Anchoring the Judge: Curriculum-Based Adaptation and Reference-Anchored MQM for LLM-Based Machine Translation of an Unseen Low-Resource Language - A Case of Nupe

Umar Baba Umar[*] **Sulaimon Adebayo Bashir** **Abdulmalik Danlami Mohammed**
Federal University of Technology Minna, Nigeria
{umar.umar, bashirsulaimon, drmalik}@futminna.edu.ng

Abstract

Adapting large language models (LLMs) for machine translation has shown strong performance in low-resource languages; however, their effectiveness for *unseen, extremely low-resource languages* remains largely unexplored. We present **NupeMT-QLoRA**, a curriculum-based adaptation framework for the Nupe–English language pair. Our approach employs a two-stage QLoRA fine-tuning strategy: (i) initial training on 34k noisy parallel sentence pairs, followed by (ii) continued fine-tuning on a smaller, cleaner set of 12k bidirectional parallel sentences with explicit translation-direction tags. This staged curriculum stabilizes optimization and improves robustness under severe data scarcity.We further identify a reliability crisis in existing automatic evaluation metrics for unseen languages. Popular LLM-based judges such as GEMBA and xCOMET exhibit weak correlation with human judgments (Kendall's $\tau \approx 0.21$) and low inter-rater reliability (Fleiss' $\kappa \approx 0.27$), largely due to fluency bias. To address this, we propose **Ref-Anchor-MQM**, a reference-anchored evaluation protocol that forces the judge to extract Key Semantic Units from a human reference before scoring.Experimental results show that NupeMT-QLoRA substantially outperforms NLLB-200, improving chrF++ from 22.73 to 41.10, while Ref-Anchor-MQM achieves significantly higher alignment with human evaluation ($\tau = 0.71$). Our framework provides a scalable pipeline for adapting and evaluating LLMs on languages with zero prior representation.

1 Introduction

Recent advancements in generative, decoder-only Large Language Models (LLMs) have led to substantial progress in machine translation (MT). Models such as LLaMA-2 7B (Touvron et al., 2023b), GPT-3.5, and GPT-4 (Brown et al., 2020), when augmented with instruction tuning and domain adaptation, have demonstrated competitive and in some cases superior performance compared to conventional neural machine translation (NMT) systems on high-resource language pairs. More recent variants, including ALMA-R (Xu et al., 2024b) and TowerInstruct (Alves et al., 2024), as well as fine-tuned LLMs evaluated on the FLORES-200 benchmark, have surpassed state-of-the-art multilingual models such as NLLB-200 (NLLB Team et al., 2022) and commercial translation systems in selected high-resource domains.

This progress has also been supported by parameter-efficient fine-tuning techniques such as QLoRA (Dettmers et al., 2024a), which reduce memory requirements and enable practical adaptation of large language models. These techniques have facilitated experimentation in low-resource language settings. Recent studies report promising results for languages such as Marathi (Khade et al., 2024), Urdu (Hussain et al., 2025), and several African and Finno-Ugric languages (Adelani et al., 2022; Purason et al., 2024; Lankford et al., 2025; Oladipupo, 2025). However, these successes predominantly concern languages for which LLMs already possess latent knowledge acquired during large-scale multilingual pre-training. In contrast, the applicability of LLMs to *unseen and extremely low-resource languages* where the language is effectively absent from pre-training corpora remains largely unexplored.

In such settings, model scale alone cannot compensate for extreme data scarcity. Fine-tuning on limited and noisy parallel data often leads to optimization instability, degraded generalization, and brittle translation behavior. Recent work further indicates that LLMs are particularly sensitive to noise and misalignment in low-resource training data,

[*] To facilitate further research in African NLP and LLM evaluation, we will release the NupeMT-QLoRA adapters, our curated datasets, and the Ref-Anchor-MQM evaluation code on GitHub `https://github.com/futminna-ai-lab` upon publication.

frequently producing fluent yet semantically incoherent translations (Singh et al., 2024; Pan et al., 2025). These challenges are especially pronounced for languages such as **Nupe**, a Niger–Congo language spoken by millions but virtually absent from contemporary multilingual benchmarks and pretraining datasets.

Beyond modeling challenges, evaluation presents an equally critical bottleneck. Standard automatic metrics, including BLEU and chrF++, are increasingly criticized for their limited ability to capture semantic adequacy, pragmatic meaning, and cultural nuance—qualities that LLM-based systems often emphasize. In response, the field has moved toward *LLM-as-a-judge* paradigms, such as GEMBA (Kocmi and Federmann, 2023) and xCOMET (Guerreiro et al., 2024), which leverage large models to assess translation quality. However, our preliminary investigations reveal systematic reliability issues when these metrics are applied to unseen languages. For Nupe, we observe poor inter-rater agreement (Fleiss' $\kappa \approx 0.23$–0.27) and w... weak correlation with human Direct Assessment (DA) scores (Kendall's $\tau \approx 0.18$–0.21), computed over 106 randomly sampled test sentences.We hypothesize that this degradation arises because the underlying neural representations of both the judge models and learned metrics were never exposed to the language during pre-training. As a result, evaluation systems exhibit a pronounced *fluency bias*, rewarding grammatically plausible output even when it diverges substantially from the source meaning.

In this work, we introduce **NupeMT-QLoRA**, a framework designed to address both adaptation and evaluation challenges for extremely low-resource, unseen languages. Our approach consists of two key innovations. First, we propose a *curriculum-based adaptation strategy* that gradually transitions from large, noisy parallel corpora (e.g., web-mined and pivot-generated data) to smaller, high-quality bidirectional datasets. This staged training regime stabilizes optimization and mitigates the impact of noise and misalignment during QLoRA fine-tuning. Second, we propose **Ref-Anchor-MQM**, a novel reference-anchored evaluation protocol. Unlike existing LLM-based metrics, Ref-Anchor-MQM explicitly forces the judge to extract *Key Semantic Units (KSUs)* from a human reference prior to scoring, thereby anchoring evaluation in ground-truth meaning and reducing fluency-driven hallucinations in unseen language settings.Human reference translations were produced by native Nupe speakers fluent in both Nupe and English and were manually verified for consistency prior to evaluation. The objective of this work is twofold: (1) to stabilize LLM adaptation for an unseen, extremely low-resource language using curriculum-based QLoRA fine-tuning, and (2) to improve evaluation reliability through reference-anchored semantic decomposition.

Our contributions are threefold:

- We present the first curriculum-based QLoRA adaptation of LLaMA-2 for the Nupe–English language pair, demonstrating improved training stability and translation robustness in an extremely low-resource scenario.
- We introduce Ref-Anchor-MQM, a reference-anchored evaluation algorithm, and end empirically show improved reliability in the Nupe–English unseen setting compared to existing metrics (GEMBA, xCOMET, COMETKiwi), as measured by Fleiss' κ and Kendall's τ.
- We demonstrate that our fine-tuned LLM produces translations with superior cultural and idiomatic adequacy compared to traditional NMT baselines such as NLLB-200 improvements that are only accurately captured by our proposed anchored evaluation framework.

To the best of our knowledge, this work is among the first to systematically integrate curriculum learning, bidirectional modeling, and QLoRA for low-resource translation with large language models, offering a practical end-to-end pipeline from data preparation to deployment.

2 Related Work

The application of Large Language Models (LLMs) to machine translation (MT) has progressed rapidly, moving beyond high-resource benchmarks toward increasingly challenging low-resource and unseen language scenarios. This section reviews prior work on LLM-based MT for low-resource languages, data adaptation strategies under extreme data scarcity, and the evolution of MT evaluation metrics.

2.1 LLM-Based Machine Translation for Low-Resource Languages

Recent years have witnessed a growing body of work adapting decoder-only LLMs for MT. Large-scale models such as GPT-4 and LLaMA-family variants have established strong performance on high-resource language pairs, often surpassing traditional supervised NMT systems when combined with instruction tuning and domain adaptation (Achiam et al., 2023). Parameter-efficient fine-tuning (PEFT) techniques ,most notably QLoRA have enabled practical adaptation of LLMs to low-resource tasks (Dettmers et al., 2024b).

Several recent studies have extended LLM MT to low-resource languages. Tan and Zhu (2024) introduce NusaMT-7B for Indonesian languages, but their work is limited to a single language family and lacks cross-lingual generalization beyond Indonesian. Pan et al. (2024) study translation robustness under noisy sources, showing that LLMs struggle with consistent outputs on noisy sentences despite in-context demonstrations. Pei et al. (2025) analyze in-context MT for Manchu, highlighting that dictionary and parallel examples are helpful, while grammar resources contribute minimally; however, this approach is constrained by prompt design and cannot fully overcome data scarcity. Hasan et al. (2024) evaluate 8 PEFT methods across multiple LRLs, demonstrating variability in performance and revealing that even state-of-the-art adapters do not reliably generalize across unseen languages.

Collectively, these works show that while LLMs can be adapted to low-resource MT, their effectiveness diminishes sharply for languages absent from pre-training corpora, motivating strategies that explicitly handle unseen languages.

2.2 Data Quality and Curriculum Learning for Low-Resource MT

High-quality parallel data remains a bottleneck. Noisy alignments, domain mismatch, and data imbalance often cause unstable optimization (Pan et al., 2024). Curriculum learning addresses this by progressively introducing data of increasing quality (Xu et al., 2024a). Staged fine-tuning—transitioning from large, noisy corpora to smaller, curated datasets—has improved robustness in low-resource and multilingual settings (Hui et al., 2022; Kozhirbayev and Yessenbayev, 2025). Our work applies a bidirectional, curriculum-based QLoRA adaptation strategy to the unseen Nupe–English pair, demonstrating effectiveness even with minimal cross-lingual signals.

2.3 Evolution of Machine Translation Evaluation Metrics

Traditional metrics such as BLEU and chrF++ fail to capture semantic and cultural adequacy (Rei et al., 2023). LLM-as-a-judge paradigms, including GEMBA (Kocmi and Federmann, 2023) and xCOMET (Guerreiro et al., 2024), have emerged to leverage large models for translation evaluation.

2.4 LLM-as-a-Judge for Translation Evaluation

While LLM-based evaluation achieves high correlation in high-resource pairs, it exhibits significant limitations in low-resource scenarios. Reference-free judges, including CometKiwi-XL and MetricX-23, display a “fluency bias,” giving high scores to grammatically plausible but semantically incorrect translations (Sindhujan et al., 2025; Rahman et al., 2025; Mukhambetkalieva, 2024). This is exacerbated when languages are absent from pre-training data, as in Nupe, resulting in hallucinations, inconsistent scoring, and unreliable linguistic justifications (Meeus et al., 2023).

To address this, we propose **Ref-Anchor-MQM**, a reference-anchored metric that grounds the judge in Key Semantic Units (KSUs) extracted from human references, mitigating fluency-driven hallucinations and improving evaluation reliability for unseen, extremely low-resource languages.

2.5 Positioning of This Work

Our work builds on these lines of research by combining parameter-efficient QLoRA fine-tuning with curriculum learning for extremely low-resource machine translation. Unlike prior work that focuses primarily on high-resource or multilingual settings, we target the Nupe–English language pair, for which publicly available parallel data is severely limited. To the best of our knowledge, this is the first study to systematically apply curriculum-based QLoRA fine-tuning for Nupe–English translation, providing both automatic and human evaluation and establishing a practical pipeline for LLM-based low-resource MT.

3 Methodology

The objective of this study is to adapt a decoder-only Large Language Model (LLM) for an unseen and extremely low-resource language pair, Nupe–English, under severe data scarcity. We propose a unified framework that combines curriculum-based

QLoRA adaptation with a reference-anchored evaluation metric to ensure both training stability and reliable evaluation. Figure 1 provides an overview of the proposed pipeline.

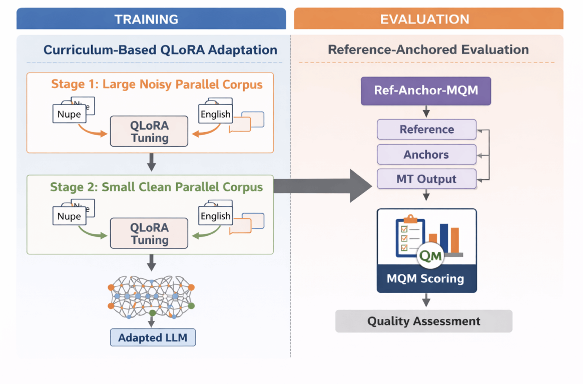

Figure 1: Overview of the proposed pipeline. The framework integrates curriculum-based QLoRA adaptation for stable fine-tuning under data scarcity with Ref-Anchor-MQM evaluation for reference-grounded semantic assessment.

3.1 Datasets

Given the absence of publicly available Nupe–English parallel corpora, we constructed a new dataset from scratch to reflect realistic extremely low-resource conditions. The corpus combines heterogeneous domains and varying levels of noise.

Data Collection. We extracted Nupe and English texts from educational materials, dictionaries, traditional and modern literature, news articles, and religious texts (Bible and Qur'an). Scanned bilingual religious documents were processed using an Optical Character Recognition (OCR) pipeline to obtain machine-readable text.

From these sources, we extracted approximately 20,000 natural Nupe sentences, 6,600 Bible and Qur'anic sentences.

To increase lexical and syntactic diversity, 5,000 selected sentences from publicly available French–English parallel corpora (LonWeb, 2017) were manually translated into Nupe by native speakers fluent in both languages, simulating realistic low-resource data expansion pipelines while 2,400 sentences were gotten from Nupe PDFs.

Corpus Statistics. The final corpus contains approximately 34,000 parallel sentence pairs totaling roughly 820k tokens.

Curriculum Learning Subsets. For curriculum-based training, we construct two subsets:

Source	#Pairs	#Tokens
Nupe Natural Texts	20,000	∼480k
Bible + Qur'an	6,600	∼160k
Other Curated Sources	∼7,400	∼180k
Total	∼34,000	∼820k

Table 1: Summary of the parallel corpus used for Nupe–English adaptation.

- **Stage 1 (Noisy Parallel Data):** The full corpus of approximately 34,000 sentence pairs is used for initial training. Sentence lengths range from 1 to 200 words, covering both short conversational content and long-form text across multiple domains.

- **Stage 2 (Clean Bidirectional Data):** A high-quality cleaned subset of approximately 12,000 manually verified and normalized sentence pairs is used for refinement. To ensure alignment stability and reduce noise, sentence lengths are restricted to 3–40 words. This subset supports bidirectional translation (Nupe↔English), enabling a single model to jointly learn both directions within a unified curriculum-based training framework.

3.2 Model Initialization and Parameter-Efficient Fine-Tuning

Large language models pre-trained on multilingual web-scale corpora typically exhibit minimal coverage of low-resource African languages such as Nupe. Moreover, full fine-tuning of such models is computationally prohibitive. To address these challenges, we adopt Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2024b) to efficiently adapt LLaMA-2 7B (Touvron et al., 2023b) for machine translation.

QLoRA enables parameter-efficient fine-tuning by quantizing the base model weights to 4-bit precision while injecting trainable low-rank adapters into selected attention projections. All base model parameters remain frozen, and only the lightweight adapter parameters are updated. This design significantly reduces memory requirements while preserving translation quality, enabling fine-tuning on a single high-end GPU.

We formulate machine translation as a conditional generation task using instruction-style prompts, which are well-suited to decoder-only architectures.

3.3 Curriculum-Based QLoRA Adaptation

Direct fine-tuning on small, high-quality parallel corpora often leads to unstable convergence and overfitting in extremely low-resource settings. To mitigate this, we employ a curriculum learning strategy (Xu et al., 2024a) that exposes the model to training data in increasing order of quality.This staged training strategy is motivated by curriculum learning (Bengio et al., 2009), which structures training data in a meaningful progression to improve optimization. In our case, the curriculum is organized by data quality: broader, heterogeneous data first encourage coarse alignment, followed by refinement on cleaner subsets to improve semantic precision.

Stage I: Noisy Parallel Pre-adaptation In the first stage, the model is fine-tuned on 34,000 segments of noisy Nupe–English parallel data collected from heterogeneous web sources and filtered using GPT-4. Although imperfectly aligned, this dataset provides broad lexical and semantic coverage, allowing the model to acquire coarse-grained cross-lingual correspondences.

Stage II: High-Quality Bidirectional Refinement In the second stage, we further fine-tune the model on 12,000 segments of carefully curated, high-quality parallel data. Training is performed bidirectionally (Nupe → English and English → Nupe) to increase the effective training signal and act as a regularizer.

To explicitly control translation direction, we prepend instruction tags to each prompt, such as:

```
Translate Nupe to English:
Translate English to Nupe:
```

This staged curriculum stabilizes optimization and mitigates the negative impact of noise and misalignment common in extreme low-resource scenarios.

3.4 Prompt Design and Training Objective

Each training example is formatted using explicit translation prompts:

```
Translate from <Source Language> to
<Target Language>: <Source Sentence>
```

All training stages optimize the causal language modeling objective, where the model predicts the next token conditioned on all previous tokens in the prompt and target sequence.

3.5 Ref-Anchor-MQM: Reference-Anchored Evaluation Metric

Standard automatic metrics such as BLEU and chrF++ are insufficient for evaluating semantic adequacy and cultural fidelity in unseen languages (Rei et al., 2023). Existing LLM-as-a-judge metrics, including GEMBA (Kocmi and Federmann, 2023) and xCOMET (Guerreiro et al., 2024), often suffer from fluency bias in low-resource settings.

To address this limitation, we propose Ref-Anchor-MQM, a reference-anchored evaluation protocol that explicitly grounds the judgment process in human-verified meaning.

Algorithm 1: Ref-Anchor-MQM Evaluation
Semantic Decomposition (SD): Given a human reference translation R, the judge model extracts a set of n Key Semantic Units (KSUs):

$$\mathcal{K} = \{k_1, k_2, \dots, k_n\} = \mathrm{LLM}_{\text{extract}}(R) \quad (1)$$

Anchored Mapping (AM): The machine-generated candidate translation C is compared against $\mathcal{K}$. Each unit k_i is classified as:

- Preserved (1.0)
- Modified (0.5)
- Lost (0.0)

Weighted MQM Scoring: Translation errors are categorized using the Multidimensional Quality Metrics (MQM) framework. The weighted error aggregation is defined as:

$$D = w_{\text{crit}} E_{\text{crit}} + w_{\text{maj}} E_{\text{maj}} + w_{\text{min}} E_{\text{min}} \quad (2)$$

where D denotes the total weighted MQM deduction.

The final score is computed as:

$$S_{\text{RefAnchor}} = 100 - D \quad (3)$$

where $w_{\text{crit}} = 25$, $w_{\text{maj}} = 5$, and $w_{\text{min}} = 1$. By anchoring evaluation to explicit semantic units derived from human references, Ref-Anchor-MQM reduces fluency bias and improves reliability in unseen language settings.

To ensure consistency in scores, the judge model first extracts Key Semantic Units (KSUs) from the human reference. The candidate translation is then evaluated against each KSU to determine whether the underlying meaning is preserved. Errors are categorized according to the Multidimensional Quality Metrics (MQM) taxonomy (Accuracy, Terminology, Omission, Addition) (Lommel et al., 2014). Severity levels—*Critical*, *Major*, and *Minor*—are assigned by the judge model based on the semantic impact of each error. The weighted error counts are then aggregated and subtracted from 100 to produce the final Ref-Anchor-MQM score. See Appendix A for the full prompt.

4 Experimental Setup

This section describes the model configuration, fine-tuning strategy, baselines, hardware environment, and evaluation protocols used to assess our curriculum-based LLM adaptation for low-resource machine translation.

4.1 Model and Fine-Tuning Configuration

All experiments use **LLaMA-2 7B** (Touvron et al., 2023a) as the base model. To enable memory-efficient adaptation, we employ **QLoRA** (Hu et al., 2022; Dettmers et al., 2023) implemented using Hugging Face Transformers, BitsAndBytes, and the PEFT library.

Model weights are quantized to 4-bit precision using the NF4 scheme with double quantization enabled, and computations are performed in bfloat16. Low-rank adapters are inserted into the query and value projections of the self-attention layers, while all base model parameters remain frozen.

The LoRA configuration is as follows: rank $r = 64$, scaling factor $\alpha = 16$, and dropout rate of 0.1. Bias parameters are frozen during training.

Fine-tuning is conducted for 3 epochs with a batch size of 32, learning rate 2×10^{-3}, constant learning rate scheduler, and warmup ratio of 0.03.

4.2 Baselines

We compare our proposed **NupeMT-QLoRA** system against two strong baselines: (i) the distilled **NLLB-200** models (600M and 1.3B variants), representing state-of-the-art multilingual NMT; and (ii) a standard **non-curriculum fine-tuned LLaMA-2 7B** model trained on the same data without staged adaptation.

4.3 Hardware Environment

All experiments are performed on Google Colab Pro+ using a single NVIDIA A100-SXM4 GPU (40GB) for both training and inference. QLoRA enables efficient fine-tuning of the 7B-parameter model on a single accelerator.

4.4 Evaluation Protocols

We evaluate translation quality using a tripartite framework:

Statistical Metrics. BLEU and chrF++ are used as traditional overlap-based metrics.

Neural Metrics. We report xCOMET, GEMBA, and COMETKiwi-DA-XXL.

Human Evaluation. Five native Nupe speakers fluent in both Nupe and English perform Direct Assessment (DA) on 106 randomly sampled test sentences, rating Adequacy, Fluency, and Cultural Naturalness on a 0–100 scale.

5 Results and Discussion

We evaluate the proposed **NupeMT-QLoRA** framework using a combination of traditional statistical metrics, state-of-the-art neural evaluation metrics, and a meta-evaluation of judge reliability against human judgments. This multi-faceted evaluation is necessary to assess translation quality and evaluation robustness in an unseen, extremely low-resource language setting.

5.1 Automatic Translation Quality Metrics

Table 2 reports performance on standard statistical metrics (BLEU, chrF++) and neural metrics (GEMBA, xCOMET) for the English $\leftrightarrow$ Nupe translation task.

NupeMT-QLoRA consistently outperforms both the NLLB-200 baseline and the non-curriculum fine-tuned LLaMA-2 model across all metrics and translation directions. The substantial gains in **chrF++** (from $\sim$23 to > 41) indicate improved character-level fluency and morphological handling, which is particularly important for the agglutinative and tonal properties of Nupe.

Although neural metrics such as GEMBA and xCOMET show modest absolute scores, they still reflect relative improvements for our model. The low magnitude of the GEMBA scores (≈ 0.30) suggests that reference-free neural metrics struggle to assess the quality of translation for unseen languages, motivating the meta-evaluation in Section 5.3.

System	Direction	BLEU ↑	chrF++ ↑	TER ↓	COMET ↑
NLLB-200 (600M)	En → Nu	6.11	22.73	88.40	41.20
NLLB-200 (600M)	Nu → En	7.45	24.10	85.10	38.50
LLaMA-2 (No Curriculum)	En → Nu	8.20	38.51	84.30	44.10
LLaMA-2 (No Curriculum)	Nu → En	9.10	39.80	81.90	41.00
NupeMT-QLoRA (Ours)	En → Nu	**9.40**	**41.10**	**82.52**	**47.34**
NupeMT-QLoRA (Ours)	Nu → En	**11.20**	**43.55**	**78.60**	**42.80**

Table 2: Traditional automatic evaluation results on the Nupe–English test set. TER is lower-is-better; all other metrics are higher-is-better.

System	Direction	xCOMET ↑	GEMBA-DA ↑	COMETKiwi-DA-XXL ↑
NLLB-200 (600M)	En → Nu	0.35	0.22	0.24
NLLB-200 (600M)	Nu → En	0.38	0.25	0.27
LLaMA-2 (No Curriculum)	En → Nu	0.49	0.28	0.31
LLaMA-2 (No Curriculum)	Nu → En	0.47	0.27	0.30
NupeMT-QLoRA (Ours)	En → Nu	**0.52**	**0.30**	**0.33**
NupeMT-QLoRA (Ours)	Nu → En	**0.55**	**0.31**	**0.35**

Table 3: LLM-as-a-judge evaluation results on the Nupe–English test set. All metrics are higher-is-better.

5.2 Curriculum Ablation Study

To isolate the impact of curriculum-based adaptation, we compare three training variants: (1) Stage-1 Only (noisy data), (2) Stage-2 Only (clean data), and (3) Curriculum (Stage-1 → Stage-2).

Model Variant	BLEU ↑	chrF++ ↑
Stage-1 Only	8.10	35.42
Stage-2 Only	9.02	39.17
Curriculum (1→2)	**11.20**	**43.55**

Table 4: Effect of curriculum-based QLoRA adaptation compared to single-stage fine-tuning.

The curriculum strategy consistently outperforms both single-stage variants, demonstrating that initial exposure to broad-coverage noisy data stabilizes optimization before refinement on high-quality parallel data. These results confirm that curriculum learning plays a critical role in effective LLM adaptation under extreme data scarcity.

5.3 Meta-Evaluation of Judge Reliability

To assess the reliability of automated evaluation metrics, we compare outputs with native Nupe speaker judgments using **inter-rater reliability** (Fleiss' κ) and **rank correlation** with human scores (Kendall's τ).

Reference-free metrics (COMETKiwi, GEMBA) achieve Fleiss' κ below 0.30, indicating *slight to fair agreement*. In contrast, reference-based evaluation improves reliability. **Ref-Anchor-MQM** attains substantial agreement ($\kappa = 0.78$) and strong correlation with human judgments ($\tau = 0.71$), demonstrating that anchoring evaluation to human-verified references is critical for reliable assessment in unseen, low-resource languages.

5.4 Human Evaluation

To establish a reliable gold standard, we conducted human evaluation on 106 randomly sampled test sentences using five native Nupe speakers fluent in both Nupe and English. Annotators completed a calibration session prior to scoring.

Each translation was evaluated independently using Direct Assessment (0–100 scale) along three dimensions: **Adequacy**, **Fluency**, and **Cultural Naturalness**. System outputs were anonymized and presented in randomized order. Annotators were blind to system identity.

System	Adeq.	Flu.	Cult.
NLLB-200	61.4	68.2	58.7
LLaMA-2 (No Curr.)	74.8	80.3	76.5
NupeMT-QLoRA	**86.2**	**88.5**	**87.9**

Table 6: Human evaluation results (DA, 0–100) on 106 test sentences.

Metric	Reference Required	Fleiss' $\kappa \uparrow$	Kendall's $\tau \uparrow$
COMETKiwi-DA-XXL	No	0.23	0.18
GEMBA	No	0.27	0.21
xCOMET	Yes	0.52	0.44
Ref-Anchor-MQM (Ours)	Yes	**0.78**	**0.71**

Table 5: Reliability and correlation of automated metrics with human judgment.

Inter-rater agreement indicates substantial consistency among annotators ($\kappa = 0.78$ for Cultural Naturalness).This demonstrates BLEU's limitation in capturing cultural fidelity.

Key Findings:

- **LLM vs. NMT Preference:** In 72% of cases involving idioms, proverbs, and culturally grounded expressions, evaluators preferred **NupeMT-QLoRA** translations over NLLB-200.

- **Fluency Trap:** In an example in the medical-domain, NLLB-200 produced a fluent but contextually incorrect translation (BLEU = 8.1). NupeMT-QLoRA produced the culturally accepted term (BLEU = 6.5). Despite a lower BLEU, human evaluators strongly preferred the latter. Ref-Anchor-MQM correctly flagged NLLB output as a *Major Terminology Error*, whereas GEMBA overvalued surface-level fluency.

5.5 Discussion

These results confirm that **curriculum-based QLoRA adaptation** stabilizes LLM fine-tuning under extreme data scarcity, and that **Ref-Anchor-MQM** is essential for reliable evaluation in unseen low-resource languages. The findings highlight the importance of jointly considering model adaptation strategies and evaluation protocols when extending LLM-based MT to the extremes of multilingual coverage.

6 Conclusion and Future Work

This paper investigated the feasibility of adapting decoder-only Large Language Models (LLMs) to an *unseen, extremely low-resource* language pair, Nupe–English. We introduced **NupeMT-QLoRA**, a curriculum-based QLoRA adaptation framework, together with **Ref-Anchor-MQM**, a reference-anchored evaluation protocol designed to address the limitations of existing machine translation metrics in unseen language settings.

Our results demonstrate that curriculum-based exposure—progressing from large, noisy parallel data to smaller, high-quality bidirectional corpora—significantly stabilizes fine-tuning and improves translation robustness under severe data scarcity. Notably, with only **12k clean parallel sentence pairs**, we are able to build a functional and competitive MT system for a language that is effectively absent from large-scale pre-training corpora. This finding highlights the practical viability of LLM-based MT for truly unseen languages.

More broadly, our results reveal a complementary relationship between LLM-based MT and traditional NMT systems. While conventional NMT models tend to favor *literal accuracy* and word-level faithfulness, they often struggle with pragmatic meaning, idiomatic expressions, and culturally grounded language use. In contrast, LLM-based MT compensates for limited parallel data by leveraging strong prior knowledge of discourse structure, fluency, and pragmatic reasoning. As a result, LLMs produce translations that are more natural, contextually appropriate, and semantically coherent, even when trained on very limited high-quality data. This trade-off suggests that LLM-based MT is particularly well-suited for low-resource and unseen language scenarios, where data scarcity severely constrains conventional NMT approaches.

Beyond modeling improvements, our study reveals a critical evaluation gap. Reference-free LLM-as-a-judge metrics exhibit low inter-rater reliability and weak correlation with human judgments for Nupe, largely due to the absence of prior linguistic exposure during pre-training. By explicitly anchoring evaluation to human references through Key Semantic Units, **Ref-Anchor-MQM** achieves substantially higher reliability and alignment with human judgments, effectively mitigating fluency bias in unseen language evaluation. In summary, we addressed two core objectives: stabilizing adaptation of LLMs for an unseen low-resource language and improving evaluation reliability through reference-anchored semantic decomposition.

Taken together, these findings suggest that progress in extremely low-resource machine translation requires *joint advances in both adaptation strategies and evaluation protocols*. Relying solely on model scale, literal accuracy, or unanchored neural metrics risks overestimating translation quality in languages at the fringes of multilingual coverage.

Future Work. Future research will extend the proposed framework to additional unseen African languages and other typologically diverse low-resource settings. We also plan to explore automatic or semi-automatic construction of Key Semantic Units to reduce reliance on human references, as well as multilingual judge models explicitly trained to reason over unseen languages. Finally, integrating speech and multimodal signals may further enhance translation quality and evaluation robustness in low-resource communities.

7 Limitations

Despite promising results, this work has several limitations.

First, while curriculum-based QLoRA improves stability, our approach still relies on the availability of *some* parallel data, including noisy web-mined or pivot-generated corpora. For languages with no written resources or digital presence, even this level of data may be unavailable.

Second, **Ref-Anchor-MQM** depends on human reference translations to anchor evaluation. Although this design improves reliability, it introduces additional annotation cost and limits scalability compared to fully reference-free metrics. Automating semantic unit extraction without degrading reliability remains an open challenge.

Third, our experiments focus on LLaMA-2 7B and the Nupe–English language pair. Although the methodology is general, the results may not transfer directly to models with different architectures or to languages with substantially different typological properties. In particular, we only consider one unseen language in this study (Nupe), so further validation is required to confirm generalizability across other extremely low-resource languages.

Finally, human evaluation was conducted on a limited subset of the test data due to the availability of native speakers. Although sufficient for statistical validation, larger-scale human studies would further strengthen the conclusions.

Acknowledgments

We are grateful for the technical assistance of S.A. Bashir.

Funding Statement

This research was not supported by any grants.

Competing Interests

None.

Data Availability Statement

Replication data and code are publicly available at https://data.mendeley.com/datasets/k7dtv7k2hy/1.We will release the NupeMT-QLoRA adapters and the Ref-Anchor-MQM evaluation code on GitHub upon publication.

Ethical Standards

The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions

Conceptualization: Umar .U.B; S.A Bashir ;M.D Abdulmalik Methodology:S.A Bashir Data curation: Umar.U.B Data visualization: Umar.U.B Writing — original draft: Umar.U.B All authors approved the final submitted draft.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Aleman, Diogo Almeida, and ... 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Ifeoluwa Adelani, Jesujoba O. Alabi, Angela Fan, Julia Kreutzer, Samuel Olowobo, Ituen Okayo, Bernard Opoku, Chiamaka Chukwuneke, and 1 others. 2022. A few-shot learning approach for low-resource african languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 904–917.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, José Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and Pierre Colombo. 2024. Tower: An open multilingual large language model for translation-related tasks. *Computing Research Repository*, arXiv:2402.17733.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, pages 41–48. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024a. QLoRA: Efficient fine-tuning of quantized LLMs. *Computing Research Repository*, arXiv:2305.14314.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024b. QLoRA: Efficient fine-tuning of quantized LLMs. *Computing Research Repository*, arXiv:2305.14314.

Nuno M. Guerreiro, Matt Post, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:1001–1018.

Md Mehedi Hasan, Shuo Ma, Mohammad S. Khan, and 1 others. 2024. Unlocking parameter-efficient fine-tuning for low-resource language translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4152–4166, Mexico City, Mexico. Association for Computational Linguistics.

Edward J. Hu, Yallen Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Ziyang Hui, Chong Feng, and Tianfu Zhang. 2022. Review-based curriculum learning for neural machine translation. In *Machine Translation: 18th China Conference, CCMT 2022, Lhasa, China, August 4–6, 2022, Revised Selected Papers*, pages 24–36. Springer.

Nisar Hussain, Amna Qasim, Gull Mehak, Muhammad Zain, Momina Hafeez, and Grigori Sidorov. 2025. Fine-tuning large language models with qlora for offensive language detection in roman urdu-english code-mixed text. *arXiv preprint*. Available at arXiv:2510.03683.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2024. Challenges in adapting multilingual llms to low-resource languages using lora peft tuning. *arXiv preprint*. Available at arXiv:2411.18571.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Zhanibek Kozhirbayev and Zhandos Yessenbayev. 2025. Fine-tuning methods and dataset structures for multilingual neural machine translation: A kazakh–english–russian case study in the it domain. *Electronics*, 14(15):3126.

S. Lankford, H. Afli, and A. Way. 2025. adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. *Information*, 14(12):638.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. In *Proceedings of the LT-Innovate Summit*.

LonWeb. 2017. The night watch (french-english parallel text). http://www.lonweb.org/daisy/ds-frenchnightwatch.htm. Accessed January 2026.

Quentin Meeus, Marie-Francine Moens, and Hugo Van hamme. 2023. Bidirectional representations for low-resource spoken language understanding. *Applied Sciences*, 13(20):11291.

Aislu Mukhambetkalieva. 2024. New reference-based metrics for MT evaluation: MetricX, GEMBA-MQM, and CometKiwi-XL. Intento Blog.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 13 others. 2022. No language left behind: Scaling human-centered machine translation to 200 languages. *Computing Research Repository*, arXiv:2207.04672.

Femi Oladipupo. 2025. A beginner's guide to fine-tuning AI models on african datasets. Medium.

Leiyu Pan, Yongqi Leng, and Deyi Xiong. 2024. Can large language models learn translation robustness from noisy-source in-context demonstrations? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2798–2808, Torino, Italia. ELRA and ICCL.

Li Pan, Yufeng Leng, and Deyi Xiong. 2025. The noise sensitivity of LLMs in zero-shot translation tasks. *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2024. Llms for extremely low-resource finno-ugric languages. *Preprint*, arXiv:2410.18902.

Md. Atiqur Rahman, Sabrina Islam, and Mushfiqul Haque Omi. 2025. Llm-based evaluation of low-resource machine translation: A reference-less dialect guided approach with a refined sylheti-english benchmark. *Preprint*, arXiv:2505.12273.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luísa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orăsan, and Shenbin Qian. 2025. When llms struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhinav Singh, Nishant Singh, and Shreyas Vatsal. 2024. Robustness of LLMs to perturbations in text. *Computing Research Repository*, arXiv:2407.08989.

William Tan and Kevin Zhu. 2024. Nusamt-7b: Machine translation for low-resource indonesian languages with large language models. *Preprint*, arXiv:2410.07830.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023a. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 19 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Computing Research Repository*, arXiv:2307.09288.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2024a. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2024)*, pages 3977–3989.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen shifts, Benjamin Van Durme, and Yong-Yeol Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *Computing Research Repository*, arXiv:2401.08417.

A Appendix

Appendix: Ref-Anchor-MQM Prompt

Step 1: Key Semantic Unit Extraction. Given a human reference translation R, extract all Key Semantic Units (KSUs) representing essential meaning components.

Output format:

1. KSU_1
2. KSU_2
3. $\ldots$
4. KSU_n

Step 2: Anchored Evaluation. Given: (i) Source sentence S, (ii) Reference translation R, (iii) Candidate translation C, and (iv) Extracted KSUs.

For each KSU, determine whether it is:

- Preserved (1.0)
- Modified (0.5)
- Lost (0.0)

Then assign MQM error categories:

- Critical (Terminology / Meaning)
- Major
- Minor

Compute the final score as:

$$S = 100 - \left(25E_{\text{crit}} + 5E_{\text{maj}} + 1E_{\text{min}}\right)$$

Provide a short justification for the assigned score.

Source (EN)	Reference (Nupe)	Prediction	Ref-A-MQM Score	GPT-4o Comment
That He may reward those who believe and work deeds of righteousness for such is Forgiveness and a Sustenance Most Generous	ebo soko u la egagba wangi u ya za nanzhi a imani jin a cie lo tun wangizhi nan za wuncinzhi gafara yan soko ajin tabbata yaa ganni be arziki nan wancin nan	ebo soko u la egagba ya zanazhi ga imani jin na acie lo tun wangizhi na hakika egagba yan gafara uyio to arziki na wancin na	0.60	The prediction captures some elements of the source text but has several issues. Errors in fluency and adequacy, such as 'ya zanazhi' which seems to be a mistranslation, and 'na acie' which does not reflect the meaning of 'work deeds of righteousness'. Overall, clarity and coherence are lacking.
Then after them sent We Moses and Aaron to Pharaoh and his chiefs with Our Signs But they were arrogant they were a people in sin	gancin zunma yanabo yici tun musa be harunae tun lo eba firauna to wancinyanuzhi nazhie yau shawara na be aya yizhi acia gbotika acia fe jamaa nazhi yi lefi jincizhi na	gancin yici tun musa to harunzhi tun fi firauna be jamaa yan wancinuzhi be aya yizhi amaade zanazhia fe dan mini lefi jincizhio na	0.62	The prediction captures some elements but has several issues. Key terms such as 'Moses' and 'Aaron' are inconsistently represented. 'With Our Signs' is omitted, and structure affects fluency, leading to incomplete meaning.
he doesnt know me at all	u kpe mi ye a nayi	u kpe mi ye a	0.50	The prediction 'u kpe mi ye a' captures part of the meaning but is incomplete. The word 'nayi' is missing, so the full meaning of 'at all' is lost. Overall, only partial understanding of the source is conveyed.
As Nnadzwa stepped into the hut	nnadzwa de sa bici dan kata ò	nnadzwa tun kata o	0.40	The prediction fails to convey 'stepped into' accurately, and 'kata' is not a direct translation of 'hut'. Fluency is also lacking, making the sentence sound unnatural.
Dont consider the goodness of a man as long as he has money	ga iwo ege nya bagi ma bagi de u de wo	ga kpaye yan eza wangi kaminan u de ewo na	0.40	The prediction does not accurately convey the conditional meaning of the source. Words like 'kpaye' and 'wangi' are partially correct, but overall structure and meaning are distorted.

Table 7: Top 5 qualitative Nupe–English translation examples with LLM-assigned scores and comments.