http://dx.doi.org/10.21622/ACE.2026.06.1.1933

LOW-RESOURCE FINE-TUNING OF LLAMA-3.1 USING QLORA FOR NIGERIAN LINGUISTIC CONTEXTS

Solomon Joseph Udoabba[1], Bliss Utibe-Abasi Stephen[1,2*], Oluseun Damilola Oyeleke[2], Philip Asuquo[1,2] and Sadiq Thomas[3]

[1] University of Uyo, Department of Computer Engineering, Nigeria
[2] TETFund Centre of Excellence in Computational Intelligence Research
[3] Global Banking School, United Kingdom
[4] Nile University, Department of Computer Engineering, Abuja, Nigeria

Emails: { blissstephen@uniuyo.edu.ng, uniqueudoabba@gmail.com, contactseun@gmail.com, philipasuquo@uniuyo.edu.ng, sadiqthomas@nileuniversity.edu.ng }

Received 31 March 2026 - Accepted 20 May 2026 - Published 23 June 2026

ABSTRACT

Large language models (LLMs) struggle to accommodate underrepresented linguistic contexts, such as Nigerian English and Nigerian Pidgin, due to Western-centric training corpora. While fine-tuning offers a solution, aggressive adaptation strategies may risk degrading general-domain performance when trained on small datasets.

This study investigates the feasibility of a conservative Quantized Low-Rank Adaptation (QLoRA) strategy to align LLaMA-3.1-8B with Nigerian contexts under extreme compute constraints. As a pilot study, we employed a dampened update scaling factor (α = 16, r = 64) on a stratified subset of 12,000 Nigerian web text samples. This configuration was deliberately chosen to prioritize model stability and preserve pre-trained general knowledge while inducing cultural alignment. Preliminary results indicate that this conservative approach successfully avoids model collapse, yielding a stable reduction in perplexity (8.20 to 7.38). Initial qualitative probing reveals a divergence between pragmatic alignment and semantic precision: the model successfully internalized the affective sentiment of Nigerian slang (e.g., associating "Sapa" with frustration), even where it lacked sufficient data to generate precise dictionary definitions. Furthermore, we discuss the economic implications of this approach, arguing that parameter-efficient fine-tuning offers a lower "inference tax" forresource-constrained deployment compared to few-shot prompting. This work provides a technical reference for "safety-first" adaptation in low-resource academic environments.

We explore whether large language models can be meaningfully adapted to Nigerian linguistic contexts using limited computational resources. Many African languages and regional dialects are underrepresented in the data used to train modern foundation models. As a result, these systems may misinterpret culturally grounded expressions or fail to capture local usage patterns. In this work, we fine-tuned LLaMA-3.1-8B using QLoRA on Nigerian web text while operating under hardware constraints typical of universities in developing regions. Instead of pursuing large-scale optimization, we focused on feasibility, training behavior, and qualitative response shifts.

Our results show that parameter-efficient fine-tuning can introduce measurable contextual adaptation without requiring full model retraining. At the same time, improvements are modest and do not eliminate semantic inaccuracies. By presenting both strengths and limitations, we aim to provide a realistic foundation for future work on culturally aligned AI systems in low-resource settings.

Keywords: *Parameter-efficient fine-tuning, natural language processing, low resource languages, Nigerian Pidgin*

http://dx.doi.org/10.21622/ACE.2026.06.1.1933

INTRODUCTION

Artificial Intelligence has emerged as a dominant computational paradigm in Natural Language Processing (NLP), with models such as OpenAI's GPT series and Meta's LLaMA series setting new performance benchmarks [1]. Specifically, Meta LLaMA 3.1 8B offers a powerful foundation for downstream applications due to its balance of performance and efficiency [2]. However, the efficacy of these models is inextricably linked to their training data, which remains predominantly Western-centric.

This data disparity creates a "representation gap" for African languages. In Nigeria, a nation with over 520 languages and a unique lingua franca (Nigerian Pidgin), standard models often fail to capture syntactic and cultural nuances [3]. For instance, cultural slang terms like "Sapa" (financial struggle) or "Jollof wars" are frequently misinterpreted by base models as generic English terms. This limitation hinders the development of effective NLP applications for local needs, from automated customer service to educational tools [4].

To address this, we investigate the fine-tuning of LLaMA 3.1 8B for Nigerian contexts using Parameter-Efficient Fine-Tuning (PEFT). Recent works by Hu et al. [5] and Dettmers et al. [6] have demonstrated that QLoRA (Quantized Low-Rank Adaptation) allows for the fine-tuning of massive models on single GPUs without significant performance degradation. Similar methodologies have been successfully applied to Vietnamese healthcare [7] and South African languages (Zulu/Xhosa) [8], validating PEFT as a viable pathway for low-resource adaptation.

The central contribution of this work is a reproducible, low-cost methodology for adapting LLMs to Nigerian linguistic contexts. We hypothesize that by employing QLoRA on a curated subset of the NaijaWeb corpus, we can induce measurable shifts in the model's cultural alignment and Pidgin fluency within a resource-constrained computational environment.

It is important to note that this work is positioned as an exploratory feasibility study rather than a comprehensive benchmark evaluation. Due to computational and dataset constraints, the objective is to examine whether parameter-efficient fine-tuning can produce measurable shifts in cultural alignment within a limited-resource academic setting. Accordingly, the evaluation focuses on preliminary quantitative indicators and structured qualitative assessment rather than large-scale comparative benchmarking.

MATERIALS AND METHODS

DATASET PREPARATION

The training data was sourced from the NaijaWeb corpus [9], a diverse collection of web-scraped Nigerian text. To accommodate the memory constraints of our hardware (Tesla T4 16GB VRAM), we selected a stratified subset of 12,000 samples. The data distribution included Nigerian English (45%), Nigerian Pidgin (30%), and code-switched text (25%).

Crucially, we reformatted the raw text into an instruction-following format (Alpaca style) to align with the supervised fine-tuning (SFT) paradigm. Each entry was structured as an input instruction (e.g., "Explain this concept") and a target output,

http://dx.doi.org/10.21622/ACE.2026.06.1.1933

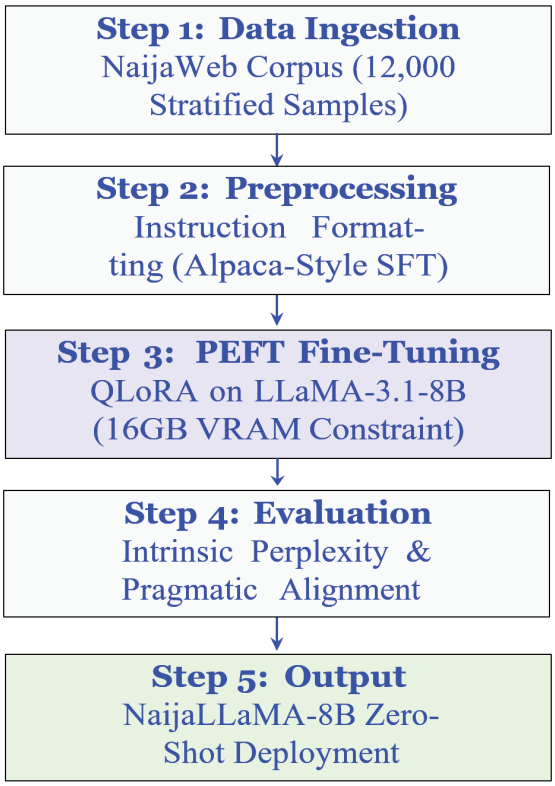

Figure 1: Proposed Methodology Framework.

The workflow illustrates the end-to-end pipeline utilized in this study: (1) Ingestion of the NaijaWeb Dataset, (2) Stratification and Alpaca-style instruction preprocessing, (3) Parameter-efficient fine-tuning via QLoRA under a 16GB VRAM constraint, (4) Model evaluation focusing on perplexity and qualitative pragmatic alignment, and (5) The proposed zero-shot deployment model. facilitating the model's ability to follow cultural prompts. A separate validation set of 500 examples was reserved for evaluation.

MODEL CONFIGURATION AND HARDWARE

We utilized unsloth/llama-3.1-8b-bnb-4bit, a pre-quantized version of Meta's LLaMA 3.1. The use of 4-bit NormalFloat (NF4) quantization was essential to fit the 8-billion parameter model into the 16GB VRAM of a single GPU provided by the Kaggle platform.

We employed the Unsloth library for optimization, which provides up to 2x faster training speeds and reduced memory fragmentation compared to standard Hugging Face implementations.

FINE-TUNING HYPERPARAMETERS

We applied QLoRA adapters to the linear layers of the attention mechanism (q proj, k proj, v proj, o proj). The specific hyperparameters were:

- LoRA Rank (r): 64
- LoRA Alpha (α): 16 (Conservative scaling to prevent overfitting)
- Dropout: 0

http://dx.doi.org/10.21622/ACE.2026.06.1.1933

- Optimizer: AdamW 8-bit
- Learning Rate: 2×10^{-4} with linear decay
- Batch Size: 2 (with 8 gradient accumulation steps for an effective batch size of 16)

The model was trained for 1 epoch to prevent overfitting on the small dataset, with checkpoints saved every 50 steps.

RESULTS

As shown in Fig 2, the training loss decreased steadily from 2.04 to 1.98, indicating stable optimization under conservative scaling. The model's adaptation was evaluated using both quantitative metrics and qualitative human evaluation to measure cultural alignment.

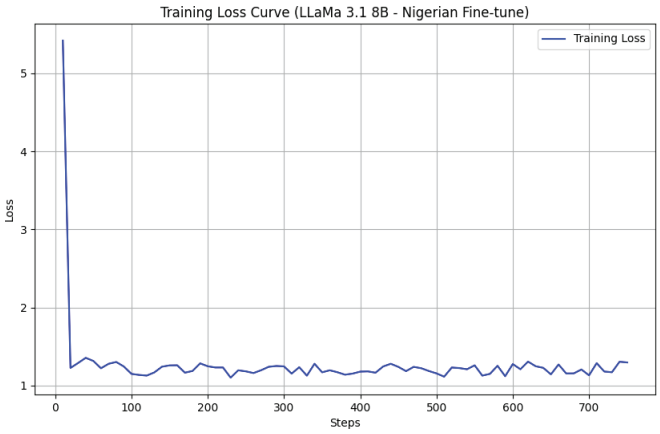

Figure 2: Training loss trajectory during QLoRA fine-tuning.

The model was trained for one epoch on 12,000 Nigerian web samples using conservative scaling ($r = 64$, $\alpha = 16$). The smooth decline indicates stable optimization dynamics during fine-tuning. While validation loss was not extensively tracked due to compute constraints, no instability was observed during training.

QUANTITATIVE EVALUATION

We observed improvements across intrinsic metrics, as detailed in Table 1. The reduction in perplexity indicates that the model became less "surprised" by Nigerian sentence structures and Pidgin syntax.

Table 1: Intrinsic Evaluation Metrics.

Metric	Baseline (LLaMA 3.1)	Fine-Tuned (NaijaLLaMA)	Improvement
Training Loss	2.04	1.98	-2.9%
Perplexity	8.20	7.38	-10.0%

Lower values indicate better performance.

STRUCTURED QUALITATIVE EVALUATION

To reduce anecdotal bias, a lightweight structured evaluation was conducted on 30 prompts spanning three categories: (1) Nigerian slang interpretation, (2) Nigerian Pidgin translation, and (3) culturally grounded knowledge queries. Prompts were sampled from culturally relevant expressions present in the training corpus.

http://dx.doi.org/10.21622/ACE.2026.06.1.1933

Each response was manually assessed by two Nigerian postgraduate students familiar with Nigerian English and Pidgin. Outputs were categorized as Correct, Partially Correct, or Incorrect based on semantic accuracy and contextual appropriateness.

Across the 30 prompts, the fine-tuned model produced 18 Correct, 7 Partially Correct, and 5 Incorrect responses. In contrast, the base model produced 11 Correct, 9 Partially Correct, and 10 Incorrect responses. While not statistically analyzed, this distribution suggests a consistent directional shift toward improved contextual alignment in culturally specific tasks. Formal statistical validation with larger evaluation sets remains necessary to confirm robustness.

These findings should be interpreted as preliminary due to limited sample size and absence of formal inter-rater reliability measurement.

Table 2: Comparative Analysis of Model Generations.

Task	Prompt	Base Model Output	Fine-Tuned Output
Slang	Explain "Sapa"	"Sapa is a slang term... for something unimportant." (Hallucination)	"Sapa is a term... used to express disapproval or disappointment... frustration with a situation."
Translation	Translate "How are you?"	"How are you doing?" (Standard English)	"How you dey?" (Syntactically Correct Pidgin)
Knowledge	Who is Fela Kuti?	(Generic Biography)	"Fela Kuti was a Nigerian multi-instrumentalist... nickname was 'the Black President'."

The examples illustrate qualitative shifts in contextual grounding following adaptation.

DISCUSSION

The central objective of this study was to determine if a state-of-the-art foundation model could be meaningfully adapted to Nigerian linguistic nuances using

consumer-grade hardware and a "conservative" adaptation strategy. The results offer three critical insights into the dynamics of low-resource fine-tuning.

STABILITY VS. PLASTICITY IN HYPERPARAMETERS

A key methodological decision in this work was the selection of a LoRA effective scaling factor of 0.25 (computed as $\alpha/r = 16/64$). Many implementations use higher effective scaling factors to maximize adaptation strength. However, given the noisy nature of web-scraped text and the limited dataset size (12,000 samples), we hypothesized that aggressive updates would risk overfitting or "catastrophic forgetting" of the model's core logic.

Our stable training loss trajectory confirms that the dampened signal acted as an effective regularizer. The model did not exhibit the erratic behavior often associated with high-rank updates on small data. This suggests that for low-resource languages where high-quality, instruction-verified data is scarce, a low-alpha strategy may provide a safer pathway to adaptation, effectively "nudging" the model's tone without overwriting its foundational reasoning capabilities.

PRAGMATIC ALIGNMENT VS. SEMANTIC DEFINITION

The model's performance on cultural concepts, particularly the term "Sapa," illuminates a distinction between pragmatic alignment and semantic precision.

http://dx.doi.org/10.21622/ACE.2026.06.1.1933

When asked to define "Sapa" (a slang term for financial struggle), the model described it as a term for "disapproval or frustration." While semantically inaccurate (it missed the specific definition of "poverty"), the model successfully captured the sentiment polarity. It understood that "Sapa" appears in distributions of text

associated with negative sentiment and complaint, even if it could not retrieve the precise dictionary definition. This phenomenon suggests that QLoRA on small, unannotated corpora is highly effective at Association Learning (capturing the "vibe" or emotional context) but less effective at acquiring new factual definitions without explicit instruction tuning. Researchers using scraped data should therefore expect improvements in tone before improvements in factual accuracy.

INFERENCE EFFICIENCY AND DEPLOYMENT VIABILITY

While recent literature suggests that Few-Shot Prompting (In-Context Learning) can achieve comparable accuracy to fine-tuning, we argue that fine-tuning may offer practical deployment advantages in settings with limited bandwidth or compute due to Inference Economics.

Few-shot prompting requires feeding long context examples with every single query, significantly increasing the token count, latency, and financial cost per interaction. In contrast, our PEFT approach "bakes" the cultural context into the adapter weights.

This allows for zero-shot deployment, where the model understands Nigerian context without requiring expensive system prompts. For deployment in rural or academic settings with limited internet bandwidth and hardware, this reduction in inference-time compute is a critical advantage that outweighs marginal gains in accuracy.

LIMITATIONS

We acknowledge that the conservative scaling factor ($\alpha = 16$) likely limited the total possible reduction in perplexity. Future work with larger, curated datasets should explore dynamic α scheduling to balance stability with more aggressive learning.

Additionally, while the model captures the syntax of Nigerian Pidgin, semantic hallucinations persist, indicating the need for the incorporation of curated lexical definition data in future dataset preparation.

CONCLUSION

In this work, we refer to the fine-tuned variant as NaijaLLaMA-8B for clarity, while acknowledging that it consists of LoRA adapters applied to the LLaMA-3.1-8B base model. By leveraging Unsloth and QLoRA, we adapted a state-of-the-art 8B parameter model on a single GPU, indicating preliminary improvements in culturally contextualized responses and Pidgin fluency.

Future work will focus on these areas: (1) scaling the dataset to include explicit definitions of slang to improve semantic precision; and (2) extending the training to include low-resource indigenous languages like Ibibio and Tiv.

http://dx.doi.org/10.21622/ACE.2026.06.1.1933

SUPPORTING INFORMATION

S1 File. Training Logs and Model Adapters. The full training logs, loss curves, and the trained LoRA adapters are available at https://huggingface.co/luminaudoabba/llama3-naijaweb-merged.

ACKNOWLEDGMENTS

We acknowledge the Department of Computer Engineering, University of Uyo, for providing the academic environment to conduct this research. We also thank the creators of the NaijaWeb corpus and the Unsloth AI team for their open-source contributions.

REFERENCES

[1] L. Qin *et al.*, "Large Language Models Meet NLP: A Survey," *arXiv.org*, 2024. https://arxiv.org/abs/2405.12819

[2] A. Dubey *et al.*, "The Llama 3 Herd of Models," *arXiv.org*, 2024. https://arxiv.org/abs/2407.21783

[3] I. Inuwa-Dutse, "NaijaNLP: A Survey of Nigerian Low-Resource Languages," *arXiv.org*, 2025. https://arxiv.org/abs/2502.19784

[4] J. Ojo *et al.*, "AfroBench: How Good are Large Language Models on African Languages?," *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 19048–19095, Jan. 2025, doi: https://doi.org/10.18653/v1/2025.findings-acl.976.

[5] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv:2106.09685 [cs]*, Oct. 2021, Available: https://arxiv.org/abs/2106.09685

[6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv.org*, May 23, 2023. https://arxiv.org/abs/2305.14314

[7] N. Bui *et al.*, "Fine-tuning Large Language Models for Improved Health Communication in Low-Resource Languages," *Computer Methods and Programs in Biomedicine*, vol. 263, pp. 108655–108655, Feb. 2025, doi: https://doi.org/10.1016/j.cmpb.2025.108655.

[8] P. W. Khoboko, V. Marivate, and J. Sefara, "Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models," *Machine Learning with Applications*, vol. 20, p. 100649, Jun. 2025, doi: https://doi.org/10.1016/j.mlwa.2025.100649.

[9] S. A. Ayanniyi, "Naijaweb: A Web Scraped Nigerian Context Dataset," *Huggingface.co*, 2021. https://huggingface.co/datasets/saheedniyi/naijaweb