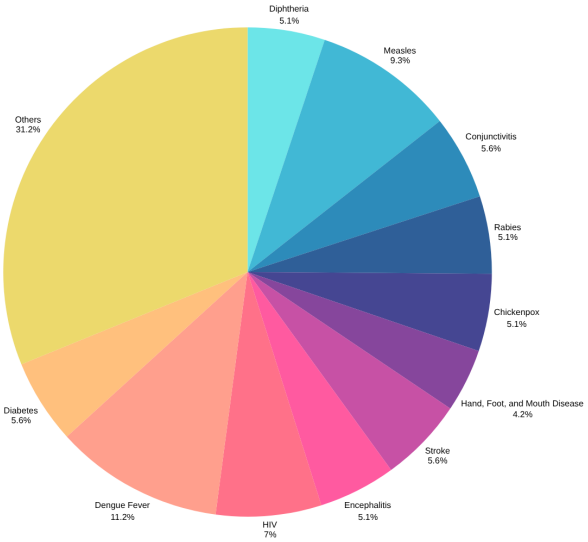

Figure 1: Distribution of the 11 prevalent diseases and other health topics covered in the 200 manually translated health articles

Table 1: Examples of common OCR errors.

Error Type	OCR Error	Correct
Loss of diacritics	kosot	kơsơt
Incorrect diacritics	gơwét	gơwèt
Misinterpretation of special characters	hở-tàng	hở-tàng
Confusion of similar characters	kơnhuài	kơnhuàl

performed by a team of three experts: a member of the VOV4 K'Ho language service (the Voice of Vietnam's ethnic minority broadcasting division) with 25 years of experience who has contributed to numerous scientific projects on K'Ho language preservation, and two K'Ho language teachers from schools in Di Linh, Lam Dong, who are also active translators for the Di Linh Parish. The 200 articles selected for manual translation focused on 11 prevalent diseases within the community, alongside other general health topics. Figure 1 illustrates the distribution of these topics, showing a significant focus on Dengue Fever (12.0% of articles), Measles (9.3%) and HIV (7.5%), which are major public health concerns in the region. Any dialectal differences in vocabulary were standardized against the style guide of Báo ảnh Dân tộc và Miền núi. The team followed a rigorous, twofold workflow:

K'Ho-to-Vietnamese Translation 100 articles sourced from the official VOV4 K'Ho website were translated into Vietnamese. In this process, the VOV4 expert performed the primary translation, which was then reviewed and validated by the two teachers.

Vietnamese-to-K'Ho Translation To overcome the scarcity of K'Ho source materials, 100 Vietnamese articles from reputable outlets (such as Sức Khỏe Đời Sống, Báo Dân Tộc, and VTV) were translated into K'Ho. For this task, each teacher translated 50 articles; the translations were then cross-reviewed by the other teacher before a final validation by the VOV4 expert.

Data from a Digitized Dictionary The corpus was further enhanced with data from a digitized 1983 Vietnamese-K'Ho dictionary, a resource provided by the Lam Dong Department of Culture and Information and digitized by Nguyen Minh Thao of the Bao Lam Medical Center. From this dictionary, which contains 7,065 words and 5,923 example sentences, we leveraged the content in two primary ways: the entire vocabulary was used to build a custom tokenizer for K'Ho intended for models that require training from scratch, while 856 health-related example sentences were extracted and filtered to incorporate into our parallel corpus.

3.2. Monolingual Corpora for Back-Translation

In addition to the ViKhoMT parallel corpus, two large-scale monolingual datasets were a crucial component of our methodology. These corpora played a key role in the back-translation technique used to augment the training data.
Vietnamese: We utilized the Vietnamese Curated Text Dataset developed by Viettel Solutions. The original dataset contains 12 million lines, of which the health domain comprises 7.42%. From this resource, we randomly extracted and preprocessed over 30,000 health-focused Vietnamese sentences.
K'Ho: To obtain monolingual K'Ho data, we proactively crawled and processed 23,000 sentences from the health section of the VOV4 news website (vov4.vov.vn/kho), a reputable source for the K'Ho community.

3.3. Data Preprocessing

To ensure the quality and consistency of the final corpus, all aggregated data underwent a multi-stage preprocessing pipeline.

First, we performed a series of general cleaning steps across the entire dataset. This included removing extraneous whitespace, residual HTML tags, and non-Latin characters that resulted from data crawling and extraction. We then segmented the text into sentences, primarily using the period (.) as a delimiter. We then expanded common Vietnamese abbreviations to their full forms (e.g., "bs" to "bác sĩ", "ths" to "thạc sĩ", "vn" to "Việt Nam", "tp" to "thành phố") and removed duplicate sentence

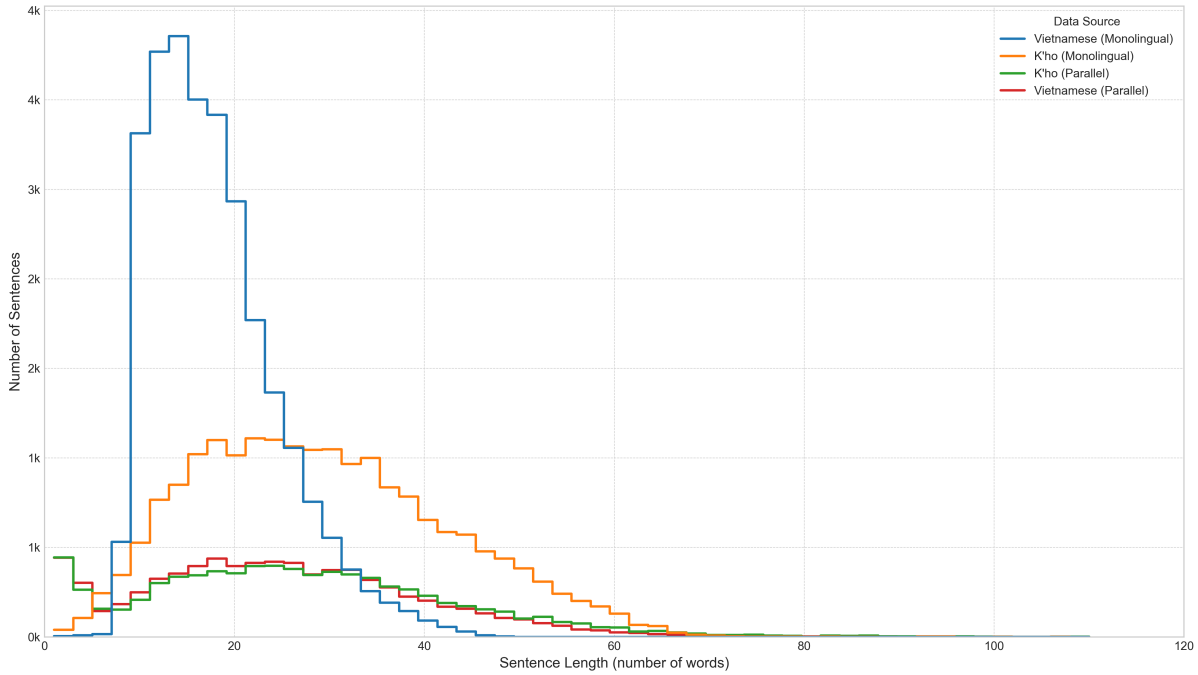

Figure 2: Sentence length distribution of the parallel (ViKhoMT) and monolingual corpora.

pairs to ensure data uniqueness. Each clean, parallel sentence was formatted as a single line with Vietnamese and K'Ho text separated by a tab character.

Next, source-specific processing was applied. For data extracted via OCR from Báo ảnh Dân tộc và Miền núi, the text first underwent segmentation and alignment. This involved a semi-automated procedure where paragraphs were manually aligned, followed by a Python script to perform one-to-one sentence alignment based on sentence length ratios and punctuation cues. After alignment, all text was normalized to Unicode NFC to resolve character inconsistencies. However, because of the specific orthography of the K'Ho language, the raw OCR output contained significant noise and errors (see Table 1). This necessitated a rigorous manual post-editing process where language experts meticulously corrected common issues to ensure the fidelity of the final data.

A final, crucial step was the orthographic unification of data from the digitized 1983 dictionary. This dictionary followed a 1982 provincial guideline that aimed to simplify K'Ho orthography by mapping its unique characters to standard Vietnamese accented letters (e.g., using é to represent ĕ, ó to ŏ, í to ĭ). However, this simplified standard is inconsistent with the orthography used in most modern, official K'Ho publications, which retain the original, more precise diacritics. To ensure consistency across the entire dataset, we implemented a reverse mapping on the dictionary data, converting the simplified characters back to the modern K'Ho orthography.

Figure 2 illustrates the sentence length distribution for all corpora after preprocessing. All four corpora exhibit a right-skewed distribution concentrated in the 5 to 40-word range, well-suited for training Transformer-based models.

4. Methodology

Our methodology is designed to address the challenges of low-resource NMT by combining the power of pre-trained models with a multi-stage data augmentation and filtering pipeline. This approach allows us to maximize the utility of our limited parallel data while leveraging large-scale monolingual corpora to improve model performance. The overall workflow, from initial data to the final bidirectional models, is illustrated in Figure 3.

4.1. Data Augmentation

To expand our limited parallel corpus, we employed two distinct data augmentation techniques: paraphrasing via synonym replacement (Wei and Zou, 2019) and large-scale back-translation.

First, we augmented the Vietnamese side of the ViKhoMT corpus using a paraphrasing technique. We utilized an AI-powered language tool to rewrite Vietnamese sentences by replacing words with synonyms while preserving the original structure and meaning. All generated paraphrases were manually reviewed by human experts before being added to the dataset. This process increased our ini-

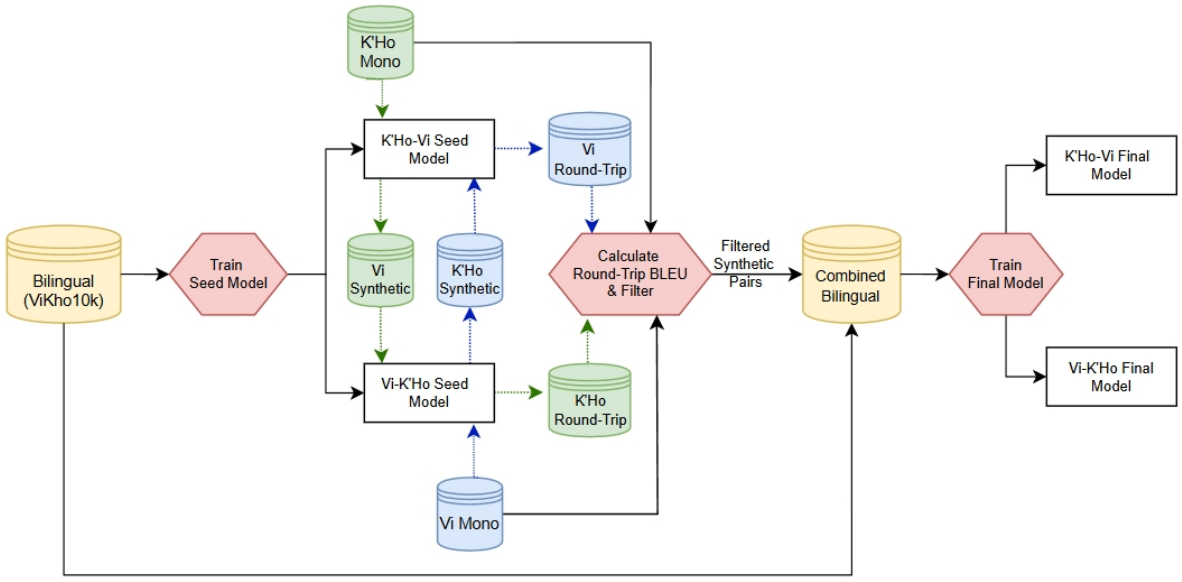

Figure 3: Our overall workflow for Vietnamese-K'Ho machine translation.

tial parallel corpus from approximately 10,000 to 11,000 sentence pairs.

The primary augmentation method was back-translation. After training initial "seed" models (detailed in Section 4.3) on the ViKhoMT corpus, we used them to generate a large synthetic parallel dataset. Specifically, the 30,000-sentence Vietnamese monolingual corpus was translated into K'Ho using the Vi-K'Ho seed model, and the 23,000-sentence K'Ho monolingual corpus was translated into Vietnamese using the K'Ho-Vi seed model. This step generated a total of 53,000 new synthetic sentence pairs. Table 2 provides a comprehensive summary of all primary and generated corpora used in our methodology, including their sources, sizes, and primary uses.

4.2. Synthetic Data Filtering

While back-translation generates a large volume of data, its quality can be inconsistent. To ensure that only high-quality synthetic data was used for final training, we implemented a rigorous data filtering process based on the round-trip translation (RTT) consistency principle (Imankulova et al., 2017). This process, illustrated in Figure 3, was applied to both monolingual corpora as follows:

- Filtering the K'Ho Monolingual Corpus (Green path in Figure 3): An original K'Ho sentence ($K'Ho_{orig}$) is first translated by the K'Ho-Vi Seed Model to generate a synthetic Vietnamese sentence (Vi_{synth}). This synthetic sentence is then immediately translated back by the Vi-K'Ho Seed Model to produce a round-trip K'Ho sentence ($K'Ho_{rtt}$). The BLEU score is then calculated between the original and the round-trip version, BLEU($K'Ho_{orig}$, $K'Ho_{rtt}$).

- Filtering the Vietnamese Monolingual Corpus (Blue path in Figure 3): Similarly, an original Vietnamese sentence (Vi_{orig}) is translated by the Vi-K'Ho Seed Model to generate a synthetic K'Ho sentence ($K'Ho_{synth}$). This is then translated back by the K'Ho-Vi Seed Model to produce a round-trip Vietnamese sentence (Vi_{rtt}). The score is calculated as BLEU (Vi_{orig}, Vi_{rtt}).

To retain only high-quality synthetic data, we implemented a percentile-based filtering strategy. Specifically, we calculated the RTT BLEU score for all generated pairs and then ranked them by score, retaining only the top 80% of pairs. This threshold was determined through a preliminary experiment comparing different filtering levels, where 80% achieved the best balance between data quality and diversity (see Appendix A.1). The corresponding absolute BLEU score cutoffs were >36.39 for Vietnamese and >51.67 for K'Ho. Illustrative examples of retained and rejected pairs under these cutoffs are provided in Appendix A.2.

4.3. Training Model

Our training methodology is based on a two-stage process designed to maximize the effectiveness of our data augmentation pipeline.

The first stage is Seed Model Training. Here, initial models are trained on the high-quality, human-curated parallel corpus (ViKhoMT). A separate pair of seed models (K'Ho-to-Vietnamese and Vietnamese-to-K'Ho) was trained for each of the five evaluated architectures described in Section

Table 2: Summary of all corpora used in this study. The "Source" column indicates the origin of the primary corpora or the method used to generate data. The synthetic data size is shown before the filtering process.

Corpus	Source	Size	Primary Use
Primary Corpora			
ViKhoMT	OCR, Manual Translation, Dictionary	10,027	Train Seed, Final Models
Vi Monolingual	Vietnamese Curated Text Dataset	30,000	Back-translation
K'Ho Monolingual	VOV4 News	23,000	Back-translation
Generated Corpus			
Paraphrased Data	Synonym Replacement	~1,000	Train Seed, Final Models
Synthetic Parallel Data	Back-translation	53,000	Final Models

5.2. The primary purpose of these "seed" models is not to be the final product, but to serve as proficient translators for the subsequent back-translation task.

The second stage is Final Model Training. After the data augmentation and filtering steps are complete, the final models are trained on a comprehensive, combined dataset. Following the data mixing strategy introduced by Sennrich et al. (2016), the final training dataset is a combination of the original parallel corpus and the high-quality synthetic data generated in the previous steps. To ensure the models prioritize learning from the more reliable human-translated data, we employ an up-sampling strategy, increasing the weight of the original corpus during training (Xu et al., 2022). This two-stage approach allows the final models to benefit from both the accuracy of the original data and the linguistic diversity of the augmented data.

5. Experimental

5.1. Experiment Setup

Dataset and Evaluation The ViKhoMT parallel corpus (consisting of 10,027 sentence pairs) was divided into three subsets. First, we held out a fixed test set of 1,000 sentence pairs for final evaluation across all experiments. The remaining 9,027 pairs were then split into a training set of 8,124 pairs (90%) and a validation set of 903 pairs (10%). This ensures a fair and consistent comparison between all models and scenarios. For the evaluation metric, we adopt BLEU (Papineni et al., 2002) as our primary metric for comparing model performance. We utilize sacreBLEU, a standardized implementation, to ensure reproducible and comparable BLEU score computations. A higher score indicates a better translation.

Implementation Details All experiments were conducted using the Hugging Face Transformers library on a single NVIDIA RTX 4090 GPU. Key hyperparameters for fine-tuning the pre-trained models were kept consistent to ensure a fair comparison, including a learning rate of 3×10^{-5}, a batch size of 16, and mixed-precision training. Seed models were trained for 20 epochs, while final models were trained for 10 epochs, both with early stopping based on validation loss to prevent overfitting.

5.2. Baselines

To comprehensively evaluate the effectiveness of our proposed methodology, we selected and compared the performance of the following five baseline model architectures:

Transformer-base This model, based on the original architecture by (Vaswani et al., 2017), was trained from scratch on our ViKhoMT corpus. It serves as a baseline to quantify the benefits of pre-training.

mBART-50 Proposed by (Tang et al., 2020), Multilingual BART is pre-trained using a denoising objective on monolingual text from 50 languages. It learns to reconstruct original text from corrupted input, enabling it to build robust multilingual representations.

M2M100 This model, introduced by (Fan et al., 2021), was the first to demonstrate many-to-many translation among 100 languages without pivoting through English. It is trained on a massive dataset mined from the web for numerous language pairs.

NLLB-200 As part of the "No Language Left Behind" project by (Meta AI, 2022), this model family scales multilingual translation to 200 languages with a particular focus on improving quality for low-resource languages. Due to hardware constraints, we selected the nllb-200-distilled-600M version over the larger 1.3B variant. This distilled model is optimized to provide a strong balance between high performance and manageable computational requirements for our experimental setup.

SeamlessM4T Short for Massively Multilingual & Multimodal Machine Translation (Communication et al., 2023), this is a single, unified model developed by Meta AI that supports multiple tasks and modalities, including speech-to-speech (S2ST), speech-to-text (S2TT), and text-to-text (T2TT) translation for up to 100 languages. Although it has audio processing capabilities, for the scope of this research, we leverage its T2TT functionality. Due to hardware constraints, we selected the SeamlessM4T-medium version (1.2B parameters) over the large version (2.3B parameters). The model is built on the UNITY architecture, integrating powerful pre-trained components for both speech and text. Including SeamlessM4T as a baseline allows for the evaluation of a state-of-the-art multimodal architecture on a specialized T2TT task.

5.3. Experimental Scenarios

To evaluate our methodology, we conducted experiments on the five chosen model architectures across two main scenarios, corresponding to the training stages outlined in Section 4.3.

Scenario 1 Seed Model Evaluation. We evaluate the performance of the "seed" models for each of the five architectures. These models were trained only on the original ViKhoMT corpus and serve as the performance baseline.

Scenario 2 Final Model Evaluation. We evaluate the performance of the "final" models. These were trained on the combined dataset of original and augmented data after the full back-translation and filtering pipeline was applied to each architecture. This scenario measures the full impact of our proposed methodology on each model.

5.4. Results

Several key observations can be drawn from the results in Table 3. First, by comparing the performance of the Seed Models, it is evident that all four pre-trained models (NLLB-200, SeamlessM4T, mBART-large-50, and M2M100_418M) significantly outperform the Transformer-base model trained from scratch. This once again confirms the immense value and indispensable role of transfer learning in the field of Neural Machine Translation (NMT) for low-resource languages like K'Ho. Among the pre-trained models, M2M100_418M consistently achieved the highest baseline performance on the original ViKhoMT dataset, reaching a BLEU score of 57.3 for K'Ho-Vi and 53.2 for Vi-K'Ho. The remaining models followed with decreasing performance: mBART-large-50, SeamlessM4T-medium, and NLLB-200-600M.

Table 3: Evaluation results (BLEU score) for all models and training stages.

Model Architecture	Training Stage	K'Ho → Vi	Vi → K'Ho
Transformer-base	Seed Model	17.2	21.5
	Final Model	24.9	25.4
NLLB-200-600M	Seed Model	48.7	42.1
	Final Model	53.9	47.2
SeamlessM4T	Seed Model	50.1	44.4
	Final Model	54.4	48.6
mBART-large-50	Seed Model	51.9	46.6
	Final Model	55.5	48.9
M2M100_418M	Seed Model	57.3	53.2
	Final Model	**60.50**	**56.42**

Second, the impact of our data augmentation pipeline is consistently positive across all tested architectures. For each pre-trained model, the Final Model—trained on the combined dataset—shows a substantial improvement in BLEU scores compared to its corresponding Seed Model. For instance, the score of M2M100 for the Vi-K'Ho direction increased from 53.2 to 56.42, while SeamlessM4T also saw a significant jump from 50.1 to 54.4 for the K'Ho-Vi direction. This validates that our back-translation and data filtering workflow is an effective method for leveraging monolingual data to boost translation quality.

Finally, the M2M100_418M model, when combined with the full data augmentation and filtering pipeline, achieved the highest overall scores among all evaluated architectures. The final model achieved a BLEU score of **60.5** for K'Ho-to-Vietnamese and **56.42** for Vietnamese-to-K'Ho, underscoring its superior capability for this specific task compared to the other evaluated models, including powerful and recent multimodal architectures like SeamlessM4T.

For contextual reference, prior works on this language pair include (Nguyen et al., 2023) who achieved an accuracy of 56.54 using OpenNMT with a private 16,000-pair dataset, and (Nguyen and Dinh, 2016) who explored EBMT with only 212 pairs. However, direct comparison with these works is limited due to differences in training data, model architectures, and evaluation sets. Our contribution is therefore best understood not as a strict performance comparison, but as the first publicly available, domain-specific benchmark for this language pair.

5.5. Optimizing Back-Translation Data Size

The volume of monolingual data used for back-translation can significantly impact model performance (Edunov et al., 2018). To determine the optimal configuration for our main experiments, we conducted a preliminary ablation study across all five model architectures. We experimented with different sizes of the Vietnamese and K'Ho mono-

lingual corpora, training a separate "final" model for each configuration on each architecture.

The results consistently showed that the configuration using 30,000 Vietnamese and 23,000 K'Ho sentences yielded the best or near-best performance across all tested models. To illustrate this trend and for the sake of brevity, we present the detailed results for the M2M100_418M model in Table 4. The results also indicate that simply increasing the amount of monolingual data does not guarantee better performance. For the Vi-K'Ho direction, increasing the Vietnamese monolingual data from 30k to 40k sentences led to a decrease in the BLEU score (from 56.42 to 54.99). We hypothesize that this may be due to the introduction of noise from the Vi-K'Ho seed model when translating a larger, potentially more out-of-domain, monolingual corpus.

To validate this hypothesis, we employed xCOMET (Guerreiro et al., 2023), a state-of-the-art metric capable of fine-grained error detection, for a detailed qualitative analysis. This analysis was conducted on a set of 100 parallel sentences withheld from the training data, which included both in-domain (health) and out-of-domain topics to assess the model's robustness. The xCOMET analysis confirmed our hypothesis, revealing a high rate of semantic errors when the seed model processed the out-of-domain source sentences. The tool frequently flagged mistranslations of specialized, non-health-related terminology, which generated noisy and nonsensical K'Ho sentences. For a side-by-side comparison of specific examples from this analysis that highlight the performance differences between the 30k and 40k models, please refer to Appendix A.3. This supports the claim that indiscriminate expansion of monolingual data can degrade the quality of the synthetic corpus, thereby negatively impacting the final model's performance. Therefore, we selected the 30k Vietnamese and 23k K'Ho configuration as the optimal balance between data volume and quality for the data augmentation pipeline in our main experiments.

5.6. Error and strength analysis

To gain a deeper, qualitative understanding of the final model's performance beyond quantitative metrics, we conducted a detailed error and strength analysis. The model's primary strength is its robust capability to accurately translate domain-specific medical terminology, particularly in the K'ho-to-Vietnamese direction, where it produces highly reliable translations. However, the analysis also identified several recurring error patterns. The Vietnamese-to-K'ho direction proved more challenging, exhibiting a higher frequency of lexical and orthographic errors, especially with the language's unique diacritics. Furthermore, both models showed a noticeable performance degradation when handling out-of-domain topics, often failing to translate metaphorical language or non-medical specialized terms. A comprehensive breakdown of these error categories, including specific examples and a side-by-side comparison, is provided in Appendix A.4.

Table 4: Ablation study on monolingual data size for back-translation using the M2M100_418M model. Scores are reported in BLEU.

Monolingual Data Sizes	K'Ho → Vi	Vi → K'Ho
10k Vi, 10k K'Ho	58.73	55.66
10k Vi, 23k K'Ho	58.95	55.62
23k Vi, 23k K'Ho	59.00	54.73
30k Vi, 23k K'Ho	**60.50**	**56.42**
40k Vi, 23k K'Ho	59.51	54.99

6. Conclusion

This paper introduces ViKhoMT, a new, high-quality Vietnamese-K'Ho parallel corpus specifically tailored for the community health domain. By leveraging this resource with a robust data augmentation and selection pipeline on pre-trained models, we established strong performance benchmarks. Our best system, based on the M2M100_418M architecture, achieved a BLEU score of 60.5 for K'Ho-to-Vietnamese and 56.4 for Vietnamese-to-K'Ho translation. ViKhoMT not only serves as the foundation for a high-performance translation system but also demonstrates the significant potential of modern NMT techniques to create practical tools that help preserve and promote the vitality of minority languages in the digital age. This work provides ViKhoMT as a valuable resource to enable future research, enhancing health communication and fostering information equity for the K'Ho community in Vietnam.

Acknowledgments

This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCM), under grant number NCM2025-26-02.

7. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On

the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual multimodal machine translation.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

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, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

N. M. Hiep, N. T. Luong, L. V. Phuong, N. T. M. Huyen, and D. V. Tuan. 2018. An application to translate from vietnamese into k'ho using stmt approach. *Dalat University Journal of Science*, 8(2):3–12.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78. Asian Federation of Natural Language Processing.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, page 48–54.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Meta AI. 2022. No language left behind: Scaling human-centered machine translation.

Long Nguyen, Tran Le, Huong Nguyen, Quynh Vo, Phong Nguyen, and Tho Quan. 2025. Serving the underserved: Leveraging BARTBahnar language model for bahnaric-Vietnamese translation. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities*, pages 32–41.

Minh Tuan Nguyen and Viet Tuan Dinh. 2016. Vietnamese–k'ho machine translation using ebmt approach. *Dalat University Journal of Science*, 6(2).

Thi Luong Nguyen, Quoc Thang La, Nhat Quang Tran, and et al. 2023. A research on vietnamese-k'ho language translation system using neural machine translation. *TNU Journal of Science and Technology*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Huong Le Thanh Thai Nguyen Quoc and Hanh Pham Van. 2022. Improving Khmer-Vietnamese Machine Translation with Data Augmentation Methods. In *Proceedings of the 11th International Symposium on Information and Communication Technology*, pages 276–282.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6382–6388.

Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shuming Shi, Lihui Chen, and Lemao Liu. 2022. On synthetic data for back translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 419–430.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *CoRR*, abs/1604.02201.

A. Appendix

A.1. Effect of Round-Trip Filtering Threshold

To determine the optimal filtering threshold for our round-trip translation (RTT) quality filtering, we conducted a preliminary experiment comparing five percentile-based thresholds using the M2M100_418M model. In each configuration, only the top X% of synthetic pairs (ranked by RTT BLEU score) were retained for final training. The results are reported in Table 5.

Threshold	K'Ho → Vi	Vi → K'Ho
60%	59.02	54.12
70%	60.33	54.39
80%	**60.50**	**56.42**
90%	59.51	54.99
100%	59.02	54.74

Table 5: Effect of round-trip filtering threshold on M2M100_418M performance (BLEU). Bold indicates the best score.

The results show that the 80% threshold yields the best performance in both translation directions. A stricter threshold (60% or 70%) removes not only noisy synthetic pairs but also useful data that contributes to lexical and structural diversity, leading to reduced generalization. Conversely, retaining all unfiltered data (100%) would introduce a substantial amount of translation noise, as demonstrated by the round-trip examples in Appendix A.2. The 80% threshold therefore provides the best trade-off between data quality and quantity, and was adopted as the default for all final experiments reported in this paper.

A.2. Synthetic Data Filtering Examples

Table 6 and Table 7 provide illustrative examples of the round-trip translation filtering process, generated by the M2M100_418M seed model, for the Vietnamese and K'Ho monolingual corpora, respectively. The "Decision" column indicates whether a synthetic pair was kept or rejected based on the round-trip BLEU score.

A.3. Qualitative Analysis for the Ablation Study on Monolingual Data Size

The quantitative results in Table 4 show that increasing the Vietnamese monolingual data from 30k to 40k sentences led to a performance drop for the Vi-K'Ho translation direction. To investigate the underlying reasons, we conducted a qualitative analysis by comparing the outputs of the model trained on 30k sentences (30k Model) against the one trained on 40k sentences (40k Model).

The table below (Table 8) presents a side-by-side comparison of their translations for identical source sentences. These examples illustrate that the 30k Model, despite using less data, consistently produces translations of higher quality. This superiority is evident not only in its semantic and grammatical accuracy but is also reflected in its sentence-level xCOMET scores. It reveals several key advantages of the 30k Model over the 40k Model:

Superior Semantic Accuracy In example ID 1, the superiority of the 30k Model is reflected in its higher xCOMET score (0.314 vs. 0.291). The 30k Model perfectly translates the difficult concepts of "ép" (to urge/force) and "mâu thuẫn" (conflict) into the correct K'ho phrase sỏr, tam lơh tai. In contrast, the 40k Model fails to grasp the meaning and instead hallucinates a nonsensical repetition, đơs wơl, đơs wơl (said again, said again), leading to a complete loss of the original intent.

Higher Grammatical Precision Example ID 2 highlights a subtle but important grammatical error. This difference in quality is clearly quantified by the xCOMET scores: the 30k Model achieves 0.529, significantly higher than the 40k Model's 0.454. The 30k Model correctly translates "khoảng" (around/approximately) to pơgăp. The 40k Model, however, produces pơgăp bơh, adding an unnecessary preposition (bơh) that makes the phrase less natural and grammatically imprecise in this context.

Greater Faithfulness (Adequacy) Example ID 3 is a special case where the 40k Model (xCOMET: 0.428) scores higher than the 30k Model (xCOMET: 0.327) despite omitting information. This can occur because the 40k Model's translation, while less complete, is fluent and lacks obvious grammatical errors, which automated metrics may prioritize. However, in terms of adequacy, a core criterion in translation, the 30k Model is clearly superior as it accurately translates the term "rủi ro" (risks) to rềs àr, whereas the 40k Model omits it entirely.

This qualitative analysis strongly supports the quantitative results. The 30k Model consistently outperforms the 40k Model by producing more semantically and grammatically accurate translations. Even when automated scores do not fully capture the quality difference (as in ID 3), a deeper analysis reveals that the 30k Model is more faithful to the source content. This confirms our hypothesis that the additional 10,000 monolingual sentences used to train the 40k Model likely introduced more "noise" than useful signal, leading to a degradation in its translation quality.

A.4. Error and Strength Analysis

Strength of Preservation of Medical Terminology A primary strength of our fine-tuned models is their robust ability to accurately translate domain-specific medical terminology. This capability is demonstrated in Table 9 and Table 10 , which present selected examples for the K'ho-to-Vietnamese and Vietnamese-to-K'ho directions, respectively. It is important to note that these examples were taken from the test set and were therefore unseen by the model during training. As shown in these tables, the models successfully handle a wide range of medical concepts—from disease names and treatments to physiological processes—producing translations that are both fluent and adequate. Notably, the performance in the K'ho-to-Vietnamese direction (Table 9) is exceptionally high, with many translations being nearly identical to the human reference, underscoring the model's reliability for in-domain tasks.

Qualitative Error Analysis While the K'Ho-to-Vietnamese model demonstrates strong overall performance, a qualitative analysis of its outputs reveals several recurring error patterns. This section provides illustrative examples from the test set, categorizing common weaknesses to offer a more nuanced understanding of the model's limitations. Table 11 presents a summary of these errors. The analysis indicates that while the K'ho-to-Vietnamese model is highly effective for many in-domain sentences, its primary weaknesses include:

- A tendency to produce literal but contextually suboptimal translations (Lexical Errors).
- Occasional but significant failures in comprehending complex phrases, leading to clinically

incorrect statements (Semantic and Hallucination Errors).

- Unreliability in accurately transcribing numerical entities (Entity Errors).

These findings suggest that while the model is a strong baseline, further work is needed to improve its contextual understanding and reliability with critical data points like numbers.

Table 6: Examples of Round-Trip Translation filtering for the Vietnamese monolingual corpus. The threshold for the top 80% was a BLEU score > 36.39.

Original Vietnamese	Round-trip Vietnamese	BLEU	Decision
với những trường hợp bị loét chân cần phải điều trị chăm sóc vết thương tích cực	với những trường hợp bị tổn thương chân phải điều trị chăm sóc vết thương tích cực	70.86	Kept
chỉ trừ một số trường hợp như bệnh vảy nến thể viêm khớp, vảy nến gây viêm thì các bác sĩ sẽ tiến hành thêm một số chẩn đoán khác để xác định bệnh	chỉ ngoài một số trường hợp như bệnh vảy nến gây viêm khớp, vảy nến gây viêm thì các bác sĩ sẽ tiến hành thêm một số xét nghiệm khác để xác định bệnh	76.12	Kept
đã vậy nhiều phòng khám tư ở đây còn tìm cách móc túi trắng trợn bệnh nhân khiến không ít người phải "dở khóc dở cười"	đã như vậy nhiều phòng khám tư nhân ở đây đang tìm cách xóa kín bệnh nhân làm không ít người phải "khổ thở"	32.19	Rejected
với người lớn có thể súc miệng thường xuyên bằng các loại nước xúc miệng có tính sát trùng	đối với người lớn có thể quấy khóc liên tục và các loại nước quấy khóc có thể bị quấy khóc	23.42	Rejected

Table 7: Examples of Round-Trip Translation filtering for the K'Ho monolingual corpus. The threshold for the top 80% was a BLEU score > 51.67.

Original K'Ho	Round-trip K'Ho	BLEU	Decision
sơm kòp niam den ngai do kòn ùr añ geh sồr rê	sơm kòp niam den tũ do kòn ùr añ geh sồr rê	73.49	Kept
ờ hết mŭt tàm kàl mìu tiah đah jum (bơh nhai 5 tus nhai 11) mơya kòp mprơm gơlik mhàm neh gơguh uă	ờ hết mut tàm kàl mìu tiah đah jum (bơh nhai 5 tus nhai 11), mơya kòp sốt xuất huyết neh gơguh uă	61.42	Kept
tàm dùl poh, bòn dờng hồ chĩ minh kung neh gŏ rơlao 300 nă cau gơtìp kòp jơng tê bơr	tàm dùl poh, òn dờng hồ chĩ minh kung neh gĩt gŏ rlau 300 nă cau bơtờp kòp jơng tê bơr	49.49	Rejected
jơh 2 ngai gŏ kòn dết ờ gơmù kòp mờ bĭc sùm gơtìp rơtơs jơng tê, ñìm, ờ bài sào mờ duh să den tàng cèng tus tàm do tai	jơl 2 ngai gŏ kòn dết ờ bời mờ bĭc sùm gơtìp rơtŏt să, ñìm sùm, ờ gơdan sào mờ duh să den tàng jun tus tàm do tai	46.87	Rejected

Table 8: Side-by-side comparison of translation outputs from the 30k and 40k models for the Vi-K'Ho direction. Key differences are highlighted in bold.

ID	Source Vi	Reference K'Ho	30k Model Translation (Superior)	40k Model Translation (Inferior)
1	...gia đình **lại ép, lại mâu thuẫn**...	...cau tàm hìu **sồr, tam lơh tai**...	...cau tàm hìu **sồr, tam lơh tai**... xCOMET: 0.314	...hìu bơnhă **đơs wơl, đơs wơl**... xCOMET: 0.291
2	**khoảng** 88-98%...	**pơgăp** 88% tus 98%...	**pơgăp** 88% tus 98%... xCOMET: 0.529	**pơgăp bơh** 88 tus 98%... xCOMET: 0.454
3	....những **rủi ro** trong quá trình sinh nở...	...ală bơta **rềs àr** ờ niam tam tũ deh dùh...	...ală bơta **rềs àr** ờ niam tam tũ deh dùh... xCOMET: 0.327	...ală bơta **ờ niam** tàm tũ deh dùh... xCOMET: 0.428

Table 9: Examples demonstrating the model's strength in translating K'ho medical terminology into Vietnamese. The **bold text highlights** the core medical terms that the model translated accurately.

Source K'Ho	Reference Vi	Machine Translation
bŭ lah broă lơh sơnơm pa neh geh uă cồng nha tàm broă sơm bời kòn se deh ờ tơl nhai in, mơya khà do ngai sơlơ gơguh neh lơh gơtìp kal ke mờ bơta lòng tus gah lơh sơnơm bal mờ cau tàm hìu kòn se dê	mặc dù **y tế hiện đại** đã có nhiều thành tựu trong việc **cứu sống trẻ sinh non**, tuy vậy tỷ lệ này ngày càng tăng đặt ra gánh nặng và thách thức cho cả **ngành y tế** và gia đình của trẻ	mặc dù **y tế hiện đại** đã có nhiều thành tựu trong việc **cứu sống trẻ sinh non**, nhưng tỷ lệ này ngày càng tăng đã gây khó khăn và thách thức cho cả **ngành y tế** và gia đình của trẻ
tềng bòn drà buôn ma thuột, tam gùl tơng poh kơnhai 7 neh dờp geh cau kòp chơt bơh kòp mprơm lik mhàm lah oh h.o.q.n	tại phường buôn ma thuột, vào trung tuần tháng 7 đã ghi nhận **trường hợp tử vong** do **sốt xuất huyết** là em h.o.q.n	tại thị trấn buôn ma thuột, vào giữa tuần tháng 7 đã ghi nhận **trường hợp tử vong** vì **sốt xuất huyết** là bé h.o.q.n.
kòp suh sur châu phi lah kòp lợn gơtờp chơi sơbơng bơh birus lơh gơlik, tũ do ờ hềt geh bắc-xin rơcăng kriăn kờp	**dịch tả lợn châu phi** là **bệnh truyền nhiễm** nguy hiểm do virus gây ra, hiện chưa có **vắc-xin phòng bệnh**	**dịch tả lợn châu phi** là căn **bệnh truyền** nhiễm nguy hiểm do virus gây ra, hiện nay chưa có **vắc-xin phòng bệnh**
ală cau gơtìp kòp rang jơh ală ndrờm ờ hềt cĩt baccine rơcăng kòp	phần lớn những **người mắc bệnh sởi** đều chưa **tiêm vaccine phòng bệnh**	những **người mắc bệnh sởi** hầu hết chưa **tiêm vắc-xin phòng bệnh**

Table 10: Examples demonstrating the model's strength in translating Vietnamese medical terminology into K'Ho. The **bold text highlights** the core medical terms that the model translated accurately.

Source Vi	Reference K'Ho	Machine Translation
khi trẻ mắc sốt xuất huyết, phụ huynh cần theo dõi chặt chẽ các dấu hiệu của trẻ	tũ kơnòm dềt bơtờp **kòp mprơm gơlik mhàm**, cau mè bèp pal jăt sền nền ală **tềl sơñio** kòn dềt dê	tũ kơnòm dềt bơtờp **kòp mprơm gơlik mhàm**, mè bèp pal jăt sền nền ală **tềl sơñio** kòn dềt dê
bệnh nhân bị đau mắt đỏ đến khoa mắt, bệnh viện đa khoa vùng tây nguyên để khám và điều trị	cau kòp gơtìp **jê măt klồng** tus gah măt, **hìu sơnơm** uă gah tiah tây nguyên nàng **kham mờ sơm**	cau kòp **jê măt klồng** tus gah măt, **hìu sơnơm** uă gah tiah tây nguyên nàng **kham mờ sơm**
dấu hiệu nhận biết sớm tăng đường huyết	tềl sơñio đal gĩt gờñ **gơguh sơdàng mhàm**	tềl sơñio đal gĩt gờñ **gơguh sơdàng mhàm**
tiêu chảy là căn bệnh thường gặp ở trẻ em nhưng cũng dễ dự phòng	**kòp jê ndul gơ jroh** là **kòp mờng** gơbàn tàm kơnòm dềt mơya kung **bươn ngan sơl rơ-cang** lài pleh	**kòp jê ndul gơ jroh** là **kòp mờng** gơbàn tàm kơnòm dềt mơya kung **bươn ngan sơl rơ-cang** sơndră lài

Table 11: Examples of common error types in the K'ho-to-Vietnamese translation direction.

Error Type	**Source K'Ho**	**Reference Vi**	**Machine Translation**	**Analysis**
Lexical Error	...pah dõ añ **rê** den cau tàm hìu sồr...	...cứ mỗi lần tôi **xuất viện** thì gia đình lại ép...	...cứ mỗi lần tôi **về** thì gia đình lại ép...	The model provides a literal translation of rê as "về" (to return). While not entirely incorrect, it misses the specific medical context, where the more appropriate term is "xuất viện" (discharged from the hospital). This shows a weakness in selecting contextually nuanced terminology.
Critical Semantic Error	...cau kòp geh ală **bơta tàm jơh să jan** bè kòp duh să...	...bệnh nhân có **phản ứng toàn thân** như sốt...	...bệnh nhân có **các cơ quan toàn thân** như sốt...	This is a critical mistranslation. The model incorrectly translates bơta tàm jơh să jan ("things in the whole body," i.e., a systemic reaction) as "các cơ quan toàn thân" (all body organs). This error completely changes the medical meaning from a description of symptoms to an incorrect anatomical statement.
Hallucination	...mờ **pal pùt chèo** di pal mờ tơl nă cau kòp...	...và **phải dùng băng cuốn** phù hợp với từng người bệnh...	...và **phải đeo khẩu trang** phù hợp với từng trường hợp bệnh nhân...	The model fails to translate the source phrase pal pùt chèo (must use a wrapping band/cuff). Instead, it hallucinates a completely unrelated instruction: "phải đeo khẩu trang" (must wear a face mask). This type of fabrication is a significant failure, as it introduces incorrect medical advice.
Entity Error	...kờp jơh geh **1 rbô 11** nă cau kòp...	...tổng số tích lũy có **1.011** bệnh nhân...	...tổng số tích lũy có **1.11** bệnh nhân...	The model incorrectly translates the number 1 rbô 11 (one thousand and eleven) as "1.11". Mistranslating numerical data is a critical error, especially in a medical context where statistics, dosages, and patient counts must be precise.