Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.0675

2026, 45(3) 114-123

Bridging the Gap: A Transformer-Based Chatbot Architecture for Low-Resource Brahui via Embedding Fusion and Transfer Learning

Naseer Ahmed [a], Mansoor Ahmed Khuhro [b], Asif Ali Laghari [a]

[a] *Department of Computer Science, Sindh Madressatul Islam University, Pakistan*

[b] *Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Pakistan*

[*] Corresponding author: makhuhro@smiu.edu.pk

Received: 22 December 2025, Revised: 18 May 2026, Accepted: 22 June 2026, Published: 01 July 2026

KEYWORDS

Low-Resource Languages
Brahui Language
Large Language Models
Transformer-Based Chatbot
Embedding Fusion
Transfer Learning

ABSTRACT

Large language models have transformed conversational artificial intelligence; however, their performance remains weak for languages with extreme resource scarcity. Brahui is a typologically distinct Dravidian language with limited annotated data and minimal representation in multilingual resources. This study presents a transformer-based chatbot architecture designed for Brahui and introduces an embedding fusion mechanism to address data scarcity. The mechanism combines multilingual embeddings with Brahui-specific embeddings, enabling the model to retain cross-lingual knowledge while learning language-specific semantic and morphological features. Experiments show substantial improvement over the baseline: semantic similarity increases from 79.91 to 98.23, perplexity decreases from 94.27 to 14.64, and the system achieves 76% accuracy in correctly answering questions. The proposed framework provides a robust and parameter-efficient approach for developing fluent generative dialogue systems in severely under-resourced linguistic contexts.

1. Introduction

Large Language Models (LLMs) contribute to the state of the art in conversational artificial intelligence for high-resource languages, but their performance drops sharply when applied to low-resource languages with limited resources [1,2]. Analogous scaling in multilingual generation, such as Cheetah, confirms that moderate coverage helps, although not entirely [3]. Low-resource adaptation continues to mature. Targeted fine-tuning, parameter-efficient adapters, and selective training significantly improve downstream performance even with minimal data. The adapted MLLM framework demonstrated that focused fine-tuning produces stable improvements for low-resource languages. Other work has shown effective transfer through monolingual adaptation, cross-lingual alignment, and unsupervised learning [4,5].

General multilingual models have difficulty handling languages such as Brahui due to their lack of corpora, unique morphology, and minimal presence in multilingual dictionaries [6]. These deficiencies result in poor tokenization, unstable embedding, and inaccurate generation. The parameter-efficient approach works only when there is some basis of knowledge in the target language by the model, which is practically non-existent for Brahui [7]. Recent approaches address this by adjusting tokenization and embeddings for previously unseen or underrepresented languages. Trans-Tokenization and vocabulary transfer techniques exploit semantically related high-resource languages to initialize missing embeddings [8]. Dynamic aggregation schemes can be used to make model fine-tuning more robust in a low-resource setting, since they calculate an aggregate of all multilingual items adjusted to their contribution [7]. These improvements point toward hybrid, language-aware solutions.

Brahui lacks sufficient corpora annotated conversational datasets and representation in existing

© Mehran University of Engineering and Technology 2026

multilingual vocabularies [9,10]. As a result, current LLMs fail to produce a reliable, coherent, and context-appropriate Brahui dialogue system. A dedicated adaptation strategy is required to improve linguistic grounding while retaining cross-lingual generalization from the neighboring languages, such as Urdu and Persian.

In the realm of low-resource natural language processing, existing models such as Recurrent Neural Networks and conventional n-gram models have been used; however, they are fundamentally incapable of handling the complex morphological dependencies of a Dravidian language, such as Brahui. In contrast, the transformer architecture makes use of a self-attention mechanism that is able to preserve contextual relationships across a dialogue sequence. In this study, Qwen2.5-0.5B-Instruct has been chosen as the base model for the following reasons. First, it is a light model having parameters of 0.5 billion, thus suitable for very low resource environments. It provides a strong, pretrained multilingual model without the need for high computational resources like large models with billions of parameters. It is a perfect blend of complexity and efficiency; thus, the proposed methods of embedding fusion and QLoRA can be effectively used to adapt the model to the Brahui language without suffering from catastrophic forgetting.

The remainder of this paper is structured as follows: Section 2 provides the relevant background in the area of multilingual language modeling, transformer-based dialogue systems, and low-resource adaptation techniques. The proposed approach, including the dataset, system design, language pre-processing, embedding fusion, and training, is discussed in detail in Section 3. Section 4 outlines the metrics for evaluating the proposed model, including semantics, relevance, and perplexity. In Section 5, the Results and Discussion are provided, where a detailed analysis of the experimental outcomes, along with a comparative analysis of the proposed and baseline models, is presented, followed by the limitations and future scope of the proposed model in Section 6, and concluding remarks in Section 7.

2. Related Work

Multilingual language models such as mBERT, XLM-R, and mT5 have enabled cross-lingual transfer, yet their performance drops sharply for languages with limited corpora or unique morphology [11,12]. Studies have shown that selective task-adaptive pretraining can improve low-resource NLP tasks, such as sentiment analysis or NER, when source languages are carefully chosen [8]. Despite improvements from scaling and pretraining, severely under-resourced languages still lag in practical NLP applications [9,13].

2.1 Transformer Architectures for Dialogue Systems

Transformer-based dialogue models have shown promise for low-resource conversational AI. Instruction-tuned models and task-adapted LLMs can provide functional chatbots for low-resource languages and dialects when combined with fine-tuning and adaptation strategies[14]. However, these approaches typically rely on languages with little digital presence, leaving extremely low-resource or typologically unique languages insufficiently addressed [11].

Table 1
Key Research Contributions for Low-Resource and Multilingual Language Modeling

S #	Author(s), Year	Work Description	Strengths	Weaknesses	Ref
1	Lankford et al., 2023	adaptMLLM: Fine-tuning multilingual LLMs for low-resource languages	Improves downstream tasks with minimal data	Limited to classification /translation; no chatbot	[1]
2	Imani et al., 2023	Glot500: Multilingual corpora for 500 languages	Broad multilingual coverage; improved low-resource performance	A gap persists for extremely under-resourced languages	[2]
3	Trautmann et al. 2025	Cheetah: NLG for 517 African languages	Demonstrates scalable multilingual generation	Evaluation mainly on generation; limited dialogue tasks	[16]
4	Sel et al. 2024	Parameter-efficient adaptation (adapters) for low-resource translation	Reduces computational cost; retains strong performance	Limited to translation; no embedding fusion	[7]

2.2 Adaptation Techniques

Parameter-efficient fine-tuning techniques, such as QLoRA or bottleneck adapters, can effectively adapt models to low-resource settings while updating only a small portion of parameters [15]. Embedding adaptation strategies, including Trans-Tokenization and cross-lingual vocabulary transfer, initialize embeddings for unseen languages using semantically similar high-resource embeddings [8]. Dynamic aggregation and federated-style fine-tuning further stabilize performance for extremely low-resource languages [16].

2.3 Gaps in Existing Research

Current approaches rarely target generative dialogue tasks for extremely low-resource languages. Embedding adaptation techniques often focus on single-source embedding rather than fusing multilingual embedding with language-specific fine-tuned embedding. A systematic approach is lacking to building transformer-based chatbots for languages with unique morphology and limited digital resources. A Summary of Recent Advances in Multilingual and Low-Resource NLP is shown below in Table 1.

Table 1 summarizes recent research on improving Multilingual Large Language Models (MLLMs), focusing on techniques like fine-tuning, corpus creation, and parameter-efficient adaptation (adapters) to enhance their performance, particularly for low-resource languages.

3. Methodology

3.1 Dataset Description

The experiments use a Brahui question-answer dataset curated from local conversational sources, consisting of exactly 5,000 Query-Response pairs. The dataset contains pairs on different topics to reflect real-life conversations, with samples of these interactions provided in Table 2. Data preprocessing included normalizing the text, removing empty entries, and structuring each item as a prompt-answer pair for causal language modeling. For this study, the dataset was divided into 90% for training and 10% for the validation set. This small yet well-chosen collection highlights the challenges of low-resource language modeling.

Table 2 shows some example sentences from the Brahui dialogue dataset, which help to illustrate the flow of interaction between the user queries and system responses. These example sentences help to highlight the structure of the sentences in the dialogue dataset, which has the ability to handle simple dialogue scenarios, such as greetings and service queries.

Table 2
Dataset Sample

Sample #	Question	Answer
1	نی امروس	ای جوان اوٹ نی امروس
2	کنے نا دارکارے	جی پار ای نے کہ انت کیو
3	خدا حافظ	منتوار

3.2 System Architecture Overview

The proposed framework is a transformer-based conversational artificial intelligence system engineered for the Brahui language. For this system, the Qwen2.5-0.5B-Instruct model was used as the base large language model and integrated into Brahui embeddings through a Transformer architecture. The architecture consists of four main components: (1) a preprocessing module intended to process Brahui text, (2) a tokenizer extended to cover Brahui characters, (3) an embedding fusion layer combining multilingual pretrained embeddings with Brahui-specific vectors, and (4) a parameter-efficient fine-tuning module based on QLoRA. This design ensures that the model maintains cross-lingual knowledge while learning language-specific semantic and morphological features of Brahui. The overall architecture of the proposed Brahui chatbot system is illustrated in Fig. 1

Fig. 1 illustrates that first, Brahui text undergoes a preprocessing-morphological analysis, tokenization, and normalization. Since Brahui is a low-resource language, the Embedding Layer applies the approach of Mean Embedding Injection: new Brahui tokens are initialized from the statistical centroid of the pretrained Qwen2.5-0.5B latent space. Then, this combined representation passes through the Transformer Architecture that makes use of self-attention for grasping deep contextual meaning. Meanwhile, the Dialogue Manager tracks the state of the conversation, intents, and entities for coherence. At the end, the Response Generation Module uses QLoRA Fine-Tuning for parameter-efficient adaptation, producing a final, contextually accurate user output.

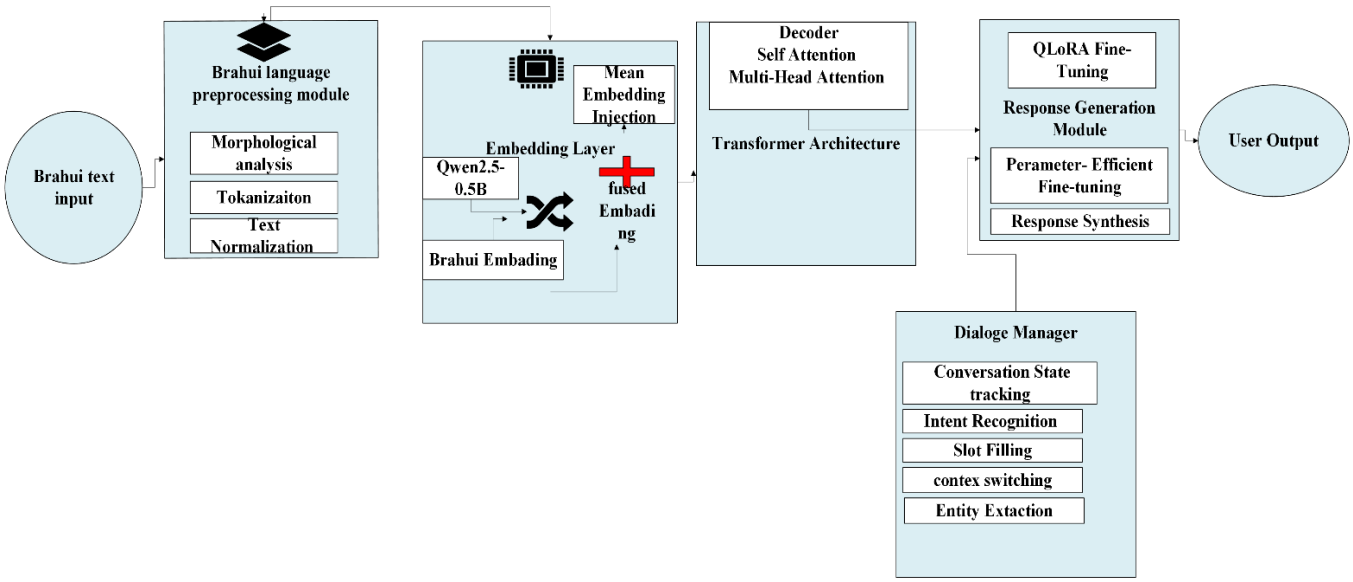

Fig. 1. System Architecture of the Brahui Embedding-Fusion Conversational Model

3.3 Language-specific Preprocessing Pipeline

Given the unique script and morphology of Brahui, a language-dependent preprocessing pipeline was developed. Raw question-answer pairs were cleaned, normalized, and tokenized. Missing or empty fields were removed to ensure high-quality data. Each input was formatted as a prompt-answer pair, i.e., "Question (نا پن دیرے؟) (Answer: (کنا پن نصیراحمداے), enabling underlying language modeling. The preprocessing ensures consistent alignment with the model's generative capabilities and reduces noise from low-resource datasets.

3.4 Embedding Fusion Technique

To accomplish the integration of Brahui-specific knowledge without hampering the ability of being multilingual, embedding fusion was implemented in the code. After adding 36 Brahui characters to the base tokenizer, the embedding matrix of the base model has been updated. The new Brahui embeddings were initialized with the mean of the existing embeddings in a multilingual manner with a small amount of noise drawn from a Gaussian distribution to promote diversification of the embeddings. This fusion strategy reduces the risk of misaligned embeddings and improves initial grounding in a language with minimal pre-existing resources.

3.5 Model Training Strategy

The model was fine-tuned using QLoRA, an efficient method that updates only a small part of the parameters. The QLoRA setup in this work targets only the transformer's query, key, value, and output projection layers. Training combined transfer learning from the base LLM with Brahui-specific adaptation: this allows the model to learn effectively from a small dataset while retaining the generalization advantage of the pretrained multilingual model.

3.6 Implementation Details

The training methodology for the model was optimized for performance on the Brahui script, while maintaining computational efficiency. Due to the limited resources of the language, the training was achieved through Parameter Efficient Fine-Tuning (PEFT), implementing the QLoRA technique, which is capable of fine-tuning the transformer model efficiently with minimal memory usage. The implementation settings are summarized in Table 3 below.

Table 3 shows that the model utilizes an extended tokenizer that features Brahui characters, an alignment of the padding tokens (PAD) and end-of-sequence tokens (EOS) to ensure script compatibility. The training is carried out on modest hyperparameters, a batch size of 2, accumulation over 8 steps with a learning rate of 2e-4, precision to FP16, and training for a total of 20 epochs. The dataset has been divided in the ratio of 90:10 for training and validation, respectively. Evaluation is done at the end of every epoch, with automatic checkpointing enabled. The system is implemented using PyTorch and Hugging Face Transformers, whereas QLoRA enables parameter-efficient fine-tuning.

© Mehran University of Engineering and Technology 2026

Table 3
Model Configuration and Training Setup for the Proposed Brahui Chatbot

Component	Description
Tokenizer	Extended with Brahui characters to support the unique script and morphology. PAD token set to EOS for compatibility.
Hyperparameters	Batch size: 2, Gradient Accumulation: 8, Learning Rate: 2e-4, FP16 precision, 20 training epochs.
Dataset Split	90% used for training, 10% used for validation.
Evaluation	Performance is monitored at each epoch. Checkpoints are saved automatically, and the final fine-tuned model is stored in a dedicated directory.
Environment	PyTorch and Hugging Face Transformers are used for training. PEFT (QLoRA) applied for parameter-efficient fine-tuning.

4. Evaluation Metrics

4.1 Semantic Similarity

Semantic similarity evaluates the extent to which a model's generated output preserves the original meaning of a reference text, as ensuring semantic fidelity is critical for the reliability of generation tasks. Following the methodology established in [17]. This relationship is mathematically defined using cosine similarity. For two textual embeddings u, v ∈ Rn the similarity is calculated as:

$$\text{Semantic Similarity } (U,V) = \left(\frac{\boldsymbol{U.V}}{\|\boldsymbol{U}\|\boldsymbol{V}\|}\right) \quad \textbf{Eq. (1)}$$

Where is the dot product

u.v denotes the dot products of the vectors.

$\| \mathbf{u} \|$ and $\| \mathbf{v} \|$ are the L_2 norms.

The value ranges from -1 to +1, where +1 signifies identical meaning, and -1 indicates completely opposed semantics for a sequence $A = (a_1, a_2, \ldots, a_N)$. Generated under context C, this measure ensures the output is semantically grounded. While this metric assesses meaning preservation, the model's ability to maintain dialogue coherence is measured by the Contextual Relevance Score, discussed in the following section.

4.2 Contextual Relevance Score

The contextual relevance score, by definition, looks at the degree of coherence of a model's output with its surrounding context, so that responses are appropriately conditioned on what has come before. The following mathematical formulation has been introduced to capture this relationship [18].

$$Rel(C,A) = \frac{1}{N}\sum_{t=1}^{N} LogP(a_t|c, a<t) \quad \textbf{Eq. (2)}$$

Where, C = previous dialogue context, A = candidate response, a_t = token at position t, and a(<t) = preceding tokens (a_1, a_(t−1). N: total number of tokens in the response. This score indicates the average log likelihood of the response, given the context. The higher the number, the more contextual relevance.

4.3 Perplexity Score

Perplexity is a measure of how well a model can predict a sequence. It may be used as the primary indicator of fluency and the quality of the language modeling. The following mathematical formula has been derived from the literature to calculate this measure, as proposed [19].

$$Perplexity = exp\left(\frac{1}{N}\sum_{t=1}^{N} -LogP(x_t|c, x<t)\right) \quad \textbf{Eq. (3)}$$

Let N be the number of tokens in the output; let xt be the token at position t, and let x_(<t) be the sequence of all tokens coming before it. P(x_t | C, x_(<t)) is the probability that the model assigns to generating token xt given the context C and the tokens that come before it. This equation first computes the mean of the

negative log-probabilities for the tokens and later raises them to a power. This produces a measure wherein smaller values imply better fluency and predictability of the model, while larger values suggest more uncertainty of token production.

5. Results and Discussion

This section contains the principal findings of the experiments conducted for training and testing the model, especially the influence of QLoRA fine-tuning using embedding fusion in terms of convergence, generalization, and performance. The tables and bar graphs presented in this section demonstrate the trends in training throughout each epoch, and the performance is clearly better than that of the baseline model.

5.1 Evaluation Metrics Setup and Scaling

The model's generation quality was measured using three key metrics: Semantic Similarity, Contextual Relevance, and Perplexity. The formulas used to calculate the Semantic Similarity measure, as described in the evaluation section, yield values ranging from -1 to +1. However, for consistency across all following figures and tables, both Semantic Similarity and Contextual Relevance are normalized on a scale of 0–100, as follows:

$$\text{Scaled Score} = \left(\frac{\text{Raw Score} + 1}{2}\right) \times 100 \quad \textbf{Eq. 4}$$

Under this scale system, 100 represents perfect congruence of semantics or topic relevance, while 0 means absolute discrepancy. The Perplexity Score, on the other hand, is a raw, unscaled score, where a smaller numeric value indicates higher model fluency.

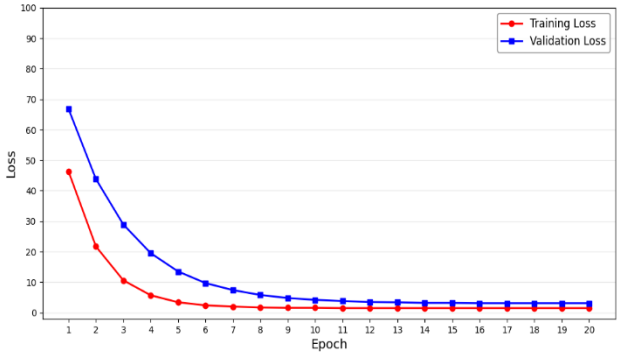

Fig. 2. Training vs. Validation Loss Trend Across Epochs

As shown in Fig. 2, in the first four epochs, the training loss decreases significantly. This indicates the rapid adaptation of the model during QLoRA fine-tuning. After that, the training loss remains stable. At the same time, the validation loss presents a smooth decreasing trend. This indicates the good cross-lingual generalization of the model without any catastrophic overfitting.

Table 4
Evaluation Metrics for Baseline vs. Embedding-Fusion Fine-Tuned Model

Model	Semantic similarity	Contextual Similarity	Perplexity Score
Qwen2.5-0.5B-Instruct model	79.91	84.90	94.27
QLoRA Fine-Tuning with Embedding Fusion	98.23	68.72	14.64

Table 4 shows a comparative analysis between the baseline model and the fine-tuned QLoRA model with embedding fusion. With regards to Semantic Similarity, we can see an increase from 79.91 to 98.23, while Perplexity decreases dramatically from 94.27 to 14.64, implying that there is a substantial improvement in fluency. However, Contextual Relevance drops from 84.90 to 68.72.

This reduction indicates a critical trade-off that is inherent in the Embedding Fusion Method when handling languages with very few resources. In doing this, by using the centroid of the multilingual latent space to initialize the Brahui token embeddings, the model quickly learns the vocabulary and structure, significantly boosting its token-level performance (Semantic Similarity). However, this immediate change in the embedding space causes noise within the attention heads of the pretrained model. While the model delivers accurate answers, it may sometimes fail to retain context during the conversation.

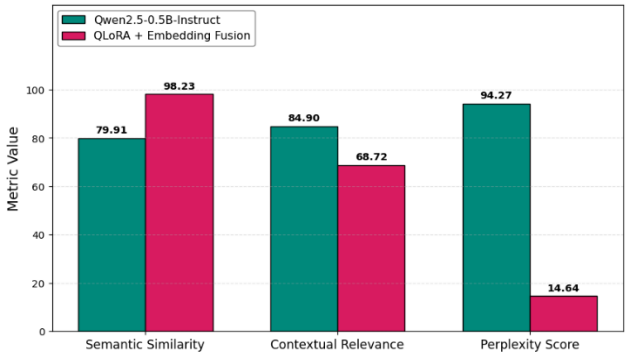

Fig. 3. Performance Comparison: Baseline vs. Fine-Tuned Model

Fig. 3 compares Qwen2.5-0.5B-Instruct and the QLoRA + Embedding Fusion model on three metrics: Semantic Similarity, Contextual Similarity, and

© Mehran University of Engineering and Technology 2026

Perplexity Score represented as unscaled values. While the QLoRA with embedding fusion yielded higher semantic alignment and significantly lower perplexity, indicating better fluency and higher predictive capability, the contextual consistency was marginally better for the baseline Qwen model. Overall, the QLoRA-fused model outperforms others in terms of linguistic precision and generation quality.

Table 5
Model Evaluation Across Semantic, Contextual, and Perplexity Metrics Without Fusion

Model	Semantic similarity	Contextual Similarity	Perplexity Score
Qwen2.5-0.5B-Instruct model	79.91	84.90	94.27

According to Table 5, the Qwen2.5-0.5B-Instruct model reaches a Semantic Similarity of 79.91, which is a mid-level alignment between generated outputs and reference responses. The strong topical coherence of the responses is represented by a Contextual Similarity score of 84.90. Its Perplexity Score of 94.27 infers high uncertainty in token prediction, hence pointing to some limitations regarding fluency and generation poise before the application of any fusion-based enhancements.

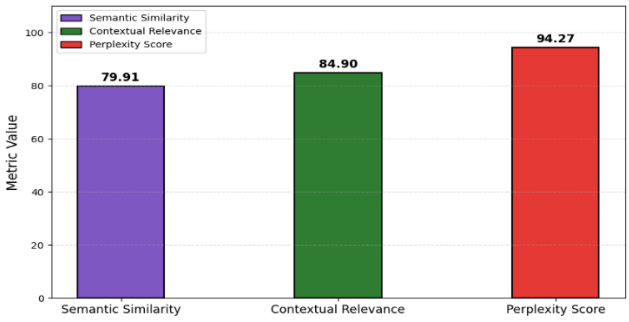

Fig. 4. Evaluation Results for the Qwen2.5-0.5B Model Without Fusion

Table 6
Evaluation Metrics for the QLoRA Fine-Tuned Model with Embedding Fusion

Model	Semantic similarity	Contextual Similarity	Perplexity Score
QLoRA Fine-Tuning with Embedding Fusion	98.23	68.72	14.64

Fig. 4 shows the evaluation metrics of the Brahui-adapted Qwen2.5-0.5B model before fusion, presented as unscaled values. It scores 79.91 on Semantic Similarity, 84.90 on Contextual Relevance, and 94.27 on Perplexity. The results set the baseline for further improvement that would be carried out through fusion.

Table 6 shows that the QLoRA Fine-Tuned model with Embedding Fusion achieves a high Semantic Similarity score of 98.23, indicating strong alignment between generated outputs and reference texts. The Contextual Similarity score of 68.72 indicates that there is a moderate degree of consistency with the target topic. A much lower Perplexity score of 14.64 reflects great enhancement regarding fluency generation and token-level predictability compared to the baseline model.

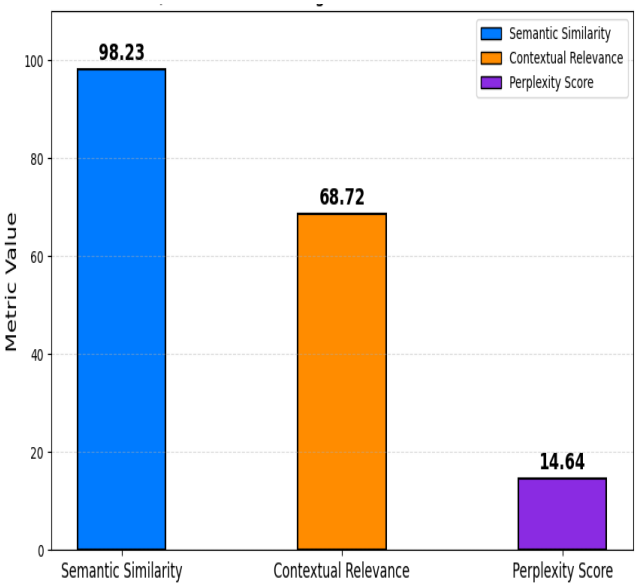

Fig. 5. Fused Model Evaluation Metrics: QLoRA and Embedding Fusion

Fig. 5 is a graph of the test scores of the Brahui-fused Qwen2.5-0.5B model, in which it scores 98.23 on the Semantic Similarity scale, 68.72 on the Contextual Relevance scale, and 14.64 on the Perplexity scale. The scores have been calculated using the raw scores of the metrics. It can be seen that there have been considerable improvements in the semantic alignment of the fused model.

Table 7 above is an indication of the sensitivity analysis that is done to determine the change from the original Qwen2.5-0.5B-Instruct model to the QLoRA fine-tuned model, which makes use of the embedding fusion method. In the original model, the sensitivity is low and does not involve the use of the MERT method, but rather the default tokenizer that has limited Brahui support. In the other model, the sensitivity is high due to the use of the embedded fusion method that

© Mehran University of Engineering and Technology 2026

enhances alignment through the use of the tokenizer that supports the Brahui environment.

Table 7
Comparative Analysis of Model Adaptation Factors

Factor	Qwen2.5-0.5B-Instruct model	QLoRA Fine-Tuning with Embedding Fusion
Reordering	Low sensitivity; uses standard attention.	Higher sensitivity due to fused embeddings and adapted representations.
Tuning (MERT)	Not specifically optimized.	Improved alignment from QLoRA fine-tuning.
Feature Fusion	None.	Present: incorporates Brahui-focused fused embeddings.
Tokenization Method	Default Qwen tokenizer; limited Brahui coverage.	An extended tokenizer with Brahui characters provides better segmentation.

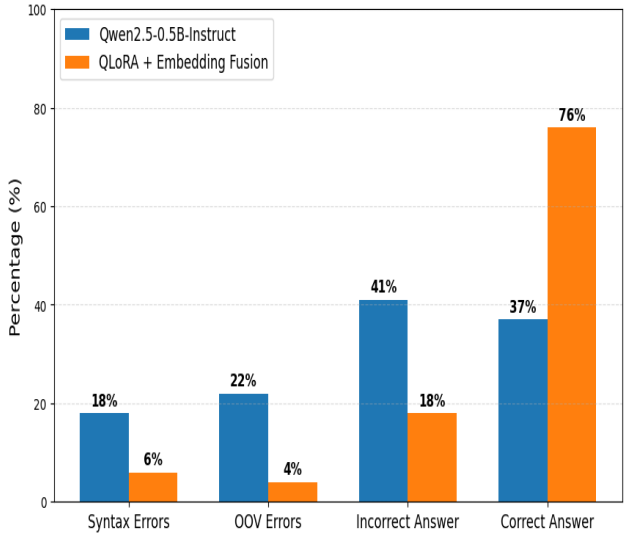

Fig. 6. Qualitative Error Analysis: Baseline vs. Fused Model

Fig. 6 Comparative Error Analysis. This chart compares performance across key error categories. This proposed approach of QLoRA and embedding fusion has a significant improvement in Out of Vocabulary (OOV) and syntax errors, resulting in a 76% correct answer rate compared to a 37% baseline.

Table 8
Qualitative Error Analysis of Baseline vs. Fused Model

Model	Syntax Errors (%)	OOV Errors (%)	Incorrect answer (%)	Correct answer (%)
Qwen2.5-0.5B-Instruct model	18	22	41	37
QLoRA Fine-Tuning with Embedding Fusion	6	4	18	76

According to Table 8, the baseline Qwen2.5-0.5B-Instruct model has much higher syntax and OOV errors due to limited Brahui coverage, thus yielding only 37% correct answers. By comparison, QLoRA Fine-Tuned with Embedding Fusion significantly reduces both structural and vocabulary errors, reaching 76% correct responses. This points to improvement in robustness and accuracy of answering outcomes for low-resource question answering.

The proposed work has managed to counter the difficulty of conversational AI in the low-resource Brahui language by proposing a transformer conversational AI model based on QLoRA fine-tuning and the Embedding Fusion Technique. The Embedding Fusion Technique, which extends the base tokenizer and creates new Brahui embeddings based on the multilingual environment, significantly enhances linguistic grounding. The empirical analysis has justified the effectiveness of the Embedding Fusion Technique by observing incredible improvements in Semantic Similarity (79.91% to 98.23%) and a remarkable drop in Perplexity value from 94.27 to 14.64, resulting in an extraordinary increase in Correct Answer Rate from 37% to 76%, thus minimizing syntax as well as Out-of-Vocabulary errors. Nevertheless, while utilizing such an effective adaptation technique in low-resource conversational AI tasks, the decline in Contextual Relevance indicates that further investigations are required to maintain overall conversational consistency.

6. Limitations and Future Work

Although the Brahui chatbot shows considerable advancement, its performance is limited by the small dataset. The limited dataset restricts the chatbot's ability to generalize across the full range of Brahui linguistic variation. Although the embedding fusion technique improves semantic accuracy and fluency,

© Mehran University of Engineering and Technology 2026

overall performance remains constrained by data scarcity.

To address these limitations, future work will focus on the following three key directions: First, any loss of context will be minimized by methods such as multi-task learning or context-aware regularization while fine-tuning to stabilize dialogue coherence and reduce the trade-off. Second, work will be directed at scaling and generalization by advanced augmentation techniques and a systematic validation of the Embedding Fusion Technique across other typologically distinctive, extremely low-resource languages in order to study its wider applicability. Lastly, future work will investigate a dynamic weighting mechanism in the fusion layer that allows the model to balance adaptively between the specialized Brahui feature representations and the generalized capabilities of the base model.

Conclusion

This paper proposes and validates a novel strategy to build a functional conversational AI for the extremely low-resource Brahui language, a domain where contemporary multilingual large language models fail to perform well. By utilizing the Qwen2.5-0.5B-Instruct model and introducing a domain-specific Embedding Fusion Technique combined with QLoRA fine-tuning, this study tackles the core problem of weak linguistic grounding caused by the limited vocabulary coverage and dataset constraints. Empirical evaluation results highlight the effectiveness of the proposed method. Concretely, perplexity significantly drops from 94.27 to 14.64, while semantic similarity considerably increases from 79.91 to 98.23. In terms of quality, the method attains a high Correct Answer Rate of 76% to emphasize the necessity of domain-specific embedding adaptation to tackle the heavy restriction imposed on the generative model. Although embedding fusion reduces contextual balance, the gain in linguistic accuracy confirms the effectiveness of Embedding Fusion in tackling the digital divide problem in severely under-resourced languages through an efficient and scientifically valid framework.

Acknowledgement

The authors would like to thank the Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, for providing the research facilities to conduct this study at the AI Lab. This lab was established under a project funded by the Sindh Higher Education Commission, under the Sindh Laboratory Strengthening Program 2023–24.

References

[1] S. Lankford, H. Afli, and A. Way, "adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds," Information (Switzerland), vol. 14, no. 12, Dec. 2023, doi: 10.3390/info14120638.

[2] A. Imani et al., "Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages," in Proc. of the 61st Annual Meeting of the Assoc. for Computational Linguistics (Vol. 1: Long Papers), Toronto, Canada, 2023, pp. 1082–1117.

[3] I. Adebara, A. Elmadany, and M. Abdul-Mageed, "Cheetah: Natural Language Generation for 517 African Languages," arXiv preprint arXiv:2401.01053, Jan. 2024.

[4] R. Khademi and L. Moradi, "Cross-Lingual Transfer in Low-Resource NLP Using Transformer Models," Data Science Journal, 2024. [Online]. Available: www.datasciencejournal.net

[5] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data," Applied Sciences (Switzerland), vol. 13, no. 2, Jan. 2023, doi: 10.3390/app13021201.

[6] N. Ahmed, M. Ahmed Khuhro, M. Ali Dootio, and S. Pakistan, "Building and Analyzing a Brahui Text Corpus: Applying DTM and TF-IDF Techniques," International Journal of Applied Engineering and Technology, 2024.

[7] I. Sel and D. Hanbay, "Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation," Mathematics, vol. 12, no. 19, Oct. 2024, doi: 10.3390/math12193149.

[8] F. Remy, P. Delobelle, H. Avetisyan, A. Khabibullina, M. de Lhoneux, and T. Demeester, "Trans-Tokenization and Cross-lingual Vocabulary Transfers: Language Adaptation of LLMs for Low-Resource NLP," arXiv preprint arXiv:2408.04303, Aug. 2024.

[9] S. Gull, N. Ahmed, and S. Ahmed Sheikh, "Deep Learning-Driven OCR System for Brahui Printed Text: Bridging the Digital Gap in Low-Resource Language Processing," Liberal

© Mehran University of Engineering and Technology 2026

Journal of Language and Literature Review, vol. 12, no. 1, 2024. [Online]. Available: https://llrjournal.com/index.php/11

[10] N. Ahmed, "A Review of Existing Machine Translation Approaches, their Challenges and Evaluation Metrics," Pakistan Journal of Engineering, Technology and Science, vol. 11, no. 1, pp. 29–44, Dec. 2023, doi: 10.22555/pjets.v11i1.1002.

[11] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural Language Processing Applications for Low-Resource Languages," Natural Language Processing, vol. 31, no. 2, pp. 183–197, Mar. 2025, doi: 10.1017/nlp.2024.33.

[12] Z. Lamin and A. A. Aziz, "Cross-Lingual Sentiment Analysis in Low-Resource Languages: A Recent Review on Tasks, Methods and Challenges," International Journal of Advanced Computer Science and Applications, 2024. [Online]. Available: www.ijacsa.thesai.org

[13] M. Arul Selvan, "Multilingual LLMs for Low-Resource Language Translation and Preservation," ResearchGate, Jun. 2025. [Online]. Available: https://www.researchgate.net/publication/392551935

[14] T. Bakkenes, D. Wang, and A. Johansson, "Fine Tuning Methods for Low-Resource Languages," arXiv preprint arXiv:2510.04139, Oct. 2025.

[15] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, May 2023.

[16] E. Trautmann, I. Hales, and M. F. Volk, "Aggregating Low Rank Adapters in Federated Fine-Tuning," arXiv preprint arXiv:2501.06332, Jan. 2025.

[17] D. Cer et al., "Universal Sentence Encoder," arXiv preprint arXiv:1803.11175, Apr. 2018.

[18] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A Diversity-Promoting Objective Function for Neural Conversation Models," in Proc. of the 2016 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT), San Diego, CA, 2016, pp. 110–119.

[19] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the Limits of Language Modeling," arXiv preprint arXiv:1602.02410, Feb. 2016.

© Mehran University of Engineering and Technology 2026