Benchmarking Encoder–Decoder and Decoder-Only Models for Extremely Low-Resource English to Myaamia Translation

Yogesh Chaudhary*, Suman Bhunia*, Arthur Carvalho†, and Douglas Troy‡
* Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio
† Farmer School of Business, Miami University, Oxford, Ohio
‡ Myaamia Center, Miami University, Oxford, Ohio
Email: chaudhy@miamioh.edu, bhunias@miamioh.edu, carvalag@miamioh.edu, troyda@MiamiOH.edu

***Abstract*—Languages represent irreplaceable vessels of cultural knowledge and identity. Critically endangered Indigenous languages are languages spoken by Indigenous communities that face existential threats to their survival. This paper investigates the use of neural machine translation (NMT) frameworks to support the revitalization efforts of the Myaamia (Miami-Illinois) language, a critically endangered Algonquian language. In this study, we train several large language models (LLMs), including both encoder-decoder and decoder-only architectures, on English–Myaamia sentence pairs compiled from archival and dictionary sources. We compare model architectures under conditions of extreme data scarcity and evaluate the impact of post-training quantization on translation quality and on-device deployment. The paper provides a benchmark for translation data and how several LLM models perform. Robust evaluation is conducted using the SacreBLEU and chrF++ metrics. Experimental evaluation demonstrates that the mt5_large.pte model variant, an ExecuTorch-exported multilingual encoder–decoder baseline attained the highest translation quality, achieving a SacreBLEU score of 39.67 and a chrF++ score of 80.37 for English to Myaamia translation. This result paves the way for applying MT systems to the revitalization of other extremely low-resource languages. Finally, we present a fully offline Android prototype, highlighting both the promise and the current limitations of LLMs for community-focused endangered language revitalization.**

***Index Terms*—Fine-tuning, large language models, machine translation, natural language processing, transformer architectures.**

I. Introduction

LLMs such as Bidirectional Encoder Representations from Transformers Multilingual Bidirectional and Auto-Regressive Transformers (mBART) [1], mT5 [2], MarianMT [3], Generative Pre-Training (GPT) [4], and Llama [5] have transformed natural language processing (NLP) by learning universal language representations from massive unlabeled corpora. These models achieve strong performance on many downstream tasks, including machine translation (MT), even with limited labeled data [6]. However, their behavior in extremely low-resource, endangered languages remains underexplored.

Despite recent advances in NMT, significant barriers remain for endangered languages such as Myaamia, the language of the Miami People. Current state-of-the-art models are predominantly designed for and trained on high-resource languages, leaving a gap in our understanding of how these architectures perform under conditions of extreme data scarcity [7]. Specifically, the efficacy of transfer learning remains largely unexplored for languages that lack representation in massive multilingual corpora.

Furthermore, the comparative suitability of different architectures for this domain is underexplored. While encoder–decoder models like MarianMT, mT5, and mBART are standard for translation, they lack comprehensive evaluation in revitalization contexts. Simultaneously, the potential of adapting decoder-only LLMs, such as Llama, for specialized translation tasks remains largely experimental. Finally, the practical feasibility of deploying these models in real-world community settings is uncertain. There is a notable lack of documentation regarding how model quantization impacts translation quality for underrepresented languages and whether on-device inference can be achieved without compromising the utility of the tool. To address these challenges, and building on these ideas, this research investigates the development, training, and evaluation of LLMs for MT in extremely low-resource settings. Specifically, it examines whether transfer learning techniques can effectively enhance translation quality for underrepresented languages, with a focus on Myaamia. The study conducts a comparative analysis of widely used encoder–decoder models, including MarianMT, mT5, and mBART, to assess their ability to leverage multilingual pretraining for improved translation performance under limited data conditions [6], [8]–[10].

As LLM inferencing requires high-performance GPUs, common LLM architecture relies on a cloud computing framework where mobile devices offload the LLM computation to a server hosted on the internet cloud. However, most of the people who speak in endangered languages live in an area where internet connectivity is not available. Thus, we need a language translator that can run on mobile devices itself without requiring any internet connectivity. This research explores on-device inference through model quantization, evaluating how various quantization strategies impact translation accuracy and computational efficiency. A decoder-only architecture, Llama, given its wide on-device support, is also incorporated into

the study. The Llama model is fine-tuned for translation and benchmarked against encoder–decoder baselines to determine its suitability for low-resource MT.

Overall, this work aims to assess the feasibility of applying modern machine learning technologies to support language revitalization efforts. By evaluating model performance before and after quantization and by comparing architectures across translation tasks, the research seeks to determine whether contemporary LLM approaches can meaningfully contribute to the preservation and revitalization of endangered languages, in this study, Myaamia.

This study faces several critical challenges, foremost among them being the severe scarcity of parallel training data inherent to the Myaamia language and its rich, complex morphology. Furthermore, the application of LLMs to the revitalization of dormant languages is a nascent domain, with limited existing literature or benchmarks for comparison. Finally, the project addresses a technical gap in edge computing: while tooling for on-device inference of decoder-only architectures (such as Llama) has proliferated, the ecosystem for quantizing and deploying encoder–decoder models (such as mBART or mT5) remains comparatively under-documented and fragmented.

Despite the inherent challenges of data scarcity, recent literature has validated the feasibility of low-resource NMT through the adoption of pre-trained multilingual models and transfer learning strategies [7], [11], [12]. These advancements have proven effective across various endangered language case studies, demonstrating that modern NMT architectures can successfully support language revitalization efforts [9], [13], [14].

The study begins by detailing the data collection process, which leverages Myaamia digital resources, specifically the Indigenous Languages Digital Archive (ILDA) dictionary and archival materials, to construct a specialized English-Myaamia parallel corpus [15]. Following the data preprocessing pipeline, we systematically fine-tune and compare established encoder-decoder architectures (MarianMT, mT5, mBART) against the decoder-only Llama model. Crucially, Llama is integrated into this comparative analysis to capitalize on its robust on-device ecosystem. Also, its extensive library support and documentation offer a more streamlined pathway for offline mobile deployment compared to traditional encoder-decoder frameworks. The paper concludes with a comprehensive analysis of the trade-offs between translation quality and the practical feasibility required for community-facing applications.

The contributions of this paper are threefold:

1) We present, to the best of our knowledge, the first systematic comparison of variants of MarianMT, mT5, mBART, and Llama for the Myaamia language.
2) We evaluate the effect of post-training quantization on translation quality for an endangered, extremely low-resource language.
3) We implement a fully offline mobile prototype that performs English to Myaamia translation on-device, demonstrating the feasibility of community-facing deployment. The source code is publicly available in an open-source GitHub repository [16].

II. Background

In this section, we discuss the context of the study. First, we discuss the language that we are focusing on: Miamia, then we discuss what machine translation is, followed by related studies and the importance of on-device inferencing.

A. Myaamia Language

This study situates the development of a English-Myaamia NMT tool within the intersecting domains of endangered language revitalization and neural language technologies. The Myaamia language is a critically endangered Algonquian language that nearly ceased to be spoken due to displacement and assimilation policies [8]. Recent revitalization efforts, led by the Miami Tribe of Oklahoma and the Myaamia Center at Miami University, have focused on reclaiming both linguistic and cultural knowledge through digital archives and educational initiatives. From an NLP perspective, Myaamia is both an endangered and a extremely low-resource language: it lacks large digital corpora and established NLP tools, but it benefits from curated bilingual resources such as the Indigenous Languages Digital Archive (ILDA) and the Miami-Peoria dictionary [15].

B. Machine Translation

Machine Translation (MT) refers to the use of computer-based systems to translate text from one natural language to another, with or without human involvement. Bilingual and multilingual translation represent two fundamental paradigms in MT, each with distinct methodologies, applications, and implications for linguistic resource availability. Bilingual translation refers to systems explicitly designed to translate between a single source-target language pair, such as English-French. In contrast, multilingual translation involves models capable of handling multiple languages within a unified architecture, either through shared parameters or language-specific modules [17], [18].

Since the recent boom in machine learning driven by transformer architectures, pretrained transformer-based models have significantly advanced the field of MT by enabling high-quality translations across a wide range of languages. These models leverage large-scale multilingual corpora and transfer learning, making them particularly effective in low-resource language scenarios [6], [13]. Among the many models developed, this research focuses on several prominent architectures such as MarianMT, mT5, mBART, and Llama due to their relevance to low-resource language translation and their strong community/technical support [6], [19].

MarianMT is a family of NMT models that leverage the MarianNMT framework, an efficient and scalable NMT engine originally developed by the University of Edinburgh [3], [20], [21]. mBART (Multilingual BART) is a multilingual sequence-to-sequence model developed by Facebook AI [1]. It is based

on the BART [22] architecture. While BART was only pre-trained for English, mBART is pretrained as a denoising auto-encoder on large-scale monolingual corpora in different sets of languages. The key terminology for mBART includes "multilingual denoising pre-training," where the model learns to reconstruct original texts from corrupted (noised) inputs across multiple languages, and "sequence-to-sequence" (Seq2Seq) learning, which means the model is designed to generate a target sequence (e.g., a translation) from a source sequence.

The mT5 model [2] is a multilingual extension of Google's T5 [23], pre-trained on data from more than 101 languages and available in five parameter scales: *mT5-Small* ($\approx$ 300M parameters), *mT5-Base* ($\approx$ 580M), *mT5-Large* ($\approx$ 1.2B), *mT5-XL* ($\approx$ 3.7B), and *mT5-XXL* ($\approx$ 13B). Its architecture and training methodology closely follow those of the original T5 model.

Large Language Model Meta AI (Llama) is a family of LLMs introduced by Meta AI in February 2023 [5]. The models span a wide range of parameter sizes, from 1 billion to 2 trillion. While the initial release was limited to a foundation model, beginning with Llama 2 Meta AI also provided instruction-tuned variants alongside the base models [5]. With the introduction of Llama 3 and a dedicated public web interface, Meta integrated Llama-based virtual assistant capabilities into Facebook and WhatsApp in selected regions. The most recent version, Llama 4, was released in April 2025; its behemoth variant contains a massive 2 trillion total parameters. [5], [24].

C. Related Work

Previous research has demonstrated promising potential in leveraging modern approaches, such as NLP and MT, for low-resource and language revitalization efforts. Foundational surveys by Hedderich [7] and Kozhirbayev [6] highlight the diverse strategies available for addressing data scarcity, with both emphasizing the critical role of multilingual pre-trained models in cross-lingual transfer.

A particularly promising avenue identified in this literature is the use of pre-trained multilingual sequence-to-sequence models to overcome the lack of parallel data [11]. For instance, Chen and Abdul-Mageed [12] demonstrated that fine-tuning bilingual and multilingual pre-trained models on Spanish–Indigenous language pairs outperforms traditional approaches, achieving state-of-the-art results. Similarly, researchers have successfully shown that fine-tuning large models, such as mBART50, significantly enhances translation quality for low-resource pairs [6], [13].

Beyond theoretical improvements, these methodologies have been applied to critically endangered languages with tangible success. Recent studies on Cherokee [14] and Colombian Indigenous languages [25] provide empirical evidence that neural machine translation can effectively support cultural preservation. In the context of the Ainu language, Miyagawa [9] achieved a robust 32.9 BLEU for Japanese(Jpn)→Ainu(Ain) translation using MarianMT. Subsequent research by Igarashi and Miyagawa [10] on multilingual models, specifically the mt5-base model, showed a significant increase in performance, achieving scores of 31.83 for Ain to Jpn and 39.06 for Jpn to Ain.

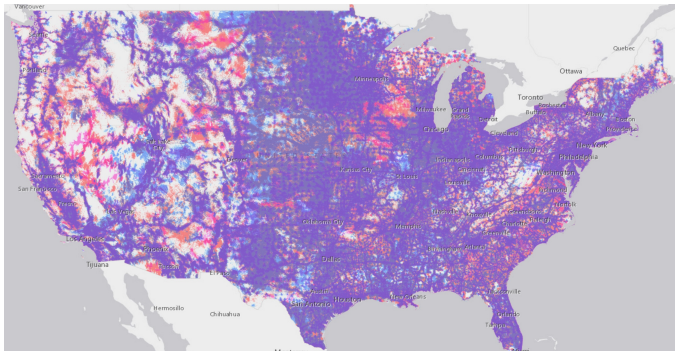


Fig. 1: A map from "Federal Communications Commission" showcasing the area with and without internet coverage by three major US cellular networks (AT&T, T-Mobile, and Verizon) [26]. Note the light gray regions where no cellular data coverage is available. Most of the indigenous tribes inhabit these coverage gaps.

Collectively, this literature builds the base for this research, providing a strong foundation for using MT and NLP technologies to preserve endangered languages like Myaamia.

D. On-Device Inference

Integrating LLMs into mobile applications transforms communication by providing instantaneous, context-aware NLP capabilities directly on personal devices. As smartphones become ubiquitous, embedding NMT models within mobile applications enables real-time translation for language learning, travel, and cross-cultural interactions, crucially extending these benefits to speakers of low-resource and endangered languages [27], [28]. Such on-device translation tools also serve as valuable language revitalization resources, enabling users to learn and engage with endangered languages like Myaamia.

Offline translation support is of particular importance. It ensures accessibility in areas with limited or unreliable internet connectivity. Fig. 1 illustrates that, even today, coverage gaps persist across parts of the United States, including regions served by one of the largest mobile network operators by subscriber count. This observation underscores the need for a fully offline translation system. Moreover, offline processing enhances user privacy by ensuring that sensitive text is handled locally. For many Indigenous communities, including the Miami people, language is sacred and deeply tied to cultural identity and heritage. As a result, community members may be reluctant to transmit linguistic data to external servers or allow it to be made public without explicit community approval. Local, on-device translation therefore aligns more closely with community preferences for privacy, data sovereignty, and cultural stewardship. This is particularly important in educational and community settings, where data security and independence from centralized infrastructure are key priorities.

III. Proposed Framework

To address the challenges of low-resource machine translation and on-device deployment for Myaamia, this study proposes a holistic experimental framework that integrates transfer learning with aggressive model compression. This section describes the data preparation pipeline, supervised and instruction-based fine-tuning strategies, post-training quantization for on-device inference, comparative evaluation metrics, and the implementation of a mobile prototype.

A. Proposed Approach

Our approach moves beyond simple model training; it systematically evaluates the trade-offs between architectural complexity (Encoder–Decoder vs. Decoder-only), translation quality, and computational efficiency. By benchmarking state-of-the-art multilingual models against fine-tuned LLMs, we aim to identify the optimal configuration for revitalizing endangered languages in resource-constrained environments. The specific components of our technical pipeline are detailed below.

1) Data Preparation and Standardization: The dataset used in this study consists of aligned English-Myaamia translation pairs compiled from diverse resources. Our dataset consists of four tab-separated (TSV) files summarized in Table I. The first three were sourced from the ILDA dictionary, which contains bilingual English-Myaamia texts curated by the Myaamia Center, including community-relevant narratives and validated translations from "Myaamia Neehi Peewaalia Kaloosioni mahsinaakani (A Miami-Peoria Dictionary)" [15]. The storybook parallel data was extracted from "Myaamia Neehi Peewaalia Aacinmoona Neehi Aalhsoohkaana (Myaamia and Peoria Narratives and Winter Stories)", the first published collection of Miami-Illinois texts. This book includes 44 narratives, with 28 presented in the Myaamia, Peoria, or Wea dialects alongside English translations edited by David J. Costa. An example Myaamia sentence and its English translation are shown below.

```
English: 'No, it is not evening.
It is still noon.'

Myaamia: 'moohci, alaakowihsiinoowi.
eehkwa maayaahkweeci'
```

To evaluate the direct adaptability of pre-trained models to raw low-resource data, minimal manual pre-processing was applied. We did not perform custom morphological segmentation or train new tokenizers from scratch. Instead, we leveraged the model-specific tokenization pipelines inherent to each architecture. For the encoder–decoder models (MarianMT, mT5, mBART) and the decoder-only model (Llama), the raw text pairs were processed using their respective pre-trained tokenizers. This approach ensures that the input data is mapped directly to the models' existing embedding spaces, allowing the fine-tuning process to adjust the weights based on the standard input representations used during the models' original pre-training. This design choice reflects real-world language revitalization efforts, which are often led by linguists and community practitioners with limited exposure to the technical intricacies of model architectures, and therefore emphasizes simple, out-of-the-box fine-tuning using raw data.

TABLE I: Summary of the Dataset Files

File Name	Entries	Description
`sentences`	1,935	Parallel sentence pairs collected from ILDA.
`command forms`	3,819	Command forms sourced from ILDA, including dictionary-validated translations.
`basic forms`	12,892	Basic forms extracted from ILDA, including dictionary-validated translations.
`story book`	455	Storybook sentences from *Myaamia and Peoria Narratives and Winter Stories*.

2) Supervised Fine-Tuning of Encoder–Decoder Architectures: We employ a transfer learning approach using three distinct encoder–decoder architectures: MarianMT, mT5, and mBART. These models' variants, listed in Table II, were selected for their proven efficacy in multilingual contexts.

MarianMT: Utilized for its lightweight architecture, serving as a baseline for efficiency [3].

mT5 & mBART: Utilized to test the hypothesis that massive multilingual pre-training (on 100+ languages) provides a "knowledge bridge" that aids in learning unseen languages like Myaamia [6], [11]. We fine-tune the weights of these pre-trained models specifically on our Myaamia corpus, optimizing for the sequence-to-sequence translation objective.

3) Instruction Tuning of Decoder-Only LLM (Llama) via Parameter-Efficient Fine-Tuning (Low-Rank Adaptation, LoRA): To evaluate the generative capabilities of modern LLMs, we incorporate the Llama architecture. Unlike traditional MT models, Llama, specifically *meta-llama/Llama-3.2-1B-Instruct*, is a decoder-only model trained using a causal language modeling objective. We adapt this model for translation through instruction tuning using parameter-efficient fine-tuning with LoRA, keeping approximately 1% of the model parameters trainable, conditioning the model to treat translation as a text generation task. This is achieved by structuring the training data as instruction–response pairs with explicit prompts (e.g., "Translate the following English text to Myaamia:"), enabling the model to learn translation behavior within its existing generative framework.

4) Post-Training Quantization (PTQ) for On-Device Inference: A critical contribution of this work is the evaluation of model compression. We apply PTQ to convert model weights from standard floating-point precision (FP32 or FP16) to lower-precision formats (4-bit).

5) Comparative Evaluation Metrics: To ensure an objective assessment of translation quality, we utilize standard automated metrics, SacreBLEU and chrF++ (Character n-gram F-score Plus Plus). chrF++ is particularly relevant for this study as it correlates better with human judgment for morphologically complex and low-resource languages than word-level metrics. We calculate these metrics across most model variations (uncompressed vs. quantized) to provide a granular analysis of performance trade-offs.

B. Prototype Implementation and Mobile Integration

To demonstrate the feasibility of bringing revitalization tools directly to the Myaamia community, we developed a functional mobile prototype. While encoder–decoder architectures are traditional standards for machine translation, we selected the decoder-only Llama architecture for the deployment phase. This decision was driven by the greater maturity of the on-device inference ecosystem for decoder-only models, which currently provides more robust tooling for model compression and mobile execution than is available for encoder–decoder counterparts.

The implementation pipeline consists of three distinct stages: Model Adaptation, Format Conversion and Quantization, and Mobile Application Integration. These are discussed in details below.

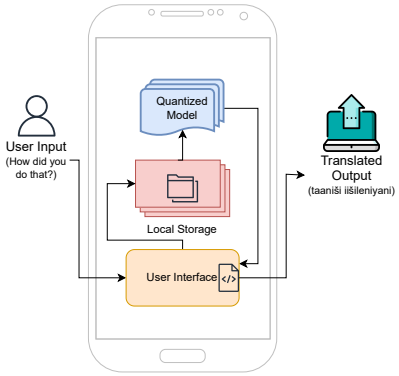
1) Model Adaptation and Prompt Engineering: The Llama model, *meta-llama/Llama- 3.2-1B-Instruct*, was fine-tuned using a supervised learning approach on the prepared English to Myaamia parallel corpus. Unlike standard encoder-decoder models, Llama operates as a text-continuation engine. Therefore, we structured the training data using specific prompt templates (e.g., "System: You are an expert translator.", "User: Translate the following English text to Myaamia: *[Input Segment]*") to condition the model for the translation task. This instruction-tuning aligns the model's generative capabilities with the specific requirements of language translation.

2) Conversion and Quantization Strategy: To transition from a research environment to a mobile runtime, we evaluated two emerging inference formats:

- ExecuTorch (.pte): PyTorch's native solution for edge-device deployment. In this work, the Optimum-ExecuTorch library was used to export the models [29].
- GGUF (GPT-Generated Unified Format): A binary format designed for fast loading and memory mapping on consumer hardware. The llama.cpp library was used for both model export and inference [30].

While we explored the .pte format for its theoretical compatibility with the PyTorch ecosystem, we prioritized the GGUF format for the final prototype due to its extensive community support [30]. Following conversion, we applied PTQ (Q4_K_M) to the f16 GGUF model. This step reduced the model's precision to lower-bit representations (4-bit), significantly compressing the file size and reducing memory bandwidth requirements without requiring retraining of the neural network [31].

3) Android Integration via React Native: The client-side application was developed using React Native, enabling a cross-platform codebase. To interface the JavaScript layer with the underlying C++ inference engine, we employed the llama.rn library, which initializes the GGUF model and performs inference directly on the Android device's CPU/GPU. The GGUF model was deployed to the device via ADB, copied to the app's private storage, and the temporary file was removed. It was then integrated and loaded within the React Native application code for on-device inference.



(a) On-Device Architecture.

(b) UI of the language translator mobile application.

Fig. 2: Proposed on-device architecture and user interface of the system.

The resulting prototype enables fully offline inference. Users interact with a standard chat interface, where English text is input, processed locally by the quantized Llama model, and returned as Myaamia translation output. This design preserves data sovereignty and ensures reliable operation in areas with limited or no internet connectivity. The on-device architecture and user interface are illustrated in Fig. 2.

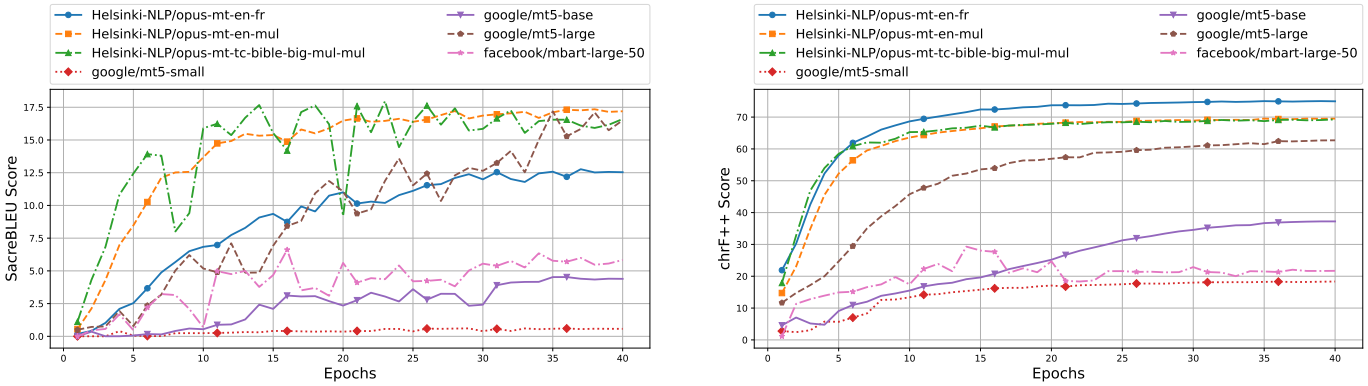
IV. Evaluation

The evaluation of NMT systems is a critical component in the development and deployment of machine translation technologies. Evaluation metrics provide quantitative and qualitative means to assess how well a system translates text from a source to a target language, guiding model selection, benchmarking, and iterative improvements. Evaluation metrics are also categorized into two parts Automatic metric and Human based evaluation [32], [33]. Automated metrics were primarily used in this study because they provide a standardized, efficient, and reproducible means of evaluating translation performance.

A. Performance Metrics

The evaluation incorporates two widely used automated metrics to assess translation quality.

1) SacreBLEU: SacreBLEU is an automated evaluation metric widely adopted for machine translation research [34]. It is a standardized implementation of BLEU (Bilingual Evaluation Understudy) [35], designed to provide consistent and reproducible scores across studies by controlling tokenization and text normalization settings. SacreBLEU measures n-gram precision by comparing the overlap between the machine-generated translation and one or more human reference translations, and it applies a brevity penalty to discourage overly short outputs. This metric is valued for its objectivity and scalability, and has become the default choice for benchmarking NMT systems due to its transparent, comparable score reporting [32], [33].



(a) SacreBLEU scores over 40 epochs.

(b) chrF++ scores over 40 epochs.

Fig. 3: Evaluation performance comparison across different models.

TABLE II: Training performance (runtime and loss) and evaluation results on the validation set (SacreBLEU and chrF++) across models for English to Myaamia translations.

Model	Runtime (s)	Loss	SacreBLEU	chrF++
Helsinki-NLP/opus-mt-en-fr	4752.70	0.38	12.77	75.02
Helsinki-NLP/opus-mt-en-mul	5920.89	0.45	17.35	69.52
Helsinki-NLP/opus-mt-tc-bible-big-mul-mul	7083.66	0.32	17.97	69.27
google/mt5-small	8755.20	3.82	0.60	18.34
google/mt5-base	13961.44	2.52	4.53	37.23
google/mt5-large	37340.53	1.04	17.19	62.70
facebook/mbart-large-50	33988.30	0.17	6.61	29.31
meta-llama/Llama-3.2-1B-Instruct	531.60	0.60	1.60	22.89

2) chrF++: chrF++ is an automatic metric for evaluating machine translation output that combines character n-gram precision and recall with additional word n-gram features. It computes an F-score by averaging over all character and word n-grams, using a default character n-gram order of 6 and a word n-gram order of 2. The final score is obtained using arithmetic mean averaging across n-gram orders. [36].

We specifically included chrF++ due to the morphological characteristics of Myaamia. As Myaamia is a polysynthetic language where a single word can contain complex grammatical information, word-level metrics like BLEU can excessively penalize minor morphological errors. chrF++ operates at the character n-gram level, providing a more granular and improved correlation with human judgment for morphologically rich languages [36].

B. Evaluation Setup

To ensure a rigorous and reproducible assessment of model performance, we implemented a standardized evaluation pipeline. The experimental design focuses on measuring semantic fidelity and convergence stability using industry-standard metrics adapted for low-resource environments.

1) Data Partitioning: Given the limited size of the available corpus, we adopted a 70/30 training–validation split instead of the conventional train-validation-test partition. This relatively large validation split, compared to standard high-resource settings, was chosen to ensure that the evaluation results were representative and robust against outliers, a common pitfall in low-resource machine translation [7]. Also, this 2-way splitting approach is commonly adopted in endangered and extremely low-resource language research where creating a three-way split would leave insufficient data for model training [9], [14]. However, unlike the other models trained for 40 epochs on the 70/30 split, the Llama model was trained on a 90/10 split for 3 epochs to accommodate its larger architecture and reduce training time.

2) Metrics Implementation: Consistent with our training pipeline, we employed the Hugging Face *Evaluate* library to initialize and compute the evaluation metrics.

3) Evaluation Protocol & Model: Model performance was monitored and recorded throughout the fine-tuning process, with evaluation performed at the end of each epoch. The validation set was used to monitor training progress at each epoch and to select the best-performing checkpoint. This epoch-level evaluation enabled monitoring of training stability, identification of convergence behavior, and early detection of overfitting, which is a critical concern when fine-tuning large-parameter models on small datasets. All reported metrics therefore reflect validation set performance.

Fig. 3 shows the SacreBLEU and chrF++ scores over 40 training epochs. It can also be observed that the model converges after approximately 10 epochs. Table II presents a comparison of fine-tuning runtimes and resulting training losses with highest achieved validation scores across all evaluated architectures.

Interestingly, the quantized model outperformed the original full-precision model across evaluation metrics, as shown in Table III. This behavior can be attributed either to the regular-

TABLE III: Evaluation results of exported/quantized models on English to Myaamia translation. *The model used for on-device inference on Realme 7 pro (Android 12).

Model	Size (MiB)	SacreBLEU	chrF++	Infer. (s)
mt5_large_Q4 _K_M.gguf	757.79	38.60	79.92	–
llama3_1b_finetune _Q4_K_M.gguf*	762.81	1.05	21.13	11.60
llama-3.2-1b-8da4w-8w.pte	914.45	0.72	17.92	–
mt5_large.pte	4367.83	39.67	80.37	–

ization effect introduced by quantization [37], which improves generalization by reducing overfitting, or to implementation differences, such as better-tuned inference libraries, superior default sampling parameters, or more efficient tokenization. To investigate this, we converted the fine-tuned *google/mt5-large* model to a 16-bit precision ExecuTorch (.pte) format and evaluated it on the same validation set used for the *google/mt5-large* model. The model achieved a SacreBLEU score of 39.67 and a chrF++ score of 80.37, suggesting that the observed improvement was not due to the regularization effect of quantization but rather the result of differences in implementation. However, this remains an active area of research and further investigation is required to validate this finding.

Based on Table II and Table III the following observations can be made :

- Best Performers: Among the non-exported models, "Helsinki-NLP/opus-mt-tc-bible-big-mul-mul" (Multilingual) achieved the highest SacreBLEU score of 17.97, while "Helsinki-NLP/opus-mt-en-fr" obtained the highest chrF++ score of 75.02. Among the exported models, the quantized "mt5_large_Q4_K_M.gguf" achieved a SacreBLEU score of 38.60 and a chrF++ score of 79.92, whereas the non-quantized "mt5_large.pte" ExecuTorch model achieved a SacreBLEU score of 39.67 and a chrF++ score of 80.37.
- Llama Performance: With SacreBLEU and chrF++ scores of 1.6 and 22.89, respectively, the "meta-llama/Llama-3.2-1B-Instruct" model illustrates that standard decoder-only LLMs perform poorly on low-resource endangered language translation relative to specialized encoder-decoder models.

Moreover, Tables II and III report three complementary performance indicators beyond translation quality of the best model checkpoint. Training runtime reflects the computational cost of fine-tuning each architecture: the Llama model completes training in 531 s due to the parameter efficiency of LoRA, whereas the larger mT5-large requires 37,341 s. Training loss measures convergence by assessing the model's error on the training data during training. Finally, on-device inference time, measured on a Realme 7 Pro smartphone running Android 12, shows that the quantized Llama model requires an average of 11.6 s per sentence, which highlights current limitations for real-time interaction.

V. Conclusion

Preserving and revitalizing critically endangered languages requires translation technologies that are both effective and accessible to the communities that use them. This study investigated the feasibility of NMT systems to support the revitalization of Myaamia, a critically endangered Indigenous language of the Miami people. In particular, the research examined the extent to which multilingual pretrained models can deliver effective translation performance in unseen low-resource settings, as well as their practicality for on-device deployment in community-oriented use. By benchmarking encoder–decoder architectures against a decoder-only LLM, we evaluated their suitability for extremely low-resource translation settings.

The experimental results indicate that multilingual encoder–decoder models, such as Google's mT5 and Helsinki-NLP, achieve promising translation performance. Notably, the ExecuTorch-exported mT5 variant attained substantially higher quality (SacreBLEU: 39.67, chrF++: 80.37) compared to the fine-tuned Llama baseline (SacreBLEU: 1.60, chrF++: 22.89) and its GGUF-exported variant (SacreBLEU: 1.05, chrF++: 21.13). This showcases that, while Llama models perform well on general-purpose tasks such as code generation or question answering, they exhibit limited effectiveness in low-resource, task-specific applications such as machine translation. In contrast, encoder–decoder models demonstrate superior performance and data efficiency in low-resource settings.

Despite the observed performance gap, the development of a React Native prototype demonstrates the technical feasibility of running quantized inference locally on mobile devices. Although current performance limits restrict immediate real-world deployment, these results suggest that on-device translation is a promising direction rather than an impractical one. Further research is needed to improve usability and translation quality for community adoption.

Future work should focus on leveraging encoder–decoder architectures for efficient on-device inference, expanding parallel data resources to improve model generalization, and further investigating evaluation discrepancies between Hugging Face models and their exported counterparts. Additionally, incorporating human evaluations from members of the Miami community will be essential to ensure that translation outputs are linguistically accurate, culturally appropriate, and supportive of ongoing revitalization efforts.

Overall, this research contributes to the broader study of NMT for low-resource and unseen languages [6], [7], [13], highlighting both the limitations and potential of modern NMT techniques in supporting language revitalization and community-centered applications.

REFERENCES

[1] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[2] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 483–498, 2021.

[3] M. Junczys-Dowmunt *et al.*, "Marian: Fast neural machine translation in C++," *CoRR*, vol. abs/1804.00344, 2018.

[4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI Blog*, vol. 1, no. 8, 2018.

[5] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[6] Z. Kozhirbayev, "Enhancing neural machine translation with fine-tuned mbart50 pre-trained model: An examination with low-resource translation pairs," *Ingenierie des Systemes d'Information*, vol. 29, no. 3, p. 831, 2024.

[7] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2545–2568, 2021.

[8] D. J. Costa, *The Miami-Illinois Language*. Studies in the Native Languages of the Americas, University of Nebraska Press, 2003.

[9] S. Miyagawa, "Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan," in *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pp. 120–124, 2023.

[10] R. Igarashi and S. Miyagawa, "Enhancing neural machine translation for ainu-japanese: A comprehensive study on the impact of domain and dialect integration," in *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pp. 413–422, 2024.

[11] E.-S. A. Lee, S. Thillainathan, S. Nayak, S. Ranathunga, D. I. Adelani, R. Su, and A. D. McCarthy, "Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?," *arXiv preprint arXiv:2203.08850*, 2022.

[12] W.-R. Chen and M. Abdul-Mageed, "Improving neural machine translation of indigenous languages with multilingual transfer learning," *arXiv preprint arXiv:2205.06993*, 2022.

[13] A. L. Tonja, H. H. Nigatu, O. Kolesnikova, G. Sidorov, A. Gelbukh, and J. Kalita, "Enhancing translation for indigenous languages: Experiments with multilingual models," *arXiv preprint arXiv:2305.17406*, 2023.

[14] S. Zhang, B. Frey, and M. Bansal, "How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language," *arXiv preprint arXiv:2204.11909*, 2022.

[15] M. Center, "Digital resources." https://miamioh.edu/centers-institutes/myaamia-center/research/digital-resources.html, 2025. Accessed: December 19, 2025.

[16] Yogesh, Suman, "Linguabridge." https://github.com/sbhunia/LinguaBridge/tree/main, 2025. Accessed: 2025-12-18.

[17] M. D. Okpor, "Machine translation approaches: issues and challenges," *International Journal of Computer Science Issues (IJCSI)*, vol. 11, no. 5, p. 159, 2014.

[18] W. J. Hutchins, "Machine translation: A brief history," in *Concise history of the language sciences*, pp. 431–445, Elsevier, 1995.

[19] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *arXiv preprint arXiv:2006.07264*, 2020.

[20] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vázquez, and S. Virpioja, "Democratizing neural machine translation with opus-mt," *Language Resources and Evaluation*, vol. 58, no. 2, pp. 713–755, 2024.

[21] Hugging Face, "Marianmt — hugging face transformers documentation." https://huggingface.co/docs/transformers/model_doc/marian, 2024. Accessed: December 19, 2025.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[24] Meta AI, "Llama 4: Multimodal intelligence." https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed: December 19, 2025.

[25] J. Prieto, C. Martinez, M. Robles, A. Moreno, S. Palacios, and R. Manrique, "Translation systems for low-resource colombian indigenous languages, a first step towards cultural preservation," in *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pp. 7–14, 2024.

[26] "4G LTE Coverage as of May 15, 2021 (AT&T Mobility, T-Mobile, UScellular, Verizon)." https://fcc.maps.arcgis.com/apps/webappviewer/index.html?id=6c1b2e73d9d749cdb7bc88a0d1bdd25b. Accessed: December 19, 2025.

[27] Y. Lin, X. Wang, Z. Zhang, M. Wang, T. Xiao, and J. Zhu, "Mobilenmt: Enabling translation in 15mb and 30ms," *arXiv preprint arXiv:2306.04235*, 2023.

[28] J. Xu, Z. Li, W. Chen, Q. Wang, X. Gao, Q. Cai, and Z. Ling, "On-device language models: A comprehensive review," *arXiv preprint arXiv:2409.00088*, 2024.

[29] Hugging Face, "Optimum ExecuTorch: Optimize and deploy hugging face models with executorch." https://github.com/huggingface/optimum-executorch, 2025. accessed December 19, 2025.

[30] ggml-org, "llama.cpp: Port of llama inference in c/c++." https://github.com/ggml-org/llama.cpp, 2023. Accessed: December 19, 2025.

[31] I. Chung, B. Kim, Y. Choi, S. J. Kwon, Y. Jeon, B. Park, S. Kim, and D. Lee, "Extremely low bit transformer quantization for on-device neural machine translation," *arXiv preprint arXiv:2009.07453*, 2020.

[32] E. Chatzikoumi, "How to evaluate machine translation: A review of automated and human metrics," *Natural Language Engineering*, vol. 26, no. 2, pp. 137–161, 2020.

[33] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim, "A survey on evaluation metrics for machine translation," *Mathematics*, vol. 11, no. 4, p. 1006, 2023.

[34] M. Post, "A call for clarity in reporting bleu scores," *arXiv preprint arXiv:1804.08771*, 2018.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[36] M. Popović, "chrF++: words helping character n-grams," in *Proceedings of the second conference on machine translation*, pp. 612–618, 2017.

[37] S. Fang, W. Ding, A. Mastropaolo, and B. Xu, "Smaller= weaker? benchmarking robustness of quantized llms in code generation," *arXiv preprint arXiv:2506.22776*, 2025.