ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Transformer Models for Hausa Personal Name Spelling Correction: A Comparative Evaluation of Byt5, Flan-T5, and AfroLlama

Bernard Ephraim[1,2*], Ajah Ifenyinwa[2]

[1]Department of Computing Sciences, Admiralty University of Nigeria, Ibusa, Delta State, Nigeria,

[2]Department of Computer Science, Ebonyi State University Abakaliki, Ebonyi State, Nigeria,

***Corresponding Author's Address**

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150800074

Received: 28 August 2026; Accepted: 02 September 2026; Published: 15 September 2026

ABSTRACT

Misspelling of personal names presents a significant challenge to data integrity, record linkage, identity verification, and information retrieval, particularly in low-resource linguistic environments. Hausa personal names exhibit substantial spelling variation arising from phonetic, orthographic, transliteration, and typographical differences, yet limited research has investigated Transformer-based approaches for their automatic correction. This study investigates the effectiveness of Transformer-based language models for automatically correcting misspelt Hausa personal names. A corpus comprising 148,710 spelling instances derived from 2,061 Hausa personal names was used, including 19,472 real human spelling attempts and synthetically generated variants. Three models, ByT5-small, Flan-T5-small, and AfroLlama, were evaluated under a three-phase cumulative curriculum-learning framework in which progressively larger edit-distance ranges were introduced during training. To reduce the possibility of canonical-name-level data leakage, the 2,061 canonical names were partitioned into mutually exclusive training, validation, and test sets containing 1,690, 185, and 186 canonical names, respectively. Model performance was evaluated using exact-match accuracy, character-level precision, recall and F1, SacreBLEU, and character error rate (CER), with McNemar's test used to assess differences in paired prediction outcomes. Fine-tuned ByT5-small achieved the strongest overall performance, correctly restoring 3,328 of 13,558 test instances (24.55%), with character precision of 92.45%, recall of 80.68%, F1 of 86.16%, SacreBLEU of 69.7286, and CER of 0.2048. Fine-tuned Flan-T5-small achieved 11.85% exact-match accuracy and 83.14% character F1. In contrast, AfroLlama declined from 6.84% to 4.46% exact-match accuracy following fine-tuning, accompanied by a reduction in character F1 from 74.93% to 69.70%. McNemar's tests indicated statistically significant differences between each pair of evaluated models ($p < 0.001$). The findings demonstrate that, under the experimental configuration examined, ByT5-small provided the strongest correction performance among the evaluated models. However, the results should not be interpreted as establishing the superiority of an entire model class. The study also highlights the need for human-versus-synthetic evaluation, curriculum ablation, architecture-specific optimization, and broader baseline comparisons in future research.

Keywords: ByT5; Flan-T5; Low-resource; Spelling correction; Transformer models.

INTRODUCTION

Personal names are fundamental components of human identity and are widely used in administrative, educational, healthcare, banking, electoral, and civil registration systems. Accurate representation of names is therefore essential for data integrity, record linkage, identity verification, and information retrieval (Christen, 2006). This is particularly important in multilingual and low-resource environments, where inconsistent orthographic conventions often produce multiple forms of the same name. In Nigeria, Hausa is widely spoken across Northern Nigeria and neighbouring West African countries, and Hausa personal names reflect indigenous, Arabic, and Islamic influences. Consequently, names may exhibit variants such as Abdulahi/Abdullahi,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Muhamad/Muhammad, and Aisha/A'isha, arising from phonetic approximations, typographical errors, transliteration, dialectal influences, and varying literacy levels (Bernard & Ajah, 2025). Such inconsistencies can result in duplicate records, failed identity verification, inaccurate analytics, and inefficient information retrieval.

Traditional spelling correction methods, including edit-distance and phonetic matching, provide useful mechanisms for identifying spelling variants but generally rely on fixed representations and may not adequately capture complex linguistic relationships (Bernard & Ajah, 2025). Recent research has consequently shifted toward Transformer-based approaches capable of learning spelling transformations directly from noisy and corrected texts. Misspellings are pervasive in digital communication and can adversely affect downstream natural language processing (NLP) tasks such as entity recognition and information extraction (Sperduti & Moreo, 2026). Moreover, large language models trained predominantly on internet-sourced data may inherit spelling variations, while multilingual models such as mT5 exhibit differing levels of robustness to phonetic and orthographic noise in real-world data (Aliakbarzadeh et al., 2025; Phonchai et al., 2025). These limitations are particularly relevant to low-resource African languages, where annotated datasets and computational resources remain comparatively scarce (Okewunmi et al., 2025).

Among recent approaches, token-free and byte-level architectures such as ByT5 have demonstrated considerable potential for noisy-text processing. ByT5 represents text as raw byte sequences rather than relying on predefined word or subword vocabularies, thereby reducing the limitations associated with conventional tokenization (Xue et al., 2021, 2022). This representation is particularly relevant to Hausa personal names, where spelling variation frequently occurs at the character level and rare names may be poorly represented by standard subword vocabularies (Kuparinen et al., 2023; Lutgen et al., 2024). Byte-level processing uses a compact universal vocabulary of 256 symbols, enabling models to represent unusual character sequences without the over-segmentation often associated with subword tokenizers (Al-Rfooh et al., 2023; Wu et al., 2024). Studies have further shown that byte-level approaches can be advantageous for underrepresented languages and noisy or out-of-distribution inputs (Aars et al., 2024; Feher et al., 2024; Limisiewicz et al., 2024; S et al., 2024; Owodunni et al., 2025).

Beyond byte-level modelling, sequence-to-sequence Transformer architectures have increasingly been applied to text normalization and spelling correction in low-resource settings. These approaches commonly fine-tune pretrained models on parallel corpora containing noisy and corrected forms, including datasets created through synthetic noise injection (Lutgen et al., 2024; Wali & Nisioi, 2025).

Such supervised approaches can move beyond static rule-based mappings by learning more nuanced spelling transformations, with evaluation commonly based on metrics such as Character Error Rate (CER) and Bilingual Evaluation Understudy (BLEU) scores (Aliero et al., 2025; Lutgen et al., 2024; Wali & Nisioi, 2025). This makes them potentially suitable for Hausa personal-name correction, where errors may involve character substitutions, insertions, deletions, spacing, and other orthographic variations.

Despite these advances, research on automatic personal-name spelling correction remains limited compared with general text spelling correction, particularly for African and other low-resource languages. Existing work has largely concentrated on high-resource languages, while the application of multiple Transformer architectures to Hausa personal names remains insufficiently investigated.

To address this gap, this study comparatively evaluates ByT5-small, Flan-T5-small, and AfroLlama (Ismail et al., 2025; Lefrandt et al., 2025) using a dedicated Hausa personal-name spelling dataset. The models are fine-tuned on correctly spelt names and their misspelt variants and evaluated using Accuracy, character-level Precision, Recall, F1-score, SacreBLEU, CER, and McNemar's significance test.

The study makes four principal contributions: (1) development of a benchmark for Transformer-based Hausa personal-name spelling correction; (2) comparative evaluation of byte-level, instruction-tuned, and African-centric Transformer architectures; (3) analysis of the strengths and limitations of the evaluated models in

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

handling Hausa orthographic variation; and (4) empirical evidence to support further research and practical development of intelligent name verification and correction systems for low-resource African languages.

MATERIALS AND METHODS

Transformer Models

This study evaluated three different LLMs for the task of automatic correction of misspelt Hausa personal names based on their architectural diversity and demonstrated effectiveness in sequence generation tasks: ByT5 small, AfroLlama, and Flan-T5 small. They were chosen for their lightweight yet great performance, their architectural differences and potential suitability for handling text generation and transformation tasks. AfroLlama is an 8B parameter instruction-tuned causal decoder-only Transformer language model obtained by enhancing Llama 3 8B for Swahili, Xhosa, Zulu, Yoruba, Hausa, and English. By incorporating African linguistic resources during pretraining and adaptation, AfroLlama aims to improve representation quality for low-resource language tasks.

The Flan-T5-small model is an instruction-tuned encoder-decoder sequence-2-sequence Transformer model which improves on the original T5 model by training on over 1,000 diverse tasks, yielding high performance on zero-shot and few-shot natural language processing tasks despite having fewer parameters than massive conversational bots. The ByT5-small model is a byte-level variant of the Text-to-Text Transfer Transformer (T5) models chosen for its robustness to noise and its suitability for spelling and pronunciation tasks. It was designed to work directly on raw UTF-8 bytes and intended to process any language (Xue et al., 2021).

The evaluation aimed to compare their effectiveness in handling spelling variations, phonetic distortions, and complex misspellings.

Evaluation Metrics

The performance of each model was evaluated using seven (7) complementary metrics:

i. Accuracy: measured the proportion of correctly predicted name spellings;

ii. Character-level Precision: measured the proportion of generated corrections that are correct;

iii. Character-level Recall: measured the proportion of correct spellings successfully recovered by the model;

iv. Character-level F1-score: provided the harmonic mean of Precision and Recall;

v. SacreBLEU: evaluated character-level similarity between generated names and reference spellings;

vi. Character Error Rate (CER): measured the normalized edit distance between predicted and reference name spellings; and

vii. McNemar's Significance Test: determined whether the performance differences observed between the Transformer architectures were statistically significant (Zanga et al., 2025).

The use of multiple evaluation metrics provides a comprehensive assessment of both exact-match performance and character-level correction quality.

Experimental Setup

Curriculum-Based Training Strategy

A curriculum-based training strategy was adopted in this study. Rather than exposing the models to the complete training corpus simultaneously, training was conducted in three sequential phases of increasing data volume. This phased approach began with high-confidence, clean name pairs to establish a strong baseline before

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

gradually introducing increasingly complex, synthetic, and noise-heavy samples to enhance model robustness (Sani et al., 2025).

The curriculum design was motivated by the hypothesis that gradual exposure to spelling variations would enable the models to learn fundamental orthographic patterns before encountering more complex name variants. This methodology mirrors common strategies in sequence-to-sequence tasks where training stability and model convergence are prioritized through systematic data scheduling (Lauc et al., 2024; Li et al., 2021).

The curriculum was designed to expose the models progressively to increasingly complex spelling distortions. Phase 1 focused on relatively small distortions with edit distances of 0–1. Phase 2 expanded the range to 0–3, while Phase 3 exposed the models to distortions extending to edit distance 9. Each phase continued from the model state obtained from the preceding phase.

Because the study did not include an otherwise identical non-curriculum training condition, the experimental design does not permit the independent causal contribution of curriculum learning to be quantified. Accordingly, the curriculum is treated in this study as the adopted training strategy rather than as an experimentally isolated causal factor.

Table 1: Dataset Distribution used during the Three Curriculum Training Phases

Dataset/Stage	Cumulative Training Instances	Real Human Instances	Canonical Names	Edit Distances
Phase 1	58,050*	4,975*	1,690	0-1
Phase 2	115,573*	11,054*	1,690	0-3
Phase 3	122,602*	15,942*	1,690	0-9
Validation	12,550	1,728	185	-
Test	13,558	1,802	186	-

*Value is cumulative from the previous phase and should not be summed. Phase 3 values represents the final values that could be summed.

Note: The canonical-name partition is maintained across all stages.

Dataset Construction and Partitioning

The study used a corpus of 2,061 Hausa personal names comprising 1,705 unique canonical names and 356 corresponding spelling variants as found in (Bernard & Ajah, 2025). From these names, a total of 148,710 spelling instances was constructed. The dataset consisted of 19,472 real human spelling attempts collected through the spelling survey and synthetically generated spelling variants designed to represent additional levels of orthographic distortion.

To minimize the possibility of canonical-name-level data leakage, the 2,061 canonical names were partitioned into mutually exclusive training, validation, and test sets. The training partition contained 1,690 canonical names, while the validation and test partitions contained 185 and 186 canonical names, respectively. Consequently, the canonical names represented in the validation and test sets were not present in the training partition. This partitioning strategy ensured that evaluation included previously unseen canonical names rather than merely unseen spelling variants of names encountered during training.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

The training data were subsequently presented using a three-phase cumulative curriculum. The curriculum progressively increased the range of spelling distortions presented to the models as shown in Table 1. Phase 1 contained 58,050 cumulative training instances corresponding to edit distances of 0–1 and included 4,975 real human spelling attempts. Phase 2 increased the cumulative training set to 115,573 instances and extended the edit-distance range to 0–3, including 11,054 cumulative human spelling attempts. Phase 3 contained 122,602 cumulative instances with edit distances ranging from 0–9 and included 15,942 cumulative human spelling attempts.

The human-attempt counts across the three training phases are cumulative rather than independent. Thus, the 15,942 human attempts reported for Phase 3 represent the cumulative human component of the training data rather than an additional 15,942 attempts introduced after Phases 1 and 2. Together with the 1,728 human attempts in validation and 1,802 human attempts in the test set, these values account for all 19,472 real human spelling attempts in the corpus.

The validation and test sets contained 12,550 and 13,558 instances, respectively. Importantly, the test set contained 1,802 real human spelling attempts, while the validation set contained 1,728. The remaining instances in these sets were synthetically generated variants. The aggregate evaluation results reported in this study therefore represent performance over mixed real-human and synthetic test instances. The complete dataset can be found at https://doi.org/10.5281/zenodo.21852788.

Fine-Tuning Configuration

ByT5-small, Flan-T5-small, and AfroLlama were evaluated using a common principal training protocol. The same canonical-name partitioning, curriculum schedule, evaluation procedure, and principal training settings were maintained across the three model configurations to provide a controlled comparison. Parameter-efficient fine-tuning was employed using LoRA/PEFT.

Each curriculum phase was trained for one epoch, resulting in three successive training phases. The principal training configuration used a training batch size of four, evaluation batch size of eight, learning rates of 2×10^{-4} in Phase 1 and 1×10^{-4} in Phases 2 and 3, with model selection based on validation loss. Training and evaluation were conducted using Google Colab GPU resources, including T4 and A100 instances. The training configuration is summarized in Table 2.

The use of a common configuration was intended to control experimental conditions rather than to imply that the selected hyperparameters were individually optimal for every architecture. Transformer architectures may differ in their sensitivity to learning rate, parameter-efficient adaptation, tokenization, batch size, and other optimization choices. Consequently, the results reported here represent the performance of the evaluated models under the common experimental protocol, while architecture-specific hyperparameter optimization remains an area for further investigation.

Table 2: Training Configuration

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Curriculum Phases	3	Training Batch Size	4
Epochs per Phase	1	Evaluation Batch Size	8
Total Effective Epochs	3	Evaluation Strategy	Steps

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Phase Learning Rate	2×10^{-4}	Checkpoint Saving	Every 200 Steps
Phase 2 Learning Rate	1×10^{-4}	Logging Frequency	Every 200 Steps
Phase 3 Learning Rate	1×10^{-4}	Model Selection Criterion	Validation Loss

RESULTS

Comparative Model Performance

The evaluation results shown in Table 3 demonstrate substantial differences among the three evaluated model configurations. Before fine-tuning, ByT5-small achieved an exact-match accuracy of 0.53%, while fine-tuning increased its accuracy to 24.55%. Its character-level F1 increased from 58.09% to 86.16%, accompanied by an improvement in SacreBLEU from 38.7581 to 69.7286 and a reduction in CER from 0.6304 to 0.2048. These changes indicate substantial improvement in the model's ability to transform distorted names toward their canonical forms.

Flan-T5-small similarly benefited from fine-tuning. Its exact-match accuracy increased from 0.63% to 11.85%, while character F1 increased from 53.81% to 83.14%. SacreBLEU increased from 25.6284 to 62.2666 and CER decreased from 1.3567 to 0.2346. Thus, both ByT5-small and Flan-T5-small demonstrated substantial gains following adaptation to the Hausa personal-name corpus. AfroLlama exhibited a different response. Its base configuration achieved 6.84% exact-match accuracy and 74.93% character F1. Following fine-tuning, exact-match accuracy decreased to 4.46% and character F1 decreased to 69.70%. Its CER increased from 0.3523 to 0.4483, although SacreBLEU increased slightly from 48.1126 to 50.7740. Therefore, fine-tuning did not produce a uniform improvement across the evaluated models. These results establish that, under the experimental conditions examined, fine-tuned ByT5-small produced the strongest overall performance, followed by Flan-T5-small. AfroLlama produced a different adaptation pattern in which fine-tuning was associated with deterioration in several evaluation measures. These findings are specific to the three evaluated model configurations and should not be generalized to all encoder-decoder or causal language models.

Table 3: Comparative Result of the Fine-tuned Models Performances

Model	Correct↑	Accuracy↑ (%)	Precision↑* (%)	Recall↑* (%)	F1 Score↑* (%)	SacreBLUE↑ (%)	CER↓
ByT5 small Base	**72**	**0.53**	**58.18**	**58.01**	**58.09**	**38.7581**	**0.6304**
ByT5 small Fine-tuned	**3328**	**24.55**	**92.45**	**80.68**	**86.16**	**69.7286**	**0.2048**
Flan-T5 small Base	**86**	**0.63**	**39.42**	**84.73**	**53.81**	**25.6284**	**1.3567**
Flan-T5 Small Fine-tuned	**1606**	**11.85**	**88.68**	**78.24**	**83.14**	**62.2666**	**0.2346**
AfroLlama Base	**928**	**6.84**	**80.68**	**69.95**	**74.93**	**48.1126**	**0.3523**
AfroLlama Fine-tuned	**605**	**4.46**	**70.06**	**69.35**	**69.7**	**50.774**	**0.4483**

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

*Character-level score

↑: better with increasing values

↓: better with decreasing values

Exact-Match Accuracy and Character-Level Performance

An important characteristic of the results is the difference between exact-match accuracy and character-level measures. Fine-tuned ByT5-small achieved an exact-match accuracy of 24.55%, corresponding to 3,328 correctly restored names out of 13,558 test instances. In contrast, the model achieved a character-level F1 of 86.16%, precision of 92.45%, recall of 80.68%, and CER of 0.2048.

These measures capture different aspects of correction quality. Exact-match accuracy requires the complete generated name to be identical to the target canonical name. Consequently, a prediction containing only a small residual character discrepancy is counted as incorrect. Character-level precision, recall, and F1, however, capture the extent to which the generated string corresponds to the target at character level. Similarly, a lower CER indicates that fewer character-level modifications are required to transform the prediction into the reference. The combination of relatively low exact-match accuracy with substantially higher character-level F1 therefore indicates that many model outputs were close to their intended canonical forms without being completely correct. This distinction is particularly important for personal-name correction, where a single remaining character error can cause an otherwise highly similar prediction to fail an exact-match evaluation.

The results should therefore not be interpreted as indicating that the model corrected only 24.55% of names in a broader sense. Rather, 24.55% represents complete-string correction, while the substantially higher character-level scores indicate considerable partial correction quality. A detailed taxonomy of residual substitution, insertion, deletion, and transposition errors was not performed in the present study; such analysis would require prediction-level error categorization and is therefore identified as an area for future work.

Effect of Fine-Tuning

The comparison between the base and fine-tuned configurations provides evidence that corpus-specific adaptation substantially changed model behaviour. ByT5-small demonstrated the largest absolute improvement in exact-match accuracy, increasing from 0.53% to 24.55%. Flan-T5-small increased from 0.63% to 11.85%. These improvements were accompanied by higher character-level F1 and SacreBLEU scores and lower CER values. The improvements are consistent with the expectation that adaptation to a domain-specific spelling-correction corpus can provide models with knowledge of the orthographic patterns and transformations represented in the training data. The effect was nevertheless model-dependent, as AfroLlama did not show the same improvement pattern. Therefore, the results support the effectiveness of fine-tuning for the evaluated ByT5-small and Flan-T5-small configurations, but they do not support the assumption that fine-tuning necessarily improves every model under an identical training procedure.

AfroLlama Fine-Tuning Behaviour

AfroLlama exhibited a distinct response to fine-tuning. Its exact-match accuracy decreased from 6.84% in the base configuration to 4.46% after fine-tuning. Character F1 similarly decreased from 74.93% to 69.70%, while CER increased from 0.3523 to 0.4483. SacreBLEU, however, increased modestly from 48.1126 to 50.7740.

The training and validation loss trajectories provide additional evidence regarding this behaviour. Training loss continued to decrease across the curriculum, whereas validation loss initially improved before increasing during the later training stage. The lowest validation loss occurred around the second phase, after which validation performance deteriorated. This pattern is consistent with a possible loss of generalization during later adaptation.

However, the observed pattern does not establish a single causal explanation for the degradation. Possible contributing factors include overfitting, interaction between the AfroLlama architecture and the selected parameter-

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

efficient fine-tuning configuration, tokenization characteristics, or sensitivity to the selected learning schedule. These possibilities were not independently isolated in the present study and should therefore be regarded as hypotheses rather than established causes.

The principal empirical conclusion is more limited but robust: under the training configuration used in this study, fine-tuning did not improve AfroLlama's performance on the evaluation measures in the same manner observed for ByT5-small and Flan-T5-small.

Generalization to Unseen Canonical Names

An important feature of the experimental design is that canonical names were partitioned before model evaluation. The training set contained 1,690 canonical names, whereas validation and test contained 185 and 186 distinct canonical names, respectively. Consequently, the evaluation was conducted on canonical names that were not represented in the training partition.

This design provides a stronger test of generalization than randomly splitting individual spelling variants while allowing the same canonical name to occur across training and testing. The reported test performance therefore reflects the models' ability to correct spelling variations associated with previously unseen canonical names under the conditions of the study.

Nevertheless, generalization should be interpreted within the scope of the available corpus. The test set represents 186 canonical names, and broader evaluation involving additional names, speakers, geographical varieties, and naturally occurring administrative records would provide further evidence of external validity.

Real Human and Synthetic Spelling Variants

The dataset combines real human spelling attempts with synthetically generated spelling variants. This design provides both authentic evidence of spelling behaviour and a mechanism for increasing coverage of distortion patterns. Of the 19,472 real human spelling attempts, 15,942 were included cumulatively in the training curriculum, while 1,728 and 1,802 were included in the validation and test sets, respectively. The aggregate results reported in this study combine human-generated and synthetic instances. Consequently, the reported test metrics should not be interpreted as a human-only estimate of model performance. Although the inclusion of real human attempts strengthens the ecological relevance of the dataset, a separate evaluation of human-generated test instances would provide a more direct estimate of performance on naturally occurring spelling behaviour.

Such stratified evaluation was not conducted in the present experimental cycle and is therefore identified as an important direction for future work.

Training Loss Analysis

The training loss curves of the ByT5 small, Flan-T5 small and AfroLlama models were plotted across training steps and phases to monitor the learning progress.

ByT5 Small Training Loss Analysis

Figures 1 and 2 illustrate the training and validation loss curves of the ByT5 small model across the three curriculum learning phases. Table 4 gives an observed trend in the training and evaluation losses data.

Table 4: High-Level Summary of ByT5 Small Curriculum Phased Training

Phase	Training Loss Trend	Validation Loss Trend	Interpretation
Phase 1	~4.6 → ~2.55	~3.64 → ~2.49	Strong learning (rapid improvement)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Phase 2	~2.61 → ~2.52	~2.48 → ~2.43	Slow refinement
Phase 3	~2.57 → ~2.50	~2.43 → ~2.40	Convergence

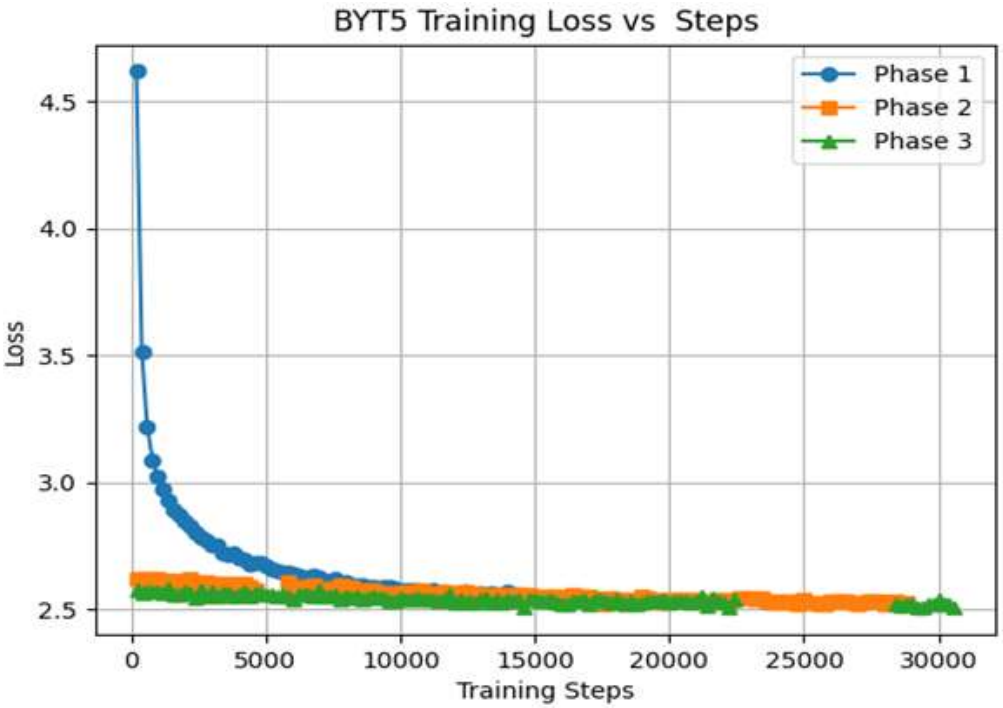

Figure 1: ByT5 Small Training Loss Curves Across all Three Phases of Curriculum Phased Training

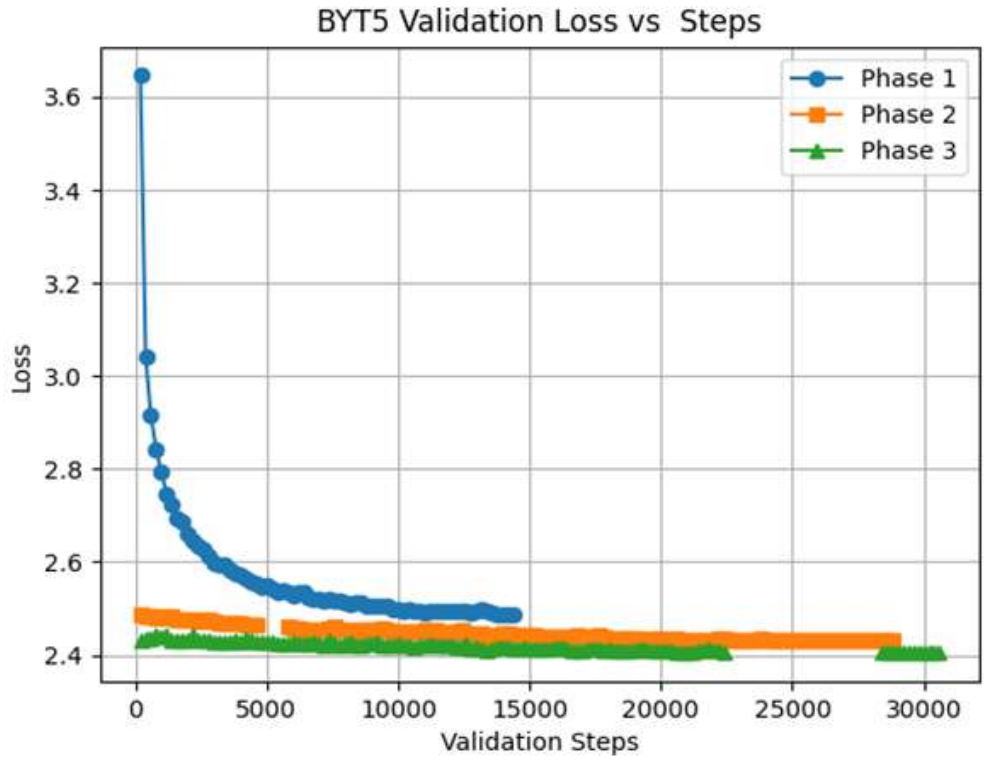

Figure 2: ByT5 Small Validation Loss Curves Across all Three Phases of Curriculum Phased Training

In Phase 1, a sharp decline in loss is observed, indicating rapid learning of fundamental name structures and character patterns. This is where the model learns spelling patterns, character sequences and common name forms.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

In Phase 2, the loss curves exhibits minor fluctuations, reflecting the introduction of more complex and noisy training data. Despite this, both training and validation losses continue to decrease gradually, demonstrating the model's ability to adapt to misspelt inputs.

In Phase 3, the loss curves stabilize, indicating convergence. The small gap between training and validation loss suggests that the model generalized well without significant over fitting.

Flan-T5 Small Training Loss Analysis

Figures 3 and 4 present the training and validation loss curves of the Flan-T5 small model across the three curriculum learning phases. Table 5 gives an observed trend in the training and evaluation loss data.

Table 5: High-Level Summary of Flan-T5 Small Curriculum Phased Training

Phase	Training Loss Trend	Validation Loss Trend	Interpretation
Phase 1	~8.61 → ~6.73	~7.51 → ~6.75	Learning is slow with high loss
Phase 2	~6.86 → ~6.71	~6.74 → ~6.64	Minimal improvement; model struggling to adapt
Phase 3	~6.76 → ~6.65	~6.64 → ~6.59	Early but poor convergence

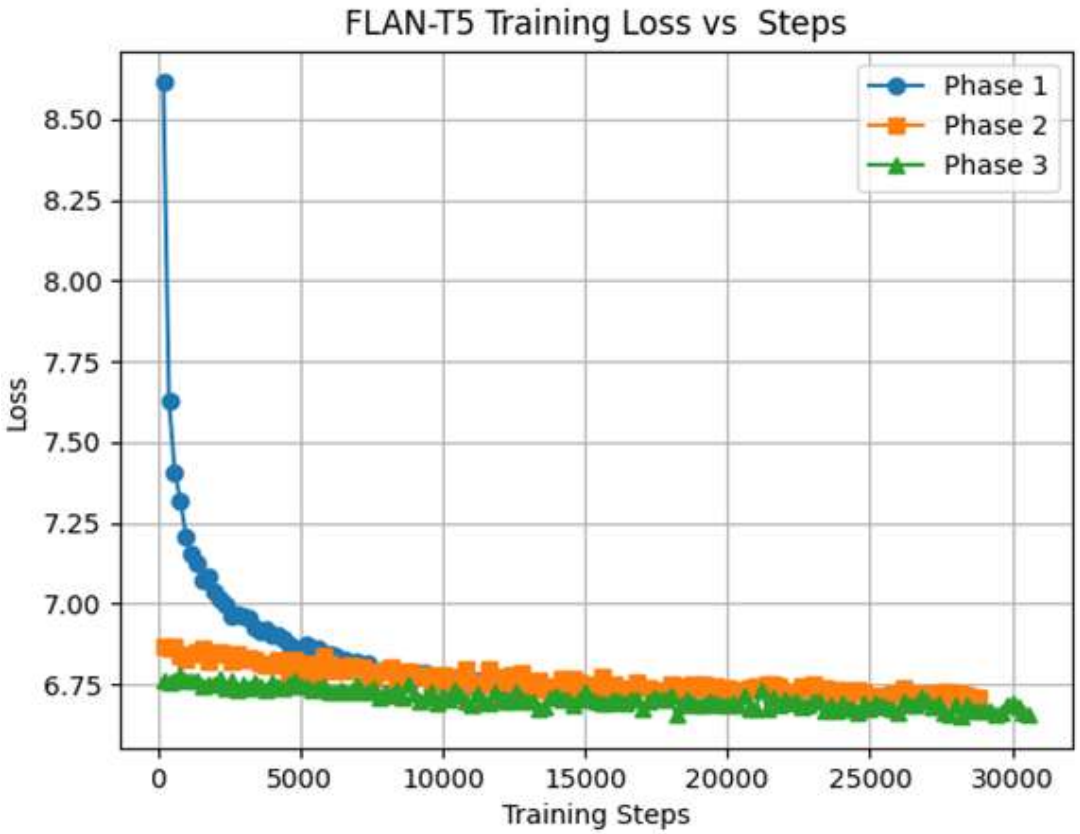

Figure 3: Flan-T5 Small Training Loss Curves Across all Three Phases of Curriculum Phased Training

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

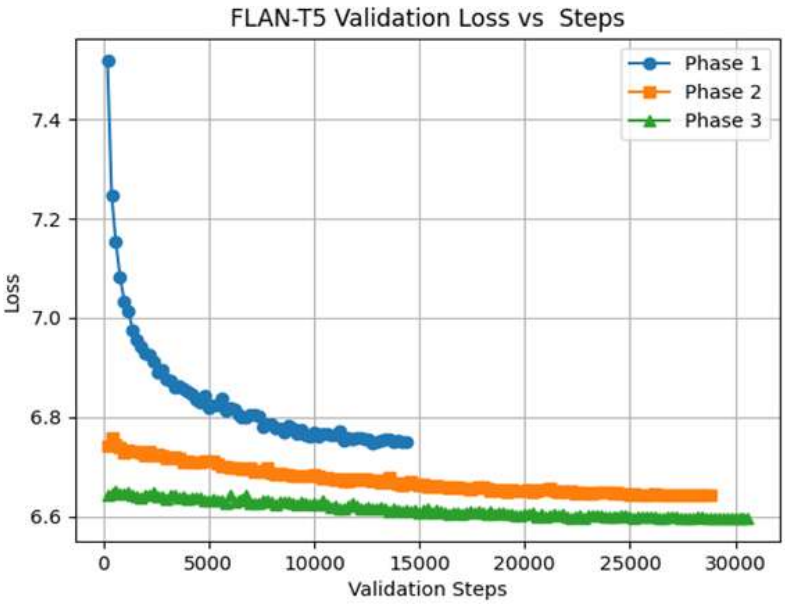

Figure 4: Flan-T5 Small Validation Loss Curves Across all Three Phases of Curriculum Phased Training

In Phase 1, the model demonstrates a gradual decrease in loss, indicating initial learning. However, the rate of decrease is relatively slow compared to ByT5 small, and the loss values remain high.

In Phase 2, the loss curves become relatively flat, suggesting limited adaptation to more complex and noisy training data. This indicates that the model struggles to learn fine-grained spelling correction patterns with a low sensitivity to noisy data.

In Phase 3, the model converges with only marginal improvements in both training and validation loss. The final loss values remain significantly higher than those observed in the ByT5 small model.

This behaviour suggests that Flan-T5 small does not effectively capture the underlying structure required for the name correction task, which explains its lower recall (78.24%) and overall performance during evaluation, as seen in Table 3. One possible root cause of this problem stems from the instruction-tuned nature of the Flan-T5 small model as opposed to ByT5, which is modelled at the character level.

AfroLlama Training Loss Analysis

Figures 5 and 6 present the training and validation loss curves for AfroLlama across the three curriculum learning phases. Table 6 gives an observed trend in the training and evaluation loss data.

Table 6: High-Level Summary of AfroLlama Curriculum Phased Training

Phase	Training Loss Trend	Validation Loss Trend	Interpretation
Phase 1	~0.48 → ~0.30	~0.54 → ~0.63*	Strong learning followed by overfitting
Phase 2	~0.39 → ~0.32	~0.51 → ~0.58*	Best generalization performance; rapid convergence
Phase 3	~0.32 → ~0.31	~0.58 → ~0.60*	Fine-tuning and convergence

* Validation loss initially decreased to its minimum value before increasing later in training

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

In all phases, training loss decreased steadily, indicating successful learning of the Hausa name spelling correction task. Validation loss also decreased initially before reaching a minimum value and subsequently increasing, suggesting the onset of overfitting as training progressed.

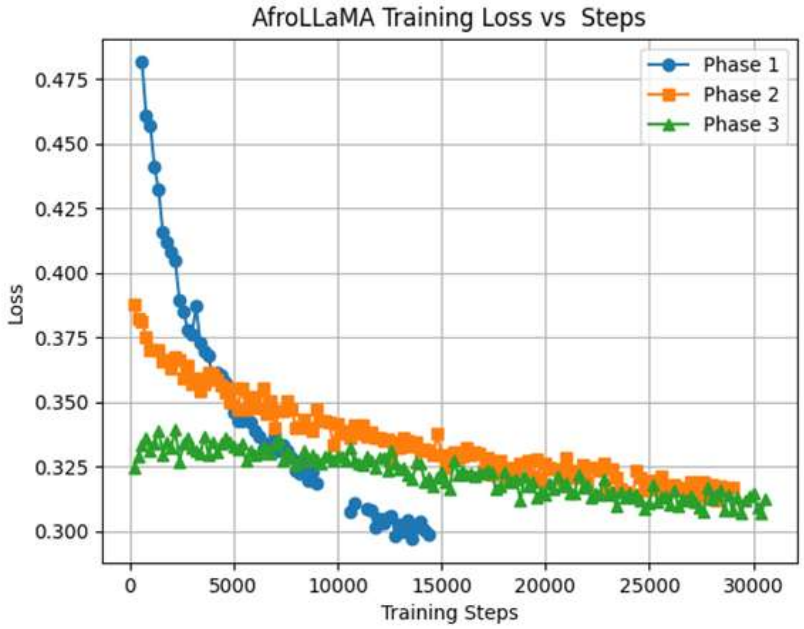

Figure 5: AfroLlama Training Loss Curves Across all Three Phases of Curriculum Phased Training

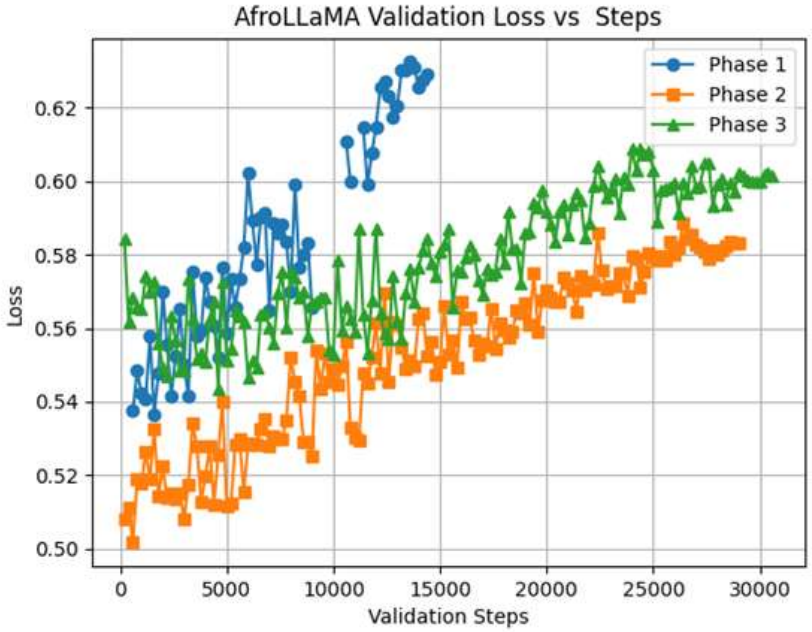

Figure 6: AfroLlama Validation Loss Curves Across all Three Phases of Curriculum Phased Training

Phase 2 achieved the lowest validation loss (0.502), outperforming Phases 1 and 3, and indicating that this stage contributed the greatest improvement in model generalization.

The rapid convergence observed in Phases 1 and 2 further suggests that AfroLlama effectively transferred knowledge between curriculum stages.

In Phase 3, both training and validation losses changed more gradually, indicating that the model was approaching convergence and primarily refining previously learned representations.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Overall, the loss curves demonstrate that the curriculum learning strategy enabled AfroLlama to learn progressively more complex spelling patterns while maintaining stable optimization behaviour. The results also highlight the importance of selecting checkpoints based on minimum validation loss rather than final training loss to maximize generalization performance.

Statistical Significance Analysis

To determine whether the observed performance differences between the evaluated Transformer models were statistically significant, McNemar's test was conducted on the prediction outcomes of the models using the test dataset.

Statistical Significance Analysis for ByT5 Small Vs Flan-T5 Small Models

Figure 7 and Figure 8 present the contingency table and the McNemar's Test for ByT5 small and Flan-T5 small models, respectively. Using the discordant pairs in Figure 7, the odds ratio is calculated thus:

i. b (ByT5 wrong and Flan-T5 correct) = 257
ii. c (ByT5 correct and Flan-T5 wrong) = 1979

Odds ratio, OR = 1979/257= 7.69

BYT5 * FlanT5 Crosstabulation

Count

		FlanT5		
		Model prediction was wrong	Model prediction was correct	Total
BYT5	Model prediction was wrong	9973	257	10230
	Model prediction was correct	1979	1349	3328
Total		11952	1606	13558

Figure 7: Cross-tabulation for ByT5 Small Vs Flan-T5 Small Models

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		<.001[a]
N of Valid Cases	13558	

a. Binomial distribution used.

Figure 8: Chi-Square Test for ByT5 Small Vs Flan-T5 Small Models

The contingency table shows that ByT5 small correctly predicted 1,979 instances that were incorrectly predicted by Flan-T5 small, whereas Flan-T5 correctly predicted only 257 instances that were incorrectly predicted by

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

ByT5 small. This yields an odds ratio that indicates that ByT5 small is 7.69 times more likely than Flan-T5 small to correctly predict the spelling of a Hausa personal name.

Moreover, the corresponding McNemar's test confirmed that the difference was statistically significant ($p < 0.001$). Therefore, the null hypothesis that both models exhibit equivalent predictive performance is rejected.

The result indicates that ByT5 small significantly outperforms Flan-T5 small on the Hausa personal name spelling correction task. The finding provides statistical support for the superior Accuracy, F1-score, SacreBLEU, and Character Error Rate achieved by ByT5 small during the experimental evaluation.

Statistical Significance Analysis for ByT5 Small Vs AfroLlama Models

Based on the discordant pairs (ByT5 small correctly predicting 3007 instances where AfroLlama failed, while wrongly predicting 284 instances where AfroLlama was correct) as shown in Figure 9, ByT5 small appears to be 10.59 times more likely to correctly predict the spelling of a Hausa personal name than the AfroLlama model.

BYT5 * afroLlama Crosstabulation

Count

		afroLlama		
		Model prediction was wrong	Model prediction was correct	Total
BYT5	Model prediction was wrong	9946	284	10230
	Model prediction was correct	3007	321	3328
Total		12953	605	13558

Figure 9: Cross-tabulation for ByT5 Small Vs AfroLlama Models

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		<.001[a]
N of Valid Cases	13558	

a. Binomial distribution used.

Figure 10: Chi-Square Test for ByT5 Small Vs AfroLlama Models

Accordingly, the corresponding McNemar's test (Figure 10) confirmed that the difference was statistically significant ($p < 0.001$). Therefore, the null hypothesis that both models exhibit equivalent predictive performance is rejected.

The result indicates that ByT5 small significantly outperforms AfroLlama on the Hausa personal name spelling correction task. The finding provides statistical support for the superior Accuracy, F1-score, SacreBLEU, and Character Error Rate achieved by ByT5 small during the experimental evaluation and further demonstrates the effectiveness of byte-level representations for character-sensitive spelling correction tasks involving Hausa personal names.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Statistical Significance Analysis for Flan-T5 Small Vs AfroLlama Models

Based on the discordant pairs (Flan-T5 small correctly predicting 1407 instances where AfroLlama failed, while wrongly predicting 406 instances where AfroLlama was correct) as shown in Figure 11, Flan-T5 small appears to be 3.47 times more likely to correctly predict the spelling of a Hausa personal name than the AfroLlama model.

FlanT5 * afroLlama Crosstabulation

Count

		afroLlama		
		Model prediction was wrong	Model prediction was correct	Total
FlanT5	Model prediction was wrong	11546	406	11952
	Model prediction was correct	1407	199	1606
Total		12953	605	13558

Figure 11: Cross-tabulation for Flan-T5 Small Vs AfroLlama Models

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		<.001[a]
N of Valid Cases	13558	

a. Binomial distribution used.

Figure 12: Chi-Square Test for Flan-T5 Small Vs AfroLlama Models

Accordingly, the corresponding McNemar's test (Figure 12) confirmed that the difference was statistically significant ($p < 0.001$). Therefore, the null hypothesis that both models exhibit equivalent predictive performance is rejected. Table 7 presents the comparative statistical significance tests for the three fine-tuned models. The results indicate statistically significant differences between all model pairs. In every comparison, the null hypothesis of equivalent predictive performance was rejected.

Table 7: Comparative McNemar's Statistical Tests for ByT5 Small, Flan-T5 Small and AfroLlama Models

Comparison	Model A Correct / Model B Wrong	Model B Correct / Model A Wrong	Odds Ratio	McNemar p-value	Result
ByT5 Small vs Flan-T5 Small	1,979	257	7.70	< 0.001	Significant

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

ByT5 Small vs AfroLlama	3,007	284	10.59	< 0.001	Significant
Flan-T5 Small vs AfroLlama	1,407	406	3.47	< 0.001	Significant

DISCUSSION

The models were evaluated using standard performance metrics, including exact-match accuracy, character-level precision, recall, F1 score, sacreBLUE, and CER scores.

The study employed a three-phase cumulative curriculum-learning strategy in which the training data progressively increased in both volume and spelling distortion. Phase 1 comprised 58,050 instances with edit distances of 0–1, Phase 2 increased the cumulative training set to 115,573 instances with edit distances of 0–3, and Phase 3 increased it further to 122,602 instances with edit distances of 0–9. Each phase continued from the model state obtained in the preceding phase.

The results demonstrate the performance obtained when the evaluated models were fine-tuned using this curriculum strategy. However, because the study did not include a matched non-curriculum training condition, the independent contribution of curriculum learning cannot be separated from the effects of model architecture, fine-tuning, and the progressively expanded training data. Accordingly, the present study does not claim that curriculum learning alone caused the observed performance differences.

Instead, curriculum learning should be regarded as the controlled training strategy under which the models were evaluated. A future ablation study comparing the proposed curriculum with an otherwise identical conventional training schedule would provide stronger evidence regarding the specific contribution of curriculum learning.

Also, the dataset used combined naturally occurring human spelling attempts with synthetically generated spelling variants. This design was adopted to provide both authentic spelling-error patterns and sufficient training diversity for the Transformer models. The complete dataset contains 19,472 real human spelling attempts, with the remaining instances generated synthetically.

The real human spelling attempts were distributed across the cumulative curriculum-training phases and the evaluation sets. Phase 1 contained 4,975 real human spelling attempts, Phase 2 contained 11,054, and Phase 3 contained 15,942. Because the curriculum phases are cumulative, these figures represent the number of real human attempts available at each successive training stage rather than independent observations. The validation and test sets contained 1,728 and 1,802 real human spelling attempts, respectively.

The inclusion of real human spelling attempts in the test set provides an important connection to the practical spelling-correction problem. Nevertheless, the reported aggregate evaluation metrics combine real human and synthetically generated instances. Therefore, the results should be interpreted as performance on the overall evaluation distribution rather than as a direct estimate of performance exclusively on naturally occurring human spelling errors. A separate stratified evaluation of human-generated and synthetic variants would provide additional evidence regarding real-world generalization and is recommended for future work.

To provide a controlled comparison, the three evaluated models were fine-tuned using the same overall training protocol, including the same dataset partitions, curriculum schedule, evaluation procedure, and principal training settings. The use of a common training configuration was intended to reduce variation attributable to differences in experimental procedure and to ensure that the models were evaluated under comparable conditions.

However, identical hyperparameter settings should not be interpreted as demonstrating that the selected configuration was optimal for every architecture. Transformer architectures can differ in their sensitivity to

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

learning rate, batch size, optimization dynamics, parameter-efficient fine-tuning configuration, and other training choices. Therefore, the present comparison evaluates the relative performance of the three models under a common experimental configuration rather than under independently optimized configurations for each architecture.

This distinction is important when interpreting the results, particularly for AfroLlama. Its lower post-fine-tuning performance may reflect, at least in part, an interaction between the model architecture and the selected training configuration. Establishing the best architecture-specific configuration would require systematic hyperparameter optimization, which was outside the scope of the present study.

Overall, the experiments demonstrate that domain-specific adaptation substantially improved the evaluated ByT5-small and Flan-T5-small configurations, with ByT5-small achieving the strongest results across the principal correction measures. The disparity between exact-match accuracy and character-level F1 further demonstrates that name correction quality cannot be fully represented by a single metric. The degradation observed for AfroLlama highlights the importance of model-specific adaptation and cautions against assuming that fine-tuning produces uniform improvements across architectures.

The canonical-name-level partitioning strengthens the validity of the evaluation by ensuring that the canonical names in validation and testing were not present in training. Nevertheless, the combined use of human and synthetic instances, absence of a curriculum ablation, and use of a common rather than architecture-optimized configuration limit the extent to which causal or universal conclusions can be drawn. These limitations define clear directions for subsequent research.

CONCLUSION

This study investigated the application of Transformer-based language models to the automatic correction of misspelt Hausa personal names using a corpus of 148,710 spelling instances derived from 2,061 canonical names. The corpus incorporated 19,472 real human spelling attempts together with synthetically generated spelling variants. To reduce the possibility of canonical-name-level data leakage, the canonical names were partitioned into mutually exclusive training, validation, and test sets containing 1,690, 185, and 186 names, respectively.

ByT5-small, Flan-T5-small, and AfroLlama were evaluated using a three-phase cumulative curriculum-learning strategy and assessed using exact-match accuracy, character-level precision, recall and F1, SacreBLEU, and character error rate. Among the evaluated configurations, fine-tuned ByT5-small achieved the strongest overall performance, obtaining 24.55% exact-match accuracy, 92.45% character precision, 80.68% character recall, 86.16% character F1, 69.7286 SacreBLEU, and 0.2048 CER. Fine-tuned Flan-T5-small achieved 11.85% exact-match accuracy and 83.14% character F1. In contrast, AfroLlama experienced a reduction in exact-match accuracy from 6.84% to 4.46% following fine-tuning.

McNemar's tests showed statistically significant differences between each pair of evaluated models ($p < 0.001$). The results therefore provide evidence that, under the experimental configuration examined, ByT5-small was the most effective of the three evaluated configurations for Hausa personal-name spelling correction.

The relatively low exact-match accuracy compared with the substantially higher character-level F1 demonstrates that complete-string correction and partial character-level correction capture different dimensions of model performance. The findings also show that fine-tuning does not necessarily produce uniform gains across model architectures, as demonstrated by the different response of AfroLlama.

The conclusions of this study are intentionally restricted to the evaluated models, dataset, and training configuration. The results do not establish the superiority of an entire class of sequence-to-sequence models over causal language models, nor do they establish the causal contribution of curriculum learning because a non-curriculum control condition was not evaluated. Similarly, the observed AfroLlama degradation identifies an empirical behaviour rather than a definitively established causal mechanism.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

LIMITATIONS AND RECOMMENDATIONS

Several limitations should be considered when interpreting the findings. First, although the dataset contains substantial numbers of real human spelling attempts, the reported aggregate metrics combine real human and synthetically generated spelling variants. Consequently, the results do not provide a separate estimate of model performance on naturally occurring human spelling errors. Future studies should report stratified performance for human-generated and synthetic variants.

Second, the three curriculum phases were designed as cumulative training stages; therefore, the present study demonstrates model performance under the proposed curriculum strategy but does not isolate the independent contribution of curriculum learning. A matched curriculum-versus-conventional-training ablation would be required to establish such a causal contribution.

Third, the models were evaluated under a common principal training configuration. Although this provides a controlled comparison, it does not establish that the selected hyperparameters were optimal for each architecture. Architecture-specific optimization may produce different relative performance.

Fourth, the study evaluates only ByT5-small, Flan-T5-small, and the selected AfroLlama configuration. Consequently, the findings should not be generalized to Transformer architectures or causal language models as classes. In particular, the observed degradation of AfroLlama after fine-tuning establishes the behaviour of the evaluated configuration but does not establish a general limitation of causal language models.

Finally, the study did not include controlled numerical baselines based on candidate retrieval using Levenshtein distance, Jaro-Winkler, or Sautex under the same test protocol. Such comparisons would be valuable in future work, particularly for determining when generative correction provides advantages over phonetic and similarity-based candidate matching.

ACKNOWLEDGEMENTS

The authors are grateful to all those who have critiqued this work in part or whole and provided valid insights to its completion.

Funding

Not applicable.

Author Contributions

Bernard Ephraim conceived and designed the various aspect of this work and carried out the experiments herein under the supervision and guidance of Ajah Ifenyinwa.

Declaration

Competing Interests

The authors declare that they have no competing interests.

Data Availability

i. The datasets generated and/or analyzed during this study are available at https://doi.org/10.5281/zenodo.21852788.
ii. The fine-tuned ByT5 small model is hosted on Hugging Face at https://huggingface.co/BernardEphraim/name-spell.
iii. This model can be accessed for inference on Hugging Face at https://huggingface.co/spaces/BernardEphraim/name-spell-api.

REFERENCES

1. Aars, C., Adams, L., Tian, X., Wang, Z., Wismer, C., Wu, J., Rivas, P., Sooksatra, K., & Fendt, M. (2024). Efficacy of ByT5 in Multilingual Translation of Biblical Texts for Underrepresented Languages. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2405.13350
2. Aliakbarzadeh, A., Flek, L., & Karimi, A. (2025). Exploring Robustness of Multilingual LLMs on Real-World Noisy Data. In *Qeios*. https://doi.org/10.32388/3x6cxv
3. Aliero, A. A., Bashir, S. A., Aliyu, H. O., Tafida, A. G., & Hussaini, M. (2025). Dual-BERT Adversarial Model for Text Normalization in Hausa User-Generated Contents. In *Research Square*. https://doi.org/10.21203/rs.3.rs-7446019/v1
4. Al-Rfooh, B., Abandah, G. A., & Al-Rfou, R. (2023). Fine-Tashkeel: Finetuning Byte-Level Models for Accurate Arabic Text Diacritization. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2303.14588
5. Bernard, E., & Ajah, A. I. (2025). Sautex: A Language-Specific Phonetic Matching Algorithm for Resolving Spelling Variations in Hausa Personal Names. *Journal of Computer, Software and Program*., *2*(2), 25–33. https://doi.org/10.69739/jcsp.v2i2.1141
6. Christen, P. (2006). A Comparison of Personal Name Matching: Techniques and Practical Issues. *The Austrilian National University*. 290–294. https://doi.org/10.1109/icdmw.2006.2
7. Feher, D., Vulić, I., & Minixhofer, B. (2024). Retrofitting Large Language Models with Dynamic Tokenization. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2411.18553
8. Ismail, K., Abdou, S., Farouk, M., & Salem, A. (2025). Transformers to the rescue: alleviating data scarcity in arabic grammatical error correction with pre-trained models. *Neural Computing and Applications, 37*(18), 13011–13038. https://doi.org/10.1007/s00521-025-11145-1
9. Khaleel, M. R., & Abandah, G. A. (2025). Efficient Stochastic Error Injection for Optimizing Large Language Models in Arabic Spelling Correction. *2025 International Conference on New Trends in Computing Sciences (ICTCS), Amman, Jordan* . 505–510. https://doi.org/10.1109/ictcs65341.2025.10989319
10. Kuparinen, O., Miletić, A., & Scherrer, Y. (2023). Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation. *Findings of the Association for Computational Linguistics: EMNLP 2023*. https://doi.org/10.18653/v1/2023.findings-emnlp.923
11. Lauc, D., Rutherford, A., & Wongwarawipatr, W. (2024). AyutthayaAlpha: A Thai-Latin Script Transliteration Transformer. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2412.03877
12. Lefrandt, M., Santoso, E. B., Gunawan, A. A. S., & Tedjasulaksana, J. J. (2025). Contextual Spelling Corrector for Indonesian Text Preprocessing: A Comparative Analysis of Large Language Models. *2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Bali, Indonesia.* 290–296. https://doi.org/10.1109/iaict65714.2025.11100636
13. Li, H., Li, J., Jiang, W., Zhang, Z., Chen, M., Wang, S., & Xiao, J. (2021). PHMOSpell: Phonological and Morphological Knowledge Guided Chinese Spelling Check. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 5958–5967. https://doi.org/10.18653/v1/2021.acl-long.464
14. Limisiewicz, T., Blevins, T., Gonen, H., Ahia, O., & Zettlemoyer, L. (2024). MYTE: Morphology-Driven Byte Encoding for Better and Fairer Multilingual Language Modeling. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2403.10691
15. Lutgen, A.-M., Plum, A., Purschke, C., & Plank, B. (2024). Neural Text Normalization for Luxembourgish using Real-Life Variation Data. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2412.09383
16. Okewunmi, P., James, F., & Fajemila, O. E. (2025). Evaluating Robustness of LLMs to Typographical Noise in Yorùbá QA. *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025.* 195–202. https://doi.org/10.18653/v1/2025.africanlp-1.29

17. Owodunni, A. T., Ahia, O., & Kumar, S. (2025). FLEXITOKENS: Flexible Tokenization for Evolving Language Models. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2507.12720
18. Phonchai, T., Siripong, S., Patterson, N., & Campbell, O. (2025). Large Language Models for Zero-Shot Multicultural Name Recognition. *arXiv (Cornell University)*. http://arxiv.org/abs/2507.04149
19. S, B. R., Suri, G., Dewangan, V., & Sonavane, R. (2024). When Every Token Counts: Optimal Segmentation for Low-Resource Language Models. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2412.06926
20. Sani, S. A., Muhammad, S. H., & Jarvis, D. (2025). Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2501.11023
21. Sperduti, G., & Moreo, A. (2026). Misspellings in natural language processing: A survey of recent literature. *Natural Language Processing., 32*(2), 113–159. https://doi.org/10.1017/nlp.2026.10020
22. Wali, A. M., & Nisioi, S. (2025). Automatic Correction of Writing Anomalies in Hausa Texts. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2506.03820
23. Wu, S., Tan, X., Wang, Z., Wang, R., Li, X., & Sun, M. (2024). Beyond Language Models: Byte Models are Digital World Simulators. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2402.19155
24. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A. P., & Raffel, C. (2022). ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics, 10*, 291–306. https://doi.org/10.1162/tacl_a_00461
25. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2021). ByT5: Towards a token-free future with pre-trained byte-to-byte models. In *arXiv (Cornell University*). Cornell University. https://doi.org/10.48550/arxiv.2105.13626
26. Zanga, A. I., Abdulrahman, S. M., Ado, A., Bichi, A. A., Jibril, L. A., Umar, A. M., Adamu, A., Muhammad, S. H., & Abubakar, B. S. (2025). HausaMovieReview: A Benchmark Dataset for Sentiment Analysis in Low-Resource African Language. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2509.16256